Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status : A longitudinal data analysis

Asare, Kennedy Opoku

2022-07

Asare , K O , Moshe , I , Terhorst , Y , Vega , J , Hosio , S , Baumeister , H , Pulkki-Råback , L & Ferreira , D 2022 , ' Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status : A longitudinal data analysis ' , Pervasive and Mobile Computing , vol. 83 . https://doi.org/10.1016/j.pmcj.2022.101621

http://hdl.handle.net/10138/354117
https://doi.org/10.1016/j.pmcj.2022.101621

Pervasive and Mobile Computing 83 (2022) 101621

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis

Kennedy Opoku Asare [a,*], Isaac Moshe [b], Yannik Terhorst [c], Julio Vega [d], Simo Hosio [a], Harald Baumeister [c], Laura Pulkki-Råback [b], Denzil Ferreira [a]

[a] Center for Ubiquitous Computing, University of Oulu, Oulu, Finland
[b] Department of Psychology and Logopedics, University of Helsinki, Helsinki, Finland
[c] Department of Clinical Psychology and Psychotherapy, Ulm University, Ulm, Germany
[d] Department of Medicine, University of Pittsburgh, Pittsburgh, USA

ARTICLE INFO

Article history:
Received 15 August 2021
Received in revised form 28 March 2022
Accepted 17 May 2022
Available online 24 May 2022

Keywords:
Smartphone
Digital biomarkers
Mental health
Depression

ABSTRACT

Depression is a prevalent mental disorder. Current clinical and self-reported assessment methods of depression are laborious and incur recall bias. Their sporadic nature often misses severity fluctuations. Previous research highlights the potential of in-situ quantification of human behaviour using mobile sensors to augment traditional methods of depression management. In this paper, we study whether self-reported mood scores and passive smartphone and wearable sensor data could be used to classify people as depressed or non-depressed. In a longitudinal study, our participants provided daily mood (valence and arousal) scores and collected data using their smartphones and Oura Rings. We computed daily aggregations of mood, sleep, physical activity, phone usage, and GPS mobility from raw data to study the differences between the depressed and non-depressed groups and created population-level Machine Learning classification models of depression. We found statistically significant differences in GPS mobility, phone usage, sleep, physical activity and mood between depressed and non-depressed groups. An XGBoost model with daily aggregations of mood and sensor data as predictors classified participants with an accuracy of 81.43% and an Area Under the Curve of 82.31%. A Support Vector Machine using only sensor-based predictors had an accuracy of 77.06% and an Area Under the Curve of 74.25%. Our results suggest that digital biomarkers are promising in differentiating people with and without depression symptoms. This study contributes to the body of evidence supporting the role of unobtrusive mobile sensor data in understanding depression and its potential to augment depression diagnosis and monitoring.

1. Introduction

Depression is a prevalent mental disorder that affects more than 260 million people worldwide [1,2]. People affected with depression typically experience recurrent episodes of symptoms which may include sadness and guilt, psychomotor retardation and agitation, unusual sleep changes, low energy levels, social and physical isolation, and loss of interest

* Corresponding author.
E-mail address: kennedy.opokuasare@oulu.fi (K. Opoku Asare).

https://doi.org/10.1016/j.pmcj.2022.101621

in daily activities [1,3]. Depression could lead to adverse health outcomes or suicide if not diagnosed and treated [1,2]. Additionally, depression has an adverse impact on the course of physical diseases such as stroke, cancer and increases the medical costs of many conditions [2,4]. Depression is the leading cause of global disease disability and morbidity; hence timely and effective mental health interventions have become an urgent priority [1,5–7].

While effective pharmacotherapy and psychotherapy for depression exist, inadequate assessment methods, unavailability of trained professionals, personal attitudes towards treatments and social stigma, among others, are all identified as barriers to the diagnosis and treatment of depression [1,8]. Clinician-rated or self-reported depression assessment methods [9–13] are applied typically with long assessment intervals, thus missing out on the temporal changes in symptoms of depression between assessment intervals [14–16]. Additionally, depression assessment methods are affected by recall biases in reconstructing past events and the tendency of respondents to provide socially desirable answers [17–19]. The ability to quantify in-situ (naturalistic environment and outside laboratories) psychological and spatial–temporal behaviour to augment traditional depression assessments has been a key research area in pervasive and mobile computing [3,18,20,21]. Such approaches are seen as promising in enabling early detection, effective diagnosis, and unobtrusive monitoring of depression (See Tables 1 and 2).

This study examines the potential of digital biomarkers and mood ratings as in-situ parameters in differentiating between participants with symptoms of depression (depressed group) and participants without symptoms of depression (non-depressed group). In this context, digital biomarkers refer to quantifiable human behavioural patterns (physiological, behavioural routines, and rhythms) that are passively and unobtrusively measured using smartphones and wearable devices [22,23]. Mood ratings refer to proactive assessments of the mood of participants using self-reports triggered at various times throughout the day. We analysed a dataset from a longitudinal study that combined smartphone and wearable sensing data to investigate the following questions:

1. Are there statistically significant digital biomarkers and mood differences between depressed and non-depressed participant groups?
2. Can digital biomarkers and mood ratings predict depressed or non-depressed groups?
3. What are the most important digital biomarkers and mood ratings in differentiating depressed from the non-depressed group?

2. Related work

2.1. Smartphone and wearable computing in mental health

Today, smartphones have become ubiquitous and have permeated every facet of our society [20,24–26]. Smartphones have embedded sensors (accelerometer, Bluetooth, GPS, Gyroscope, among others) that allow for the continuous and unobtrusive collection of datasets that provide insights into human behaviour [19]. Similarly, wearable devices like the Oura ring or Fitbit wristbands have embedded sensors that enable the collection of physiological datasets that provide granular measurements of sleep, heart rate variability (HRV), body temperature, Electrodermal Activity (EDA), and physical activity [19,21]. Smartphones and wearable devices also enable the collection of additional contextual information from users in naturalistic settings through Experience Sampling Methods (ESM) [17]. ESMs are proactive self-reports typically triggered throughout the day at multiple times, based on specific events or contexts [17].

Mobile Computing and Mental Health researchers have developed several mobile sensing applications such as AWARE [27], Beiwe [20], and Insights [24], given the opportunities that smartphones and wearables provide. These mobile sensing applications enable continuous, passive, unobtrusive data collection from smartphones and wearable embedded sensors or active data collection from users via ESM. These mobile sensing applications, coupled with the advancement of human behaviour modelling through statistical analysis and machine learning methods, have enabled the quantification of digital biomarkers [22,23,28], for the detection and monitoring of mental health symptoms and the development of timely mental health interventions [19].

A growing body of research in mental health has resulted in identifying digital biomarkers for the symptoms of mental disorders (Major depressive disorder, Bipolar disorder, Seasonal affective disorder, Post-traumatic stress disorders, and others). For instance, Location-based biomarkers, such as circadian rhythm [6,14], location regularity [18,29], time spent at home, time spent at specific location clusters, and location variance [6], inferred from GPS, WiFi, and Bluetooth sensors, has been shown to be related to symptoms of depression. In a large scale study with $N = 1765$ participants (mean age $= 18.94 \pm 2.22$, 60.49% female) over a two weeks, Müller et al. [29] investigated the relationship between depression and digital biomarkers of GPS mobility behaviours (distance travelled, entropy, and irregularity). The results of the study demonstrated a negative correlation ($r = -0.12$ 95% CI$[-0.20, -0.04]$, $p = 0.003$) and a negative association between depression and irregularity in mobility patterns.

It has been shown that sleep disturbances (such as insomnia, hypersomnia, irregular sleep time and wake-up time) are common manifestations of mental disorders [30,31]. Decades of sleep studies in mental health research using polysomnography have resulted in digital sleep biomarkers such as Total Sleep Time (TST), Random Eye Movement (REM), and Wake After Sleep Onset (WASO) [30]. Wang et al. [32], in their study with 48 participants over ten weeks, found statistically significant correlations between depression and sleep duration ($r = -0.360$, $p = 0.025$). In addition

Table 1
Summary of previous work on the relationship between digital biomarkers and depression in comparison to the current study.

