Invited Paper

Smartphone-based Mental State Estimation: A Survey from a Machine Learning Perspective

Yusuke Fukazawa[1,a)] Naoki Yamamoto[1] Takashi Hamatani[1] Keiichi Ochiai[1]
Akira Uchiyama[2] Ken Ohta[1]

Received: May 9, 2019, Accepted: October 24, 2019

Abstract: Monitoring mental health has received considerable attention as a countermeasure against the increasing occurrence of mental illness worldwide. However, current monitoring services incur costs because users are required to attach wearable devices or answer questions. To reduce such costs, many studies have used smartphone-based passive sensing technology to capture a user's mental state. This paper reviews those studies from the perspective of machine learning and statistical analysis. Forty-four studies published since 2011 have been reviewed and summarized from three perspectives: designed features, machine learning algorithm, and evaluation method. The features considered include location and mobility, activity, speech, sleep, phone usage, and context features. Tasks are classified as correlation analysis, regression tasks, and classification tasks. The machine learning algorithm used for each task is summarized. Evaluation metrics and cross validation methods are also summarized. For those who are not necessarily machine learning experts, we aim to provide information on typical machine learning framework for smartphone-based mental state estimation. For experts in the field, we hope this review will be a helpful tool to check for potential omissions.

Keywords: smartphone, sensing, mental health, machine learning

1. Introduction

The growing number of mental disorders and mental illnesses has become a global social issue, particularly for students [1] and workers [2], [3], [4] who are continually pressured to perform rationally and efficiently. The World Health Organization (WHO) reported that greater than 90% of suicides are due to mental disorders [5]. The WHO considers mental illness to be a global challenge and an economic problem [4]. In Japan, the importance of worker mental health has been recognized, and the Japanese government has sponsored some preventive systems, such as a self-reporting stress check program [6]. Considering the widespread use of smartphones [7], there is growing interest in the use of such devices to monitor a user's mental state and promote self-care [8]. The WHO's Mental Health Action Plan 2013–2020 [9] recommended "the promotion of self-care, for instance, through the use of electronic and mobile health technologies."

Technology-assisted mental health monitoring can be categorized into the following three types. The first approach relies on specialized or wearable devices to measure physiological signals [10], [11], [12]. Some studies have shown that anxiety or stress can be estimated by measuring amylase [13], cortisol in saliva [13], [14], [15], blood pressure [16], [17], heart rate[17], heart rate variability [17], [18], [19], nitric oxide (NO)-related signals [20], and skin conductance [21], [22]. Estimations that rely on physiological indices are reliable; however, consistent daily measurement is impractical as users must attach wearable devices to their bodies continuously. The second approach relies on self-assessment using smartphone applications [23], [24], [25]; however, frequent self-assessment is difficult to maintain because it requires time and effort. The last approach is to use the passive sensing features of smartphones to detect mental states [8], [26]. Sensing via a smartphone is less intrusive than specialized wearable devices [27]. Due to increasing computational power and pervasiveness, most smartphones are equipped with multiple sensors that can capture complex and meaningful behavioral features. Smartphones can also capture various operational data, such as call, text, and application activity [28]. In this paper, we investigate and summarize studies related to sensing via smartphones.

The research challenge of using smartphones' passive sensing features to detect mental states is to minimize errors between the correct mental state and the estimated state. Numerous approaches, such as designing mental health related features using smartphone data and selecting machine learning algorithms that improve accuracy, have been proposed. There are two notable survey papers [8], [29] in this research area. In these papers, studies were categorized by mental health condition (e.g., stress, anxiety) or mental illness (e.g., bipolar disorder, depression, schizophrenia) and the sensor used (e.g., GPS [*1], call logs, accelerometer). However, these survey papers did not focus on statistical analysis or machine learning models. We found that past studies take totally different approaches to feature design, se-

[1] NTT DOCOMO, Inc., Chiyoda, Tokyo 100–6150, Japan
[2] Osaka University, Suita, Osaka 565–0871, Japan
[a)] fukazawayu@nttdocomo.com

[*1] Global Positioning System.

lection of machine learning algorithm, and evaluation setting. To learn and apply best practices, in this paper, we summarize previous studies from the perspective of a machine learning framework.

In the next section, we describe a general machine learning framework for smartphone-based mental health estimation.