Study	Demographics	Data	Findings
Müller et al. [29]	$N = 1765$, mean age = 18.94 ± 2.22, 60.49% female	GPS mobility	Location irregularity has a negative association with depression
Saeb et al. [7]	$N = 28$, 20 females, 8 males, age range 19–58	GPS mobility	Circadian movement, variance in the location visited are negatively correlated with depression. Significant differences in mobility patterns and phone usage between depressed and non-depressed participants
Opoku Asare et al. [39]	$N = 629$, from 56 countries, ICC 0.7584	Phone usage	Entropy of smartphone screen unlocks has a positive association with depression
Moshe et al. [15]	$N = 55$, age range 24–68, 25 males, 30 females	GPS mobility	Variance in locations visited has a negative association with depression. Total sleep time and time in bed has a positive association with depression
Wang et al. [32]	$N = 48$, 10 females, 38 males	Sleep, accelerometer and GPS mobility	Distance Travelled had a positive correlation with depression. Sleep duration and physical activity duration had a negative correlation with depression
Rozgonjuk et al. [37]	$N = 101$, mean age = 19.53 ± 4.31, 77 females	Mood ratings, Phone usage	Mood ratings has a positive correlated with depression
Wang et al. [3]	$N = 83$, mean age 20.13 ± 2.31, 40 male and 43 female	Phone usage and accelerometer	Stationary time and phone unlock duration have a positive correlation with depression. A significant difference in phone unlock duration between depressed and non-depressed participants
Current study	$N = 54$, mean age 43.04 ± 11.58, 30 females, 24 males	Sleep, Physical activity, Phone Usage, GPS location	Participants with symptoms of depression showed less mobility, more sleep time, less physical activity, and low mood. GPS mobility and Mood ratings showed the highest effect size in the statistically significant difference between depressed and non-depressed participants

to sleep disturbances, it is known that psychomotor retardation, restlessness and reduced mobility are associated with mental disorders [1,3]. Previous research has quantified these symptoms of depression from datasets such as smartphone keystrokes, touch screen patterns, accelerometer, and voice analysis [33,34].

Finally, previous research has demonstrated that mood changes are related to depression [35–38]. Several depression scales identify mood changes as a common symptom of depression. For example, the Patient Health Questionnaire (PHQ) [11] measures mood with questions such as *Little interest or pleasure in doing things?*, *Feeling down, depressed, or hopeless?*. The Depression Anxiety and Stress Scale (DASS) [10] has similar questions such as *I could not seem to experience any positive feeling at all.* Bowen et al. [38], in a study with 137 participants (59 in the non-depressed group and 78 in the depressed group), found that average mood ratings correlated with depression scores in the depressed group (r=0.38 95% CI[0.17, 0.55]). With the advancement of mobile computing, wearable technologies and ESMs, it is possible to collect mood ratings multiple times in a day, triggered periodically, or based on other contextual information like location or number of phone screen unlocks.

Taken together, previous work suggests that moment-to-moment smartphone and wearable datasets have the potential to show insightful trends in understanding depression. In addition to correlations, it is possible to understand the significant differences in digital biomarkers when comparing depressed and non-depressed participants. For example, in [7], the study found statistically significant differences in mobility patterns and phone usage frequency when comparing depressed and non-depressed participants. Similarly, in [3], the study found significant differences in phone usage (mean phone unlock duration and mean phone unlock duration at study places) when comparing depressed and non-depressed participants. We summarise key previous work on the relationship between digital biomarkers and depression in Table 1.

2.2. Predicting depression with mood and digital biomarkers

Beyond understanding the relationship between digital biomarkers, mood ratings and depression, previous work has demonstrated the feasibility of predicting depression symptom severity and depression status using statistical and machine learning analysis of digital biomarkers and mood ratings (See Table 2). In predicting depression symptom severity, previous studies have used statistical methods such as Linear mixed models or machine learning methods such as Random Forest Regression to predict depression severity in a continuum rather than a dichotomy. For example, Zulueta et al. [34], in their study with 16 participants (mean age 48.67 ± 9.63, 8 females) collected smartphone keystroke metadata and accelerometer datasets for 8 weeks. Clinicians rated the depression symptom severity of participants at weekly intervals using the 17-item Hamilton Depression Scale (HDRS) [40] and the Young Mania Rating Scale (YMRS) [13].

Table 2
Summary of previous work on predicting depression with mood ratings and digital biomarkers compared to the current study. RF = Random Forest, LR = Logistic Regression, KNN = K-Nearest Neighbour, SVM = Support Vector Machine, XGB = XGBoost.

Study	Demographics	Data	Analysis method	Results
Zulueta et al. [34]	$N = 16$, mean age 48.67 ± 9.63, 8 females	HDRS, YMRS, 8 weeks of phone keystrokes	Linear Mixed-Effect model	Keystroke pattern predicted depression
Pedrelli et al. [21]	$N = 31$, age range 19–73, mean age 33.7 ± 14	HDRS, 9 weeks of phone usage, GPS, HRV, temperature, and EDA	AdaBoost with RF regression	Smartphone and physiological datasets predicted the clinician-rated HDRS depression
Farhan et al. [41]	$N = 79$, age range 18–25, 73.9% female, and 26.1% male	PHQ, GPS, Physical activity	SVM and RF classifiers	GPS location and Physical activity biomarkers could predict participants' baseline depression status
Jacobson et al. [42]	$N = 23$, mean age 42.8 ± 11.0, 57% male)	MADRS, 2 weeks of Actigraphy data	XGB classifier	Actigraphy based biomarkers predicted baseline depression status with high accuracy and correlation
Jacobson and Chung [36]	$N = 31$, age range 18–27, mean age 19.13, 64.52% females	PANAS , DASS, 8 days of GPS, phone calls, weather, HRV	XGB classifier and RF regression	Population-level and participant level models with biomarker could predict observed depressed mood at a high correlation
Mastoras et al. [33]	$N = 25$,mean age 23.86 ± 4.44, 10 women, 15 men	PHQ, 124 days of smartphone keystroke patterns	SVM, RF and Gradient Boosting classifiers	Typing based biomarkers could predict participants' depression status
Kim et al. [35]	$N = 47$, mean age 78 ± 5.24	HDRS, 2 weeks of Actigraphy data	LR, Decision Tree, Boosted Trees, RF classifiers	Actigraphy based biomarkers predicted baseline depression status at high accuracy
Xu et al. [16]	$N = 138$	BDI, 106 days of Phone usage, Call, Bluetooth, GPS, Sleep, and Steps	AdaBoost with Decision Tree classifiers	Smartphone-based, sleep and steps biomarkers predicted depression at high accuracy.
Current Study	$N = 54$, mean age 43.04 ± 11.58, 30 females, 24 males	DASS, 30 days of mood ratings, Sleep, Physical activity, Phone Usage, GPS location	SVM, LR, RF, KNN, XGB classifiers	Digital biomarkers predict depression status as good as actively assessed mood plus digital biomarkers

Statistical analysis (Linear mixed-effects models) of participants' keystroke patterns showed a strong relationship between depression and smartphone keystroke metrics (R squared $= 0.63$, $p = 0.01$).

Likewise, in predicting depression status, previous studies train classification based models such as Support Vector Machine, Random Forest, Logistic Regression to classify participant's depression status. For example, Xu et al. [16] collected smartphone screen, call, Bluetooth, GPS location and Fitbit sleep and step count datasets from 138 participants in a study that lasted for 106 days. Participants provided self-reported symptoms of depression severity using the Beck Depression Inventory (BDI) [9] at the beginning and end of the study. AdaBoost (with Decision Tree classifiers) could predict participants' end-of-study depression status with an accuracy of 81.3%, precision of 84.3%, recall of 86.6% and F1 of 84.3%.

To sum up, the findings from previous work suggests the feasibility of applying machine learning and statistical methods to digital biomarkers quantified from smartphone and wearable sensors and mood ratings collected via ESMs, for the automatic classification of depression status and the continuous monitoring of depression symptoms severity. We summarise key previous work on predicting depression with digital biomarkers in Table 2.

3. Methods

In this paper, we obtained and further analysed the raw dataset from the study by Moshe et al. [15]. Moshe et al. [15] analysed the data to understand the correlation and predictive effect of digital biomarkers and mood on depression, anxiety and stress symptoms severity, using Pearson's correlation analysis and Multilevel regression model. This paper analyses the dataset to understand the statistically significant differences and effect sizes of mood and digital biomarkers when comparing depressed and non-depressed participants. Additionally, we employ machine learning methods to predict depression status with mood and digital biomarkers. We also use feature importance analysis to understand the most important biomarkers when differentiating depression status (depressed and non-depressed). In the subsequent paragraphs, we explain the data collection procedures and our analysis methods.

3.1. Participants

We briefly recap the participant recruitment protocol in [15]. Moshe et al. [15] recruited an initial batch of 60 participants with posts on online communities. The inclusion criteria were participants who: (1) owned an Oura ring, (2) were 18 years and above, (3) owned an iPhone with access to the internet, and (4) could read, understand and speak English.

Participants were taken through an informed consent process, in which they voluntarily signed an online informed consent form. An email was then sent to the participants with a URL to install a custom iOS application (study application) developed with AWARE [27] on their iPhone for 30 days. Participants' demographic data was collected via the study application after joining the study. Five (5) participants withdrew from the study: (2 for the burden of data collection, 1 for privacy reasons and 2 for unknown reasons). However, for the analysis in this article, we excluded one (1) additional participant because of insufficient data after data cleaning (described later in 3.4). Thus, we included 54 participants in our analysis.

3.2. Mental health measures

The symptoms of depression of participants were assessed, via the study application, at the study onboarding (baseline), at 2-weeks into the study (mid-point), and at the end of the study (end-point) via a self-reported 21-item DASS [10] . DASS has three subscales for assessing depression, anxiety and stress in the past week with example items such as "*I felt that I had nothing to look forward to*", "*I was unable to become enthusiastic about anything*". All items are rated on a 4-point Likert scale (*0 = Did not apply to me at all, 3=Applied to me very much or most of the time*). The score of each DASS subscale ranges from zero(0) to twenty-one (21), with higher scores indicating more severe symptoms. The internal consistency of DASS has been demonstrated with Cronbach's alpha of 0.96 for depression and 0.97 for the total scale [36,37].

This study used the DASS depression subscale score to measure depression symptom severity. We created two participant groups based on the baseline DASS depression scores with cut-offs proposed in the DASS protocol [10]. The participant groups were the non-depressed group (baseline DASS depression score 0–9) and the depressed group (baseline DASS depression score greater than 9). To capture the participants' depression status for the entire duration of the study, we examined the baseline, mid-point and end-point DASS depression scores. We observed that 8 out of 54 participants' depression status moved between depressed and non-depressed status during the study. We assigned each of these 8 participants to their most common depression status.

3.3. Daily mood ratings

For the dataset, daily mood ratings were collected with the Circumplex Model of Affect (CMA) [43]. The CMA conceptualises mood in two dimensions, that is, Valence (Negative/Positive) and Arousal (Negative/Positive), with each dimension rated on a 9-point Likert scale from −4 to 4 (low to high). An ESM notification was sent to participants' phones via the study application at approximately 9:00, 14:30 and 20:00 during the day. The ESM had a single item question, "How are you feeling right now?" with two Likert response scales (−4 to 4, zero as default mode) for Valence and Arousal. For our analysis in this study, we aggregate (average) the Valence and Arousal rating separately per day for each participant.

3.4. Digital biomarkers from smartphone and wearable sensor data

During the 30-day data collection period, participants were required to wear their Oura ring all the time. After the study, participants shared their Oura ring data for the 30-day duration with the research team. The Oura ring dataset included day level aggregate biomarkers of Sleep: Total Sleep Time (TST), Rapid Eye Movement (REM), Sleep Onset Latency (SOL), Wake After Sleep Onset (WASO), Sleep Efficiency (SE), Heart Rate Variability (HRV), Physical activity: Step count and Metabolic Equivalent for Task (MET). Within the same 30-day data collection period, the study application passively collected time-stamped GPS location every 5 min, phone screen unlocks (phone usage as detected by AWARE), and the participants' timezone every 60 min. Detailed information about the sensors (permissions, configuration), data encryption and obfuscation, secure data transfers over SSL from the participant's device to a secured MySQL database server, and user authentication can be found in the AWARE framework documentation [27].

For our analysis in this study, we computed 52 day-level (24 h from midnight to midnight) biomarkers from the passively sensed smartphone dataset using RAPIDS [28] — a tool for data pre-processing and biomarker computation. For GPS mobility biomarkers, we computed Location Variance, Total Distance, Location Entropy, Normalised Location Entropy [6,29], Log Location Entropy, Average speed, Circadian Movement, Number of Significant Places, Radius of Gyration, Moving to Static Ratio, Time at home [14,16], and Location Regularity (routine) Index [18,29]. From Phone screen unlock events, we computed Phone Usage Duration and Phone Usage Frequency. We applied the Circadian Rhythm [6,14], and Regularity Index [29,44] methods to compute Screen Regularity Index and Screen Circadian Rhythm (for weekdays, weekends and all days) [29].

We first excluded all features with zero variance and more than 15% of missing data to clean the computed biomarkers. We then dropped the entire participants' data with less than 15 days of data. We dropped days of participants with more than 30% of missing data. After the data cleaning process, we had 54 participants with 1556 days (mean 28.81, range 21–30 per participant) and 49 biomarkers. On average, participants were missing 7.56% (range 0% −12.98%) of data values.

3.5. Statistical analysis

First, we tested the normal distribution assumptions of the features with the Shapiro–Wilk test. Two-sided Mann–Whitney's U test, a non-parametric test, was used to test the statistically significant difference between the *depressed* and *non-depressed* group in mood ratings, GPS mobility, Phone Usage, and Wearable sensor biomarkers. We use the bootstrapping method with 1 million iterations and resampling of biomarkers with replacement as a non-parametric method of computing the differences in the group means. We also tested whether the confidence intervals for both groups' mean overlapped with the bootstrapping method. The difference between groups was assumed to be valid if the 2.5% and 97.5% quantiles of the resampled group means did not overlap. With the bootstrapping method, we also computed the effect size (Cohen's d) and the confidence interval of the effect size.

3.6. Predictive analysis with machine learning

In our predictive analysis with machine learning, we modelled (1) participant groups (depressed group with label 1 and non-depressed group with label 0) as a function of mood ratings (Valence and Arousal), Demographics (Gender and Age) and digital biomarkers (GPS mobility, Phone usage, Sleep, and Physical activity), and (2) participant groups as a function of demographics and digital biomarkers without mood ratings. We labelled each participant's daily digital biomarkers, mood ratings, age and gender with their depression status. Supplementary Figure 1 shows a diagrammatic representation of the data labelling.

To this end, we created population-based models leveraging five (5) supervised machine learning (ML) algorithms: Support Vector Machines (SVM), Random Forest (RF), XGBoost (XGB), K-Nearest Neighbour (KNN), and Logistic Regression (LR) [19]. We chose these ML algorithms because they are widely used in supervised classification, easy to train, interpretable [19]. These ML algorithms have also been used in previous related mental health research in Table 2.

We followed a robust ML model training approach with nested cross-validation to reduce the chances of model overfitting. We used a time-series aware leave-one-participant-day-out and stratified three-fold cross-validation for the outer and inner cross-validation, respectively. For each iteration of the nested cross-validation, one participant's day is designated as the test set, and the rest of the participants' dataset are designated as the training set. For time-series awareness, all training set samples recorded after the test set are removed from the training set. This ensures that future dataset is not used to predict the past, as this scenario is untenable in the real world.

The inner cross-validation was for missing data imputation, feature scaling, feature selection and hyperparameter optimisation of the classifiers. We optimised the hyperparameters of the classifiers using grid search over a predefined set of parameters, listed in Supplementary Table 3. Missing data imputation was done separately for each nested cross-validation iteration. We imputed missing data in the train set (participants' data for training the model) separately per participant. For training set imputation, we used a Bayesian Ridge Regression iterative feature imputation process [45] that uses all other features with no missing data as predictors. Gender was converted from a categorical to a numerical feature using one-hot encoding. We imputed missing values in the test set (i.e. one record of a participant's day) with the mean of the corresponding feature in the training set. Similar imputation of the test set and training set has been done in Low et al. [46] and Poulos et al. [47].

All features were scaled with min–max scaling. We applied feature scaling on the test set using the min–max parameters of the training set [46,47]. To mitigate biases in the output of the ML models, we handled the imbalanced training set by oversampling the minority class with the synthetic minority over-sampling technique (SMOTE). We then selected the 45 best features based on the mutual information between the features and the target (participants' depression status). We used the SHAP (SHapley Additive exPlanations) [48] method to compute feature importance.