2. Machine Learning Framework

In this section, we describe the overall procedure of smartphone-based mental state estimation. The purpose of this procedure is to learn the complex relationships among features that can be captured by smartphone sensors and the mental state. As a large variety of features can be considered, a machine learning framework is widely used to learn the relationships efficiently. **Figure 1** shows a general machine learning framework in smartphone-based mental state estimation. The framework consists of five steps. First, we collect raw smartphone data for each participant. Second, we collect the state of mental illness/mental health for each participant as their ground truth via self-reported questionnaires or conducting psychological scale-tests performed by a clinical psychologist. Third, we design features based on a heuristic assumption about the relationship between mental state and human behavior and then calculate the feature values from the raw data collected from each participant's smartphone. Fourth, we learn the complex relationships among features and the mental state through one of three tasks; classification of mental state (e.g., stressed or not stressed), regression of mental score (e.g., prediction of the self-assessment score), and correlation analysis (e.g., correlation between feature score and self-assessment score). Finally, we design the evaluation setting, e.g., evaluation metrics and cross-validation policy, and conduct the evaluation.

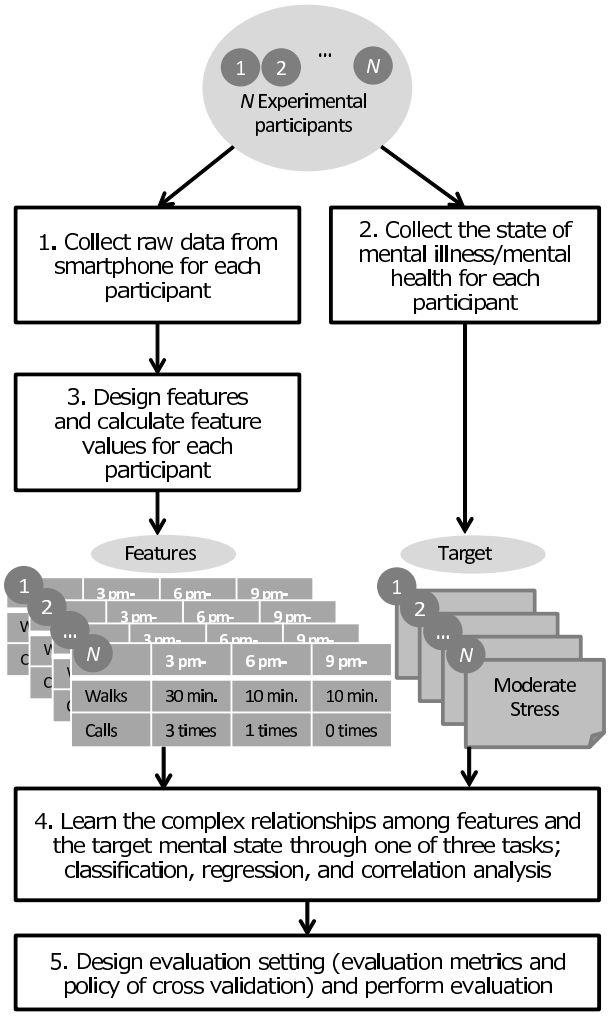

Fig. 1 Machine learning framework for smartphone-based mental state estimation. The framework consists of five steps. Past surveys summarize previous studies from the perspective of the sensor used (step 1) and the type of mental health or mental illness (step 2). This survey summarizes previous studies from the perspective of feature design (step 3), machine learning algorithm selection (step 4), and evaluation setting (step 5) in a machine learning framework.

To thoroughly investigate recent trends in this field, we first selected papers that match the machine learning framework shown in Fig. 1 from those mentioned in the survey papers [8], [29]. We also conducted a web search to collect additional papers using the query "smartphone, mental health, machine learning." In total, we considered 44 papers, which are listed in Tables 6, 7, and 8. In the following sections, we investigate and summarize the previous studies from the perspective of feature design (step 3), machine learning algorithm selection (step 4), and evaluation setting (step 5) in the machine learning framework. Summaries of steps 3, 4, and 5 are provided in Sections 3, 4, and 5, respectively. In Section 6, we discuss recommendations for this research field and limitations. Conclusions are presented in Section 7.

3. Features

3.1 Building Hypothesis

To create smartphone-based features related to mental health, most studies proposed hypotheses about the characteristic behavior taken by people with mental illness based on the assessment metrics used for diagnosis or the known findings. With reference to previous studies [30], [31], Saeb et al. developed a hypothesis about the association of depression with several behaviors, such as reduction in activity, psychomotor retardation, and changes in sleep patterns. According to the hypothesis, they developed a location and mobility feature and a phone usage feature [32]. With reference to past findings [33], [34], Beiwinkel et al. [35] hypothesized that higher levels of physical activity and social communication measured by a smartphone represent lower levels of depressive symptoms and temporal increases in manic symptoms for patients with bipolar disorder. According to the hypothesis, they create location and mobility and phone usage features. With reference to the Patient Health Questionnaire (PHQ) [36], Canzian et al. [37] developed a hypothesis related to behavioral patterns associated with depression, such as reduced mobility [38] and limited willingness to perform different activities. According to the hypothesis, they created location and mobility features. Wang et al. [39] proposed features that are proxies for depression symptoms defined in the DSM-5 [40], such as changes in sleep patterns, diminished ability to concentrate, and diminished interest or pleasure in activities. They developed phone usage, location and mobility, activity, speech, and sleep features. In the following sections, the features developed in previous studies are summarized. Features are categorized into

six types: location and mobility, activity, phone usage, context, speech, and sleep (38 features in total).