We evaluated the predictive performance of the ML models, with the area under the receiver operating characteristic curve (AUC), F1, F1 Macro, Accuracy, Recall, and Precision metrics. We used three (3) baseline classifiers as a benchmark for the performance of the ML algorithms. The baseline classifiers were: (1) a naive classifier that predicts only the Majority Class (MC), (2) a Decision Tree (DT) classifier trained (same training approach as ML classifiers) with only the demographic dataset, and (3) a Random Weighted Classifier (RWC), that is, ten thousand randomly generated predictions according to the multinomial distribution of the depression and non-depressed group labels. We report the F1, Precision, Recall metrics for depressed (F11, Precision1, Recall1) and non-depressed groups (F10, Precision0, Recall 0).

3.7. Ethical considerations

The data collection and management followed all the local ethical guidelines in research. This study was exempt from formal ethical review board approval, based on the local ethical review board guidelines [49], since (1) the dataset used in this study follows the informed consent process; (2) all participants were above 18 years old; (3) the study does not intervene in the physical integrity of the participants; (4) our study does not expose participants to strong stimuli; (5) there is no intervention or a foreseeable potential for mental harm to the participants that exceed the limits of the participant's everyday life.

Table 3
Participants ($N = 54$) demographic. Participant groups were compared with a two-sided Mann–Whitney U test, except for Gender, which was compared with the Chi-squared test (χ^2). Group means for Age, and Cohen's d effect size were computed with the bootstrapping method.

Factors	Depressed group ($N = 14$)	Non-depressed group ($N = 40$)	P	Effect size Cohen's d or Phi
Age, M	38.51	44.39	<0.001	−0.53
Gender			1(χ^2)	0.00 (phi)
Female, n(%) Male, n(%)	8 (57.14) 6 (42.86)	22 (55) 18 (45)		

4. Results

4.1. Participants

Of $N = 54$ participants (mean age 43.04 ± 11.58, range 24–68), 30 (55.6%) were females and 24 (44.4%) were males. Participant groups included 14 participants (8 females, 6 males) in the depressed group (see Section 3.2) and 40 participants (22 females, 18 males) in the non-depressed group. Table 3 details the participant demographics. When participant groups were compared, we found a statistically significant difference in age ($p < 0.001$, $d = -0.53$). There was no statistically significant difference in gender.

4.2. Statistical difference between participant groups

The digital biomarkers and mood ratings were not normally distributed (Shapiro–Wilk test: $p < 0.05$). Table 4 details the statistically significant difference and effect sizes of digital biomarkers when comparing participant groups. Table 4 only presents digital biomarkers that showed; (1) significance difference ($p < 0.05$) between depressed and non-depressed groups in the two-sided Mann–Whitney's U test, (2) non-overlapping confidence intervals of the depressed and non-depressed group means, and (3) small to high effect size (absolute cohen's d >0.2). Extended results of the Shapiro–Wilk test for normality, the Two-sided Mann–Whitney's U significant difference test is shown in Supplementary Table 1. In addition, the extended results of the Mean difference and Effect Sizes in biomarkers between the Depressed group and the Non-Depressed group are shown in Supplementary Table 2.

Of 49 digital biomarkers, GPS mobility biomarkers showed the largest effect size, with location routine index having $d = 0.66$ 95%CI[0.550, 0.766], followed by physical activity, with step count having $d = -0.486$ 95%CI[−0.589, −0.383]. The depressed group showed statistically significantly less mobility than participants in the non-depressed group. Particularly, the depressed group tends to have lower location entropy (i.e. location entropy, $M_{Diff} = -0.237$, $p < 0.001$, $d = -0.504$ 95%CI[−0.617, −0.391]), visits fewer different locations (i.e. numberofsignificantplaces, $M_{Diff} = -2.861$, $p < 0.001$, $d = -0.486$ 95%CI[−0.582, −0.389]) and spends more time at single location clusters (i.e. meanlengthstayatclusters, $M_{Diff} = 128.406$, $p < 0.001$, $d = 0.439$ 95%CI[0.310, 0.572]). For mood ratings, the depressed group showed statistically significantly lower Valence ($M_{Diff} = -1.172$, $p < 0.001$, $d = -0.803$ 95%CI[−0.941, −0.669]) and significantly lower Arousal ($M_{Diff} = -0.867$, $p < 0.001$, $d = -0.488$ 95%CI[−0.609, −0.369]). In addition, the depressed group showed statistically significantly more total sleep time ($M_{Diff} = 916.377$, $p = 0.039$, $d = 0.218$ 95%CI[0.089, 0.345]) and time in bed ($M_{Diff} = 1424.033$, $p < 0.001$, $d = 0.296$ 95%CI[0.169, 0.422]). The depressed group also showed statistically significantly less physical activity (Table 4).

4.3. Predicting participants' depression status

The predictive performance of ML classifiers trained with demographics, mood ratings and digital biomarkers and the performance of the three baseline models are summarised in Table 5. Only the XGB and SVM classifiers outperformed all three baseline models. The XGB was the best performing classifier. XGB could predict whether a participant belongs to the group with symptoms of depression at 81.43% accuracy (AUC=82.31%, Precision1 = 69.97%, Recall1 = 50.49%, F11 = 58.66%, Precision0 = 84.09%, Recall0 = 92.35%, F10 = 88.02%). Likewise, SVM could predict whether a participant belongs to the group with symptoms of depression at an 75.90% accuracy (AUC=74.89%, Precision1 = 69.97%, Recall1 = 48.77%, F11 = 51.36%, Precision0 = 82.54%, Recall0 = 85.48%, F10 = 83.98%).

The predictive performance of ML classifiers trained with demographics and digital biomarkers only (without mood ratings) are also summarised in Table 5. Only the XGB and SVM classifiers outperformed all three baseline models. Compared with ML classifier performances when mood ratings were included as features, the SVM classifier obtained marginal improvements in performance while the XGB obtained decreases in performance (AUC = −1.6%, Precision1 = −5.68%, Recall1 = −5.68%, F11 = −4.66%). However, the XGB was still the best performing classifier. Feature importance analysis showed that while valence was the 12th most important feature in SVM, valence was the 8th most

Table 4
Mean difference in biomarkers between Depressed group (DG) and Non-Depressed group (NDG). Two-sided Mann–Whitney test showed $p < 0.001$ for all biomarkers except TST ($p = 0.039$).

Biomarkers	Depressed group M_{DG} [95% CI]	Non-depressed group M_{NDG} [95% CI]	Mean difference M_{Diff}	Effect size Cohen's d [95% CI]
GPS location				
locationroutineindex	6.086 [5.142, 7.100]	2.454 [2.325, 2.589]	3.632	0.66 [0.550, 0.766]
loglocationvariance	−5.628 [-5.839, −5.416]	−4.552 [-4.657, −4.446]	−1.076	−0.561 [−0.690, −0.434]
location entropy	0.338 [0.296, 0.382]	0.575 [0.548, 0.603]	−0.237	−0.504 [−0.617, −0.391]
normalizedlocationentropy	0.176 [0.157, 0.197]	0.283 [0.270, 0.295]	−0.106	−0.498 [−0.613, −0.384]
numberofsignificantplaces	5.07 [4.615, 5.539]	7.931 [7.575, 8.296]	−2.861	−0.486 [−0.582, −0.389]
meanlengthstayatclusters	282.678 [248.453, 318.123]	154.272 [139.221, 169.933]	128.406	0.439 [0.310, 0.572]
numberlocationtransitions	11.367 [9.930, 12.919]	19.325 [18.124, 20.575]	−7.957	−0.402 [−0.495, −0.305]
maxlengthstayatclusters	600.891 [572.273, 629.632]	496.931 [482.685, 511.303]	103.96	0.399 [0.275, 0.526]
minlengthstayatclusters	202.942 [165.803, 241.514]	88.508 [72.897, 104.889]	114.434	0.37 [0.240, 0.503]
outlierstimepercent	0.019 [0.016, 0.022]	0.031 [0.029, 0.033]	−0.011	−0.343 [−0.445, −0.237]
Mood				
valence	0.931 [0.762, 1.098]	2.103 [2.019, 2.186]	−1.172	−0.803 [−0.941, −0.669]
arousal	0.204 [0.027, 0.381]	1.071 [0.963, 1.179]	−0.867	−0.488 [−0.609, −0.369]
Sleep				
TIB	30810.06 [30257.282, 31381.813]	29386.026 [29102.752, 29668.928]	1424.033	0.296 [0.169, 0.422]
WASO	4257.23 [4034.579, 4487.724]	3749.391 [3609.507, 3892.737]	507.838	0.226 [0.107, 0.348]
TST	26552.725 [26062.355, 27052.295]	25636.348 [25389.039, 25882.485]	916.377	0.218 [0.089, 0.345]
Physical activity				
step_count	8344.938 [7886.364, 8813.376]	11090.244 [10723.914, 11460.563]	−2745.306	−0.486 [−0.589, −0.383]
MET	1.454 [1.435, 1.473]	1.568 [1.552, 1.585]	−0.115	−0.462 [−0.555, −0.368]
Phone usage				
screen_firstuseafter	13594.697 [12157.786, 15058.356]	18173.29 [17387.954, 18961.385]	−4578.593	−0.329 [−0.451, −0.208]

Table 5
Performance of population models classifying depression status. The model performances are compared with three baseline classifiers (MC = Majority Class, RWC = Random Weighted Classifier, DT = Decision Tree).

Models	Accuracy %	AUC %	FI macro %	Precision1 %	Recall1 %	F11 %	Precision0 %	Recall0 %	F10 %
Baseline1: MC	74.07	50.00	42.55	0.00	0.00	0.00	74.07	100.00	85.11
Baseline2: DT	59.26	46.96	46.96	21.43	21.43	21.43	72.50	72.50	72.50
Baseline3: RWC	61.68	50.13	49.88	26.07	26.14	25.75	74.15	74.12	74.01
Models with Demographics, Digital biomarkers and Mood ratings as features									
Logistic regression	64.91	67.26	60.30	38.71	59.11	46.78	82.26	66.96	73.83
Random forest	70.82	68.08	62.11	44.06	43.84	43.95	80.21	80.35	80.28
K-nearest neighbours	68.25	69.35	64.12	42.93	65.76	51.95	85.12	69.13	76.30
Support vector machine	75.90	74.89	67.67	54.25	48.77	51.36	82.54	85.48	83.98
XGBoost	81.43	82.31	73.34	69.97	50.49	58.66	84.09	92.35	88.02
Models with Demographics and Digital biomarkers features									
Logistic regression	63.50	65.74	58.58	36.81	55.67	44.31	80.89	66.26	72.85
Random forest	69.60	67.07	61.41	42.26	45.07	43.62	80.14	78.26	79.19
K-nearest neighbours	69.22	64.80	63.25	43.02	55.42	48.44	82.48	74.09	78.06
Support vector machine	77.06	74.25	68.67	57.10	48.52	52.46	82.74	87.13	84.88
XGBoost	79.31	80.71	70.33	64.29	46.55	54.00	82.81	90.87	86.65

important feature in XGB and hence the changes in performance when mood ratings were removed from the features (See Supplementary Figure 2).

The twenty (20) most important biomarkers for XGB and SVM classifiers in the models with no mood ratings are illustrated in Fig. 1. The most important biomarkers also included phone usage (sum, average, standard deviation of screen unlock duration, and count of screen unlocks), GPS mobility (mean length of stay at significant places, number of significant places, location routine index, and normalised location entropy), and Sleep (TST, SE). Fig. 1 lists the digital biomarkers on the y-axis in descending order of importance on the model's output. Each dot represents the SHAP value of one participant's biomarker, with blue and red colours representing low and high values of that biomarker.

Worthy of note is that the interpretation of digital biomarker importance and impact as illustrated with the SHAP values does not denote causal relationships between the digital biomarkers and the classifier output. Furthermore, Fig. 1 only interprets the models at the population level, and the interpretation when looking at individual participants may differ. Finally, feature importance is dependent on the machine learning classifier [50]. An important feature for SVM

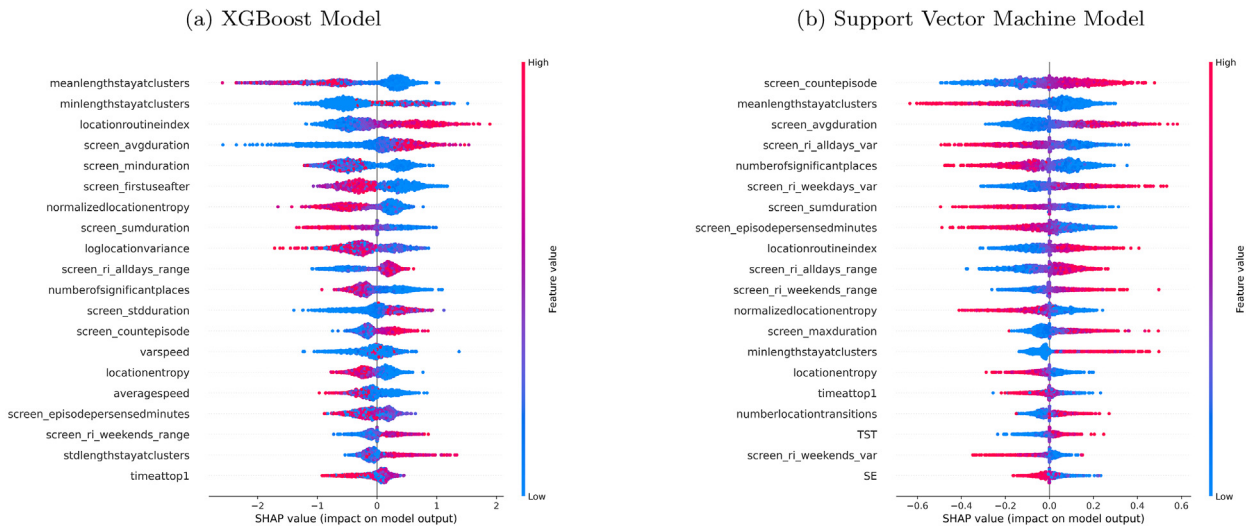

Fig. 1. Density scatter plot of SHAP values for XGBoost (a) and SVM (b) models, illustrating biomarker importance and impact on the model output. The biomarkers are listed in descending order of importance.

may be unimportant or less important in XGBoost. As seen in Fig. 1 the important digital biomarkers and the order of importance differs for both SVM and XGBoost. Therefore, it is recommended to use multiple explanation techniques, in our case multiple classifiers for global interpretation and local interpretation of individual classification, to enhance the overall trust in digital biomarker importance in classification [50].

For instance, SHAP dependence plots in Fig. 2 reviewed interesting interactions between important digital biomarkers, which further enhances the understanding of the impact of the biomarkers on the model's output. Each subplot in Fig. 2 is a scatter plot that shows the biomarker on the X-axis and the SHAP values on the Y-axis. The red and blue colour corresponds to a second biomarker that interacts with the X-axis biomarker. In Fig. 2(a), the number of significant places has a negative correlation with the model's output. For most participants, a lower number of significant places and increased average length of stay in significant places increased the likelihood for the classifiers to classify the participant under the group with symptoms of depression.

Location routine index, in Fig. 2(b), has a positive correlation with the model's output. Participants who usually stay at the same significant location, at the same time of day, across several days (high location routine index) with high location entropy are more likely to be classified under the depressed group. Similarly, Normalised location entropy in Fig. 2(c), has a negative correlation with the model's output. Participants who, on average, spend time at different locations within a day (high normalised location entropy) with a high location routine index are more likely to be classified as non-depressed.

Similar interactions between biomarkers were also observed in phone usage. For example, in Fig. 2(d), an increased number of screen unlocks per day with a lower average unlock duration per unlock increases the likelihood of the participant being classified as depressed. Conversely, participants who unlock their phones less in a day and spend more screen time per unlock were more likely to be classified as non-depressed.