3.2 Location and Mobility Feature

Location and mobility features are generated from smartphone GPS and Wi-Fi scan data. **Table 1** lists the location and mobility features.

Feature LOC-1, travel distance, represents the distance between two locations. Many studies used total travel distance per day [32], [41], [42], [43], which is calculated by accumulating the distances between the location samples in a single day. Some studies created conditional total distance, such as approximate distance covered by foot [44] and the distance a student travels inside buildings identified by Wi-Fi scan logs [45], [46]. Canzian et al. used the maximum distance between two significant locations [37].

Next, we describe the features LOC-2, time spent at significant location, and LOC-3, number of places visited feature, which are related to visiting significant locations such as home, campus or office. Some studies identified the home location based on several heuristics, e.g., the home location is the location most visited between 12am and 6am [32], [41], [42], [43], [47], and the home is among the first to the third most visited location clusters [32], [41], [42], [43]. Some studies identified significant locations by applying a clustering algorithm, such as *k*-means, to participants' raw GPS data [32], [41], [42]. Boukhechba et al. identified semantic locations, e.g., restaurants, campus areas, and shops, by combining spatiotemporal clustering results and a map database [48]. After significant locations are identified, we calculate the stay time in each location for feature LOC-2, and count the number of locations for feature LOC-3. For example, Farhan et al. used the amount of time that a participant spends in the top three clusters as the feature LOC-2 [43]. Exler et al. used the distribution of cumulative stay time in each semantic location as the feature LOC-2 [49]. The number of location clusters found by *k*-means is used as the feature LOC-3 [32], [41], [42].

Finally, we explain the remaining location and mobility features listed in Table 1. Feature LOC-4, transition time, represents total time elapsed during travel [50]. Some studies calculated this feature by dividing the number of GPS location samples in transition states by the total number of samples [32], [41], [42], [43]. Feature LOC-5, routine index, represents the extent to which participants' sequence of locations followed a circadian rhythm [32], [41], [42]. If a participant left home for work and returned home from work at approximately the same time each day, the circadian rhythm was high. In contrast, a participant with a more irregular pattern of moving between locations had a lower circadian rhythm. Canzian et al. defined a routine index that represents the extent to which places visited on a particular day differ from those visited on other days [37]. Feature LOC-6, location variance, represents the variability in a participants' GPS location. The sum of the statistical variances of the latitude and longitude components of the location data is used as a feature LOC-6 [32], [41], [42], [43]. Feature LOC-7, the living area size, represents an imaginary circle encompassing the various locations that a user traveled across on a particular day [21]. Canzian et al. proposed the radius of gyration, which outputs the radius of an imaginary circle by weighting the contribution of each place by time spent in that place [37]. Feature LOC-8, speed, represents the mean of the speed obtained at each GPS data point. Here, speed is calculated as the change in latitude and longitude values over time [42].

Table 1 Location and mobility features.

	Feature	Reference
LOC-1	Travel distance	[21], [32], [35], [37], [41], [42], [43], [44], [45], [46], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61]
LOC-2	Time spent at significant location (home, clinic, office, school, etc.)	[32], [39], [41], [42], [43], [47], [48], [49], [53], [55], [56], [58], [59], [60], [61], [62], [63], [64], [65], [66]
LOC-3	Number of places visited	[32], [37], [39], [41], [42], [48], [53], [55], [56], [57], [60], [61], [67]
LOC-4	Transition time	[32], [37], [41], [42], [43], [50], [61], [64], [66]
LOC-5	Routine index	[32], [37], [41], [42], [60]
LOC-6	Location variance	[32], [41], [42], [43], [67]
LOC-7	Size of living area	[21], [37], [60]
LOC-8	Speed	[42], [61]

Table 2 Activity features.

	Feature	Reference
ACT-1	Duration of activity time	[39], [43], [45], [46], [47], [49], [51], [58], [60], [62], [67], [68], [71], [72]
ACT-2	Activity state transition	[62]
ACT-3	Acceleration magnitude	[53], [57], [65], [67], [69], [70]
ACT-4	Acceleration frequency analysis	[43], [55], [56]
ACT-5	Physical orientation	[69]
ACT-6	Walking steps	[70]

3.3 Activity Feature

Activity features are generated using the multi-axial accelerometers embedded in the smartphone. Activity features are listed in **Table 2**.