5. Discussion

This paper analysed a longitudinal dataset with statistical and machine learning methods to investigate the relationship between digital biomarkers, mood ratings, and depression. We investigated whether digital biomarkers and mood ratings are different between depressed and non-depressed participants. To reflect on the research questions (see the end of the Introduction section), we found statistically significant differences in mood ratings and digital biomarkers of sleep, physical activity, phone usage and GPS mobility when depressed and non-depressed participants are compared (see Statistical difference between participant groups and Table 4). We show that it is possible to accurately predict the depression status of participants using only digital biomarkers as predictors or using both digital biomarkers and mood ratings as predictors. (see Predicting participants' depression status and Table 5). Finally, the most important digital biomarkers in differentiating depressed and non-depressed participants were related to phone usage, GPS mobility, and sleep.

Our findings are consistent with previous research on the relationship between digital biomarkers and depression (see Table 1). Previous work found statistically significant differences in mood and digital biomarkers when comparing depressed and non-depressed participants. More specifically, people with symptoms of depression showed significantly lower mood [35,37,38], reduced physical activity [35,51], longer sleep time [16], reduced location mobility [7] and increased phone usage [3,7] compared to people without symptoms of depression. In our findings, the magnitude of significant differences (effect size) in digital biomarkers between the depressed and non-depressed participants are

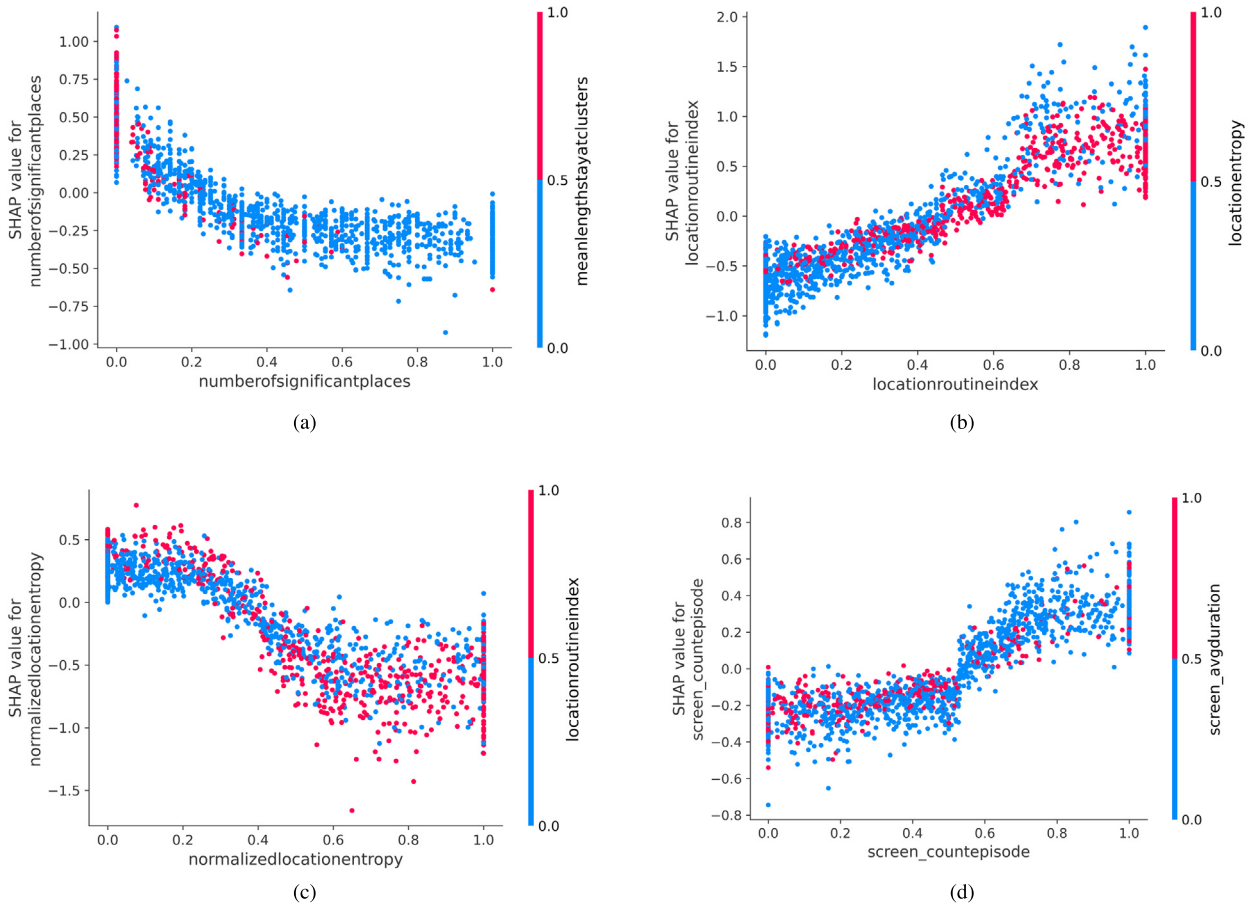

Fig. 2. SHAP dependence plots showing interactions between important biomarkers of the XGBoost model.

interpreted as small to medium effect sizes (see Section 4.2). In line with our findings, medium effect sizes of significant differences in physical activity biomarkers have been reported in [51]. Our results suggest that digital biomarkers quantified from passively sensed smartphone and wearable data could be used to differentiate between the depressed and non-depressed to support current clinical care by recognising and monitoring depression without relying on the patient's ability to recall their behaviour and mood. Our findings are in line with previous research on the potential of using Machine Learning models to predict depression with digital biomarkers and mood ratings (see Table 2). The best performing classifier could predict the depression status at 81.43% accuracy and AUC of 82.31% when digital biomarkers and mood ratings are used as predictors. Alternatively, using digital biomarkers alone, the best performing classifier could predict depression status at 79.31% accuracy and AUC of 80.71% . The ability to predict depression status with only digital biomarkers offer a better user experience because self-reported mood ratings may pose a considerable data collection burden on participants, and that affects participant retention and increases study drop-out rate [17]. Our findings mirror the results of previous studies on predicting depression status with a reported accuracy of 72.7% [42], 81.8% [16], and AUC 80.9% [3] and 75.4% [35]. These findings suggest that digital biomarkers quantified from passively sensed smartphone and wearable datasets could be used to predict depression status to augment the current depression diagnosis methods.

The findings also demonstrate that it is feasible to understand which digital biomarkers influence the participant's depression status. For instance, our findings suggest that phone usage, GPS mobility, and sleep digital biomarkers were among the most important in predicting depression status. Phone usage, GPS mobility were reported among the important digital biomarkers associated with brain functionalities known to be associated with depression [52]. Phone usage has been associated with depression in previous studies [3,21,32,39]. Likewise, GPS mobility has also been associated with depression [7,29,32]. These findings could be useful in clinical implementations to augment current diagnosis and monitoring of depression symptom trajectories, and course of treatment [22,23]. For instance, knowing which digital biomarkers influence a person's depression status could help clinicians personalise interventions for improving a patient's mental health, for example, with actionable goals like improving physical activity and step count per day.

We acknowledge several limitations in the study. First, the participants in our dataset ($N = 54$, average 28.21 days of data per participant) are from a non-clinical population, with an unequal distribution of gender, depressed and non-depressed groups. However, about 25.9% of the participants belonged to the depressed group. A systematic analysis [5] of the disease burden in 195 countries from the year 1990 to 2017 showed that there is a high prevalence of depression even in general populations. Moreover, the participants were recruited in 2019 during the first wave of the coronavirus disease (COVID-19) [15]. The COVID-19 is known to have contributed to an increase in the prevalence rate of mental health conditions in general populations [53].

Second, we focused on building population-based models given the limited sample size. Future work could replicate this study with a clinical population and a balanced sample in terms of gender, depressed and non-depressed groups to investigate whether the results of this study generalises in both population and individual models. The dynamics in the relationship between digital biomarkers and depression for individuals and similar participant subgroups based on gender, age group, or personality trait would be interesting to investigate, leading to a deeper understanding for personalised mental health interventions. It would be relevant to understand how different digital biomarkers, combinations of digital biomarkers, and the number of monitored days of these digital biomarkers differ in predictive power for depression status across age groups, gender, level of education, personality traits and time.

Third, the 21-item DASS scale used in assessing the depression symptoms of our participants is not a direct diagnostic tool compared to the Patient Health Questionnaire (PHQ) [11,54]. However, the DASS scale is predominantly used as a screening tool for accessing the severity of Depression, Anxiety and Stress in clinical settings [54,55]. We used the established cut-off subscales of the DASS-21 proposed by Lovibond and Lovibond [10] for categorising the participants into depression groups. The depressed and non-depressed groups are not clinical diagnoses of depression but participants with or without symptoms of depression. While these categorisations are not a clinical diagnosis of depression, studies have established a high correlation between the DASS depression subscale and clinical diagnostics of depression with PHQ [54].

The strengths of this study include the use of longitudinal and passively collected smartphone and wearable data, in addition to daily mood ratings in a naturalistic setting outside the confinements of a laboratory. We quantified digital biomarkers of sleep, physical activity, phone usage and GPS mobility from our longitudinal dataset. We utilised robust statistical methods to analyse the statistical differences in participant demographics and digital biomarkers from participants with or without depression. We trained our Machine Learning models with state-of-the-art cross-validation methods (time series awareness, within-fold imputation, scaling and feature selection). We use feature importance analysis to understand the importance and impact of digital biomarkers on the model predictions.

This study demonstrates that digital biomarkers and mood ratings could be useful in differentiating and predicting the depression status of individuals. The current study builds upon previous research and contributes to the compelling evidence that digital biomarkers and mood ratings could be continuously collected and monitored to understand the complex vectors of depression and augment the traditional depression assessment method for effective diagnosis and monitoring of depression.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the Academy of Finland [Grants 316253 - SENSATE, 320089-SENSATE] and the Infotech Institute at the University of Oulu, Finland. KOA was supported by the Nokia Foundation, Finland. LP-R was supported by The Jenny and Antti Wihuri Foundation and the Yrjö Jahnsson Foundation, Finland. IM was supported by the Finnish Foundation for Psychiatric Research, Finland.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.pmcj.2022.101621.

References

[1] WHO, World health organisation | depression, 2020, URL https://www.who.int/news-room/fact-sheets/detail/depression.
[2] H. König, H.-H. König, A. Konnopka, The excess costs of depression: a systematic review and meta-analysis, Epidemiol. Psychiatr. Sci. 29 (2019) e30, http://dx.doi.org/10.1017/S2045796019000180.
[3] R. Wang, W. Wang, A. daSilva, J.F. Huckins, W.M. Kelley, T.F. Heatherton, A.T. Campbell, Tracking depression dynamics in college students using mobile phone and wearable sensing, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2 (1) (2018) 43, http://dx.doi.org/10.1145/3191775.
[4] H. Baumeister, N. Hutter, J. Bengel, M. Härter, Quality of life in medically ill persons with comorbid mental disorders: a systematic review and meta-analysis, Psychother. Psychosom. 80 (5) (2011) 275–286.

[5] S.L. James, D. Abate, K.H. Abate, S.M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, et al., Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017, Lancet 392 (10159) (2018) 1789–1858, http://dx.doi.org/10.1016/S0140-6736(18)32279-7.

[6] S. Saeb, M. Zhang, M. Kwasny, C.J. Karr, K. Kording, D.C. Mohr, The relationship between clinical, momentary, and sensor-based assessment of depression, in: 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2015, pp. 229–232, http://dx.doi.org/10.4108/icst.pervasivehealth.2015.259034.

[7] S. Saeb, M. Zhang, C.J. Karr, S.M. Schueller, M.E. Corden, K.P. Kording, D.C. Mohr, Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study, J. Med. Internet Res. 17 (7) (2015) e175, http://dx.doi.org/10.2196/jmir.4273.

[8] L.H. Andrade, J. Alonso, Z. Mneimneh, J. Wells, A. Al-Hamzawi, G. Borges, E. Bromet, R. Bruffaerts, G. De Girolamo, R. De Graaf, et al., Barriers to mental health treatment: results from the WHO world mental health surveys, Psychol. Med. 44 (6) (2014) 1303–1317.

[9] A.T. Beck, R.A. Steer, G. Brown, Beck depression inventory–II, Psychol. Assess. (1996) http://dx.doi.org/10.1037/t00742-000.

[10] P.F. Lovibond, S.H. Lovibond, The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories, Behav. Res. Ther. 33 (3) (1995) 335–343, http://dx.doi.org/10.1016/0005-7967(94)00075-u.

[11] K. Kroenke, R.L. Spitzer, J.B.W. Williams, B. Löwe, The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review, Gen. Hosp. Psychiatry 32 (4) (2010) 345–359, http://dx.doi.org/10.1016/j.genhosppsych.2010.03.006.

[12] R. Hartmann, F.M. Schmidt, C. Sander, U. Hegerl, Heart rate variability as indicator of clinical state in depression, Front. Psychiatry 9 (2019) http://dx.doi.org/10.3389/fpsyt.2018.00735.

[13] R.C. Young, J.T. Biggs, V.E. Ziegler, D.A. Meyer, A rating scale for mania: reliability, validity and sensitivity, Br. J. Psychiatry: J. Ment. Sci. 133 (1978) 429–435, http://dx.doi.org/10.1192/bjp.133.5.429.

[14] P. Chikersal, A. Doryab, M. Tumminia, D.K. Villalba, J.M. Dutcher, X. Liu, S. Cohen, K.G. Creswell, J. Mankoff, J.D. Creswell, et al., Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection, ACM Trans. Computer-Human Interact. 28 (1) (2021) 3:1–3:41, http://dx.doi.org/10.1145/3422821.

[15] I. Moshe, Y. Terhorst, K. Opoku Asare, L.B. Sander, D. Ferreira, H. Baumeister, D. Mohr, L. Pulkki-Råback, Predicting symptoms of depression and anxiety using smartphone and wearable data, Front. Psychiatry 12 (2021) http://dx.doi.org/10.3389/fpsyt.2021.625247.

[16] X. Xu, P. Chikersal, A. Doryab, D.K. Villalba, J.M. Dutcher, M.J. Tumminia, T. Althoff, S. Cohen, K.G. Creswell, J.D. Creswell, et al., Leveraging routine behavior and contextually-filtered features for depression detection among college students, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3 (3) (2019) 116:1–116:33, http://dx.doi.org/10.1145/3351274.

[17] N. van Berkel, D. Ferreira, V. Kostakos, The experience sampling method on mobile devices, ACM Comput. Surv. 50 (6) (2017) 93:1–93:40, http://dx.doi.org/10.1145/3123988.

[18] L. Canzian, M. Musolesi, Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2015, pp. 1293–1304, http://dx.doi.org/10.1145/2750858.2805845.

[19] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K.J. Oedegaard, J. Tørresen, Mental health monitoring with multimodal sensing and machine learning: A survey, Pervasive Mob. Comput. 51 (2018) 1–26, http://dx.doi.org/10.1016/j.pmcj.2018.09.003.

[20] J. Torous, M.V. Kiang, J. Lorme, J.-P. Onnela, New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research, JMIR Ment. Health 3 (2) (2016) e5165, http://dx.doi.org/10.2196/mental.5165.

[21] P. Pedrelli, S. Fedor, A. Ghandeharioun, E. Howe, D.F. Ionescu, D. Bhathena, L.B. Fisher, C. Cusin, M. Nyer, A. Yeung, L. Sangermano, D. Mischoulon, J.E. Alpert, R.W. Picard, Monitoring changes in depression severity using wearable and mobile sensors, Front. Psychiatry 11 (2020) http://dx.doi.org/10.3389/fpsyt.2020.584711.

[22] L.M. Babrak, J. Menetski, M. Rebhan, G. Nisato, M. Zinggeler, N. Brasier, K. Baerenfaller, T. Brenzikofer, L. Baltzer, C. Vogler, L. Gschwind, C. Schneider, F. Streiff, P.M.A. Groenen, E. Miho, Traditional and digital biomarkers: Two worlds apart? Digit. Biomark. 3 (2) (2019) 92–102, http://dx.doi.org/10.1159/000502000.

[23] A. Coravos, S. Khozin, K.D. Mandl, Developing and adopting safe and effective digital biomarkers to improve patient outcomes, Npj Digit. Med. 2 (11) (2019) 1–5, http://dx.doi.org/10.1038/s41746-019-0090-4.

[24] C. Montag, H. Baumeister, C. Kannen, R. Sariyska, E.-M. Meßner, M. Brand, Concept, possibilities and pilot-testing of a new smartphone application for the social and life sciences to study human behavior including validation data from personality psychology, J 2 (2) (2019) 102–115.