Feature ACT-1, duration of activity time, represents the duration of each activity. We estimated the types of activities based on data acquired from multi-axial accelerometers. Some previous studies built a physical activity classifier to infer stationary, walking, running, driving, and cycling activities based on features extracted from accelerometer streams [45], [46]. Recently, Google, Inc. has provided an activity recognition service API [*2] to estimate users' physical activities, such as walking, running, staying still, and using vehicles. Some studies identified an activity using recently developed APIs and calculate the total time spent on each activity for each full day [49], [60], [68]. Feature ACT-2, activity state transition, represents the number of times participant changes his activity to the other activity such as still to walking. Feature ACT-3, acceleration magnitude, is calculated by the square root of sum of squares for individual acceleration axes. Doryab et al. used minimum, median, maximum, average,

[*2] Both Google, Inc. and Apple Inc. have started to provide basic activity recognition services, which are implemented as ActivityRecognitionApi on Android OS and CMMotionActivity on iOS.

and SD [*3] values of the acceleration magnitude as features [53]. For feature ACT-4, acceleration frequency analysis, Grünerbl et al. used Fourier transformation to output frequency centroid and frequency fluctuation as features [55], [56]. Feature ACT-5, physical orientation, which is the angle of the display to the ground, has been calculated using smartphone rotation sensor data [69]. Feature ACT-6, walking steps, represents the number of steps per minute and is calculated using the tri-axial accelerometers embedded in a smartphone [70].

3.4 Phone Usage Feature

Phone usage features are generated from the smartphone's user operation log. **Table 3** lists phone usage features.

First, we explain the features related to communication tools, such as calls, SMS [*4] and E-mail. Feature PHO-1, number of communication tool usage, represents the total number of incoming, outgoing, and missed phone calls, and the number of incoming and outgoing text messages (SMSs and E-mails). The mean and SD of the number of communications are also used [52]. Feature PHO-2, duration of calls, represents the total duration of incoming and outgoing calls. Mean, median, and SD of the duration of calls are also used [50], [54]. Asselbergs et al. calculated call duration only for the top five caller contacts [71]. Gjoreski et al. defined call duration deviation by the duration of the longest call in the last two days relative to the average duration of calls of the past [58]. Feature PHO-3, number of people interacted with via communication tools, represents the total number of individuals with whom a participant interacted through calls, SMS, and E-mails.

Next we explain features related to screen on/off events. Feature PHO-4, duration of screen on, represents the sum of each session of screen time. Note that a user session begins when the device screen is powered on and ends if the screen goes off. Mean and SD of screen-on duration is also used [54]. Farhan et al. measured the total duration that a participant's phone is locked in a single day [43]. Feature PHO-5, number of times screen on/off, represents the total number of times the screen was turned on and off per day. As feature PHO-6, time stamp of call, SMS, screen on/off is a continuous value, the timestamp is converted to a discrete feature by splitting several time intervals per day (e.g., 12am–3am, 3am–6am, 6am–9am, 9am–12pm, 12pm–3pm, 3pm–6pm, 6pm–9pm, and 9pm–12am) [52]. Another discretization of timestamp is calculated by mean, SD, and median of the time of each call [21].

Third, we explain features related to smartphone application usage such as feature PHO-7, duration of app usage, and feature PHO-8, number of times each app is launched. As smartphone applications are diverse, to avoid the data sparsity problem, most studies do not use the exact names of executed apps. Instead, apps are categorized into one of several categories. Some studies used Google Play Store categories, such as communication, entertainment, finance, games, office, social, travel, and utilities [60], [71]. Own categories are created such as "information, system, health, social, entertainment and work" [68], and "social networking, browser, mail, and entertainment" [69]. After apps are categorized, both PHO-7 and PHO-8 are calculated. Some studies measured app usage duration and the number of times apps are launched by category [68], [69], [73]. Unlike the above approach, some studies created features PHO-7 and PHO-8 by setting some conditions rather than categorizing apps. Asselberg et al. used the usage duration of the top five apps as feature PHO-7 [71]. Mehrotra et al. used the number of unique applications launched as feature PHO-8 [74]. Asselberg et al. used the number of photos taken per day by summarizing phone camera logs as feature PHO-8 [71]. Sano et al. considered the total duration of internet access as feature PHO-7 [52].

Finally, we explain other features. For feature PHO-9, number of calendar events, Wahle et al. assumed that too many calendar events could influence depression levels and tracked the number of stored calendar events [67]. To define feature PHO-10, keypress features, Zulueta et al. considered various metrics, such as average time between keystrokes, number of backspace keypresses, number of autocorrect events, average accelerometer amplitude while typing, number of keypress sessions, and average keypress session length [75]. Mehrotra et al. defined the number of normal clicks and long clicks on the phone screen as a feature [74]. For feature PHO-11, amount of network traffic, Stütz et al. split the traffic into four categories, i.e., received, transmitted, mobile network, and wireless Network, and calculated the amount of network traffic by category [68]. Doryab et al. defined