[25] E. Peltonen, E. Lagerspetz, J. Hamberg, A. Mehrotra, M. Musolesi, P. Nurmi, S. Tarkoma, The hidden image of mobile apps: geographic, demographic, and cultural factors in mobile usage, in: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, 2018, pp. 1–12, http://dx.doi.org/10.1145/3229434.3229474.

[26] E. Peltonen, P. Sharmila, K. Opoku Asare, A. Visuri, E. Lagerspetz, D. Ferreira, When phones get personal: Predicting big five personality traits from application usage, Pervasive Mob. Comput. 69 (2020) 101269, http://dx.doi.org/10.1016/j.pmcj.2020.101269.

[27] Y. Nishiyama, D. Ferreira, Y. Eigen, W. Sasaki, T. Okoshi, J. Nakazawa, iOS crowd-sensing won't hurt a bit!: AWARE framework and sustainable study guideline for iOS platform, in: International Conference on Human-Computer Interaction, Springer, 2020, pp. 223–243, http://dx.doi.org/10.1007/978-3-030-50344-4_17.

[28] J. Vega, M. Li, K. Aguillera, N. Goel, E. Joshi, K. Khandekar, K.C. Durica, A.R. Kunta, C.A. Low, Reproducible analysis pipeline for data streams: Open-source software to process data collected with mobile devices, Front. Digit. Health 3 (2021) doi: 10/hkrc.

[29] S.R. Müller, H. Peters, S.C. Matz, W. Wang, G.M. Harari, Investigating the relationships between mobility behaviours and indicators of subjective well–being using smartphone–based experience sampling and GPS tracking, Eur. J. Pers. 34 (5) (2020) 714–732, http://dx.doi.org/10.1002/per.2262.

[30] I. Kobayashi, J.M. Boarts, D.L. Delahanty, Polysomnographically measured sleep abnormalities in PTSD: A meta-analytic review, Psychophysiology 44 (4) (2007) 660–669, http://dx.doi.org/10.1111/j.1469-8986.2007.537.x.

[31] A. Gershon, W.K. Thompson, P. Eidelman, E.L. McGlinchey, K.A. Kaplan, A.G. Harvey, Restless pillow, ruffled mind: Sleep and affect coupling in interepisode bipolar disorder, J. Abnorm. Psychol. 121 (4) (2012) 863–873, http://dx.doi.org/10.1037/a0028233.

[32] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, A.T. Campbell, Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14, Association for Computing Machinery, 2014, pp. 3–14, http://dx.doi.org/10.1145/2632048.2632054.

[33] R.-E. Mastoras, D. Iakovakis, S. Hadjidimitriou, V. Charisis, S. Kassie, T. Alsaadi, A. Khandoker, L.J. Hadjileontiadis, Touchscreen typing pattern analysis for remote detection of the depressive tendency, Sci. Rep. 9 (1) (2019) 1–12.

[34] J. Zulueta, A. Piscitello, M. Rasic, R. Easter, P. Babu, S.A. Langenecker, M. McInnis, O. Ajilore, P.C. Nelson, K. Ryan, et al., Predicting mood disturbance severity with mobile phone keystroke metadata: A BiAffect digital phenotyping study, J. Med. Internet Res. 20 (7) (2018) e241, http://dx.doi.org/10.2196/jmir.9775.

[35] H. Kim, S. Lee, S. Lee, S. Hong, H. Kang, N. Kim, Depression prediction by using ecological momentary assessment, actiwatch data, and machine learning: Observational study on older adults living alone, JMIR MHealth and UHealth 7 (10) (2019) e14149, http://dx.doi.org/10.2196/14149.
[36] N.C. Jacobson, Y.J. Chung, Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones, Sensors 20 (12) (2020) 3572, http://dx.doi.org/10.3390/s20123572.
[37] D. Rozgonjuk, J.C. Levine, B.J. Hall, J.D. Elhai, The association between problematic smartphone use, depression and anxiety symptom severity, and objectively measured smartphone use over one week, Comput. Hum. Behav. 87 (2018) 10–17, http://dx.doi.org/10.1016/j.chb.2018.05.019.
[38] R. Bowen, E. Peters, S. Marwaha, M. Baetz, L. Balbuena, Moods in clinical depression are more unstable than severe normal sadness, Front. Psychiatry 8 (2017) http://dx.doi.org/10.3389/fpsyt.2017.00056.
[39] K. Opoku Asare, Y. Terhorst, J. Vega, E. Peltonen, E. Lagerspetz, D. Ferreira, Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: Exploratory study, JMIR MHealth and UHealth 9 (7) (2021) e26540, http://dx.doi.org/10.2196/26540.
[40] M. Hamilton, A rating scale for depression, J. Neurol. Neurosurg. I Psychiatry 23 (1) (1960) 56–62, http://dx.doi.org/10.1136/jnnp.23.1.56.
[41] A.A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, B. Wang, Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data, in: 2016 IEEE Wireless Health (WH), IEEE, 2016, pp. 1–8, http://dx.doi.org/10.1109/WH.2016.7764553.
[42] N.C. Jacobson, H. Weingarden, S. Wilhelm, Digital biomarkers of mood disorders and symptom change, Npj Digit. Med. 2 (1) (2019) 1–3, doi: 10/gjgc2g.
[43] J.A. Russell, A circumplex model of affect, J. Personal. Soc. Psychol. 39 (6) (1980) 1161–1178, http://dx.doi.org/10.1037/h0077714.
[44] L. Canzian, M. Musolesi, Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2015, pp. 1293–1304.
[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (85) (2011) 2825–2830.
[46] C.A. Low, M. Li, J. Vega, K.C. Durica, D. Ferreira, V. Tam, M. Hogg, H.Z. Iii, A. Doryab, A.K. Dey, Digital biomarkers of symptom burden self-reported by perioperative patients undergoing pancreatic surgery: Prospective longitudinal study, JMIR Cancer 7 (2) (2021) e27975, http://dx.doi.org/10.2196/27975.
[47] J. Poulos, R. Valle, Missing data imputation for supervised learning, Appl. Artif. Intell. 32 (2) (2018) 186–196, http://dx.doi.org/10.1080/08839514.2018.1448143.
[48] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (11) (2020) 56–67, http://dx.doi.org/10.1038/s42256-019-0138-9.
[49] TENK, Guidelines for ethical review in human sciences, 2020, URL https://tenk.fi/en/advice-and-materials/guidelines-ethical-review-human-sciences,
[50] M. Saarela, S. Jauhiainen, Comparison of feature importance measures as explanations for classification models, SN Appl. Sci. 3 (2) (2021) 272, http://dx.doi.org/10.1007/s42452-021-04148-9.
[51] S.V. George, Y.K. Kunkels, S. Booij, M. Wichers, Uncovering complexity details in actigraphy patterns to differentiate the depressed from the non-depressed, Sci. Rep. 11 (1) (2021) 13447, http://dx.doi.org/10.1038/s41598-021-92890-w.
[52] M. Obuchi, J.F. Huckins, W. Wang, A. daSilva, C. Rogers, E. Murphy, E. Hedlund, P. Holtzheimer, S. Mirjafari, A. Campbell, Predicting brain functional connectivity using mobile sensing, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4 (1) (2020) 1–22, http://dx.doi.org/10.1145/3381001.
[53] J. Xiong, O. Lipsitz, F. Nasri, L.M.W. Lui, H. Gill, L. Phan, D. Chen-Li, M. Iacobucci, R. Ho, A. Majeed, R.S. McIntyre, Impact of COVID-19 pandemic on mental health in the general population: A systematic review, J. Affect. Disord. 277 (2020) 55–64, http://dx.doi.org/10.1016/j.jad.2020.08.001.
[54] L. Peters, A. Peters, E. Andreopoulos, N. Pollock, R.L. Pande, H. Mochari-Greenberger, Comparison of DASS-21, PHQ-8, and GAD-7 in a virtual behavioral health care setting, Heliyon 7 (3) (2021) e06473, http://dx.doi.org/10.1016/j.heliyon.2021.e06473.
[55] I.N. Beaufort, G.H. De Weert-Van Oene, V.A. Buwalda, J.R.J. de Leeuw, A.E. Goudriaan, The depression, anxiety and stress scale (DASS-21) as a screener for depression in substance use disorder inpatients: a pilot study, Eur. Addict. Res. 23 (5) (2017) 260–268, http://dx.doi.org/10.1159/000485182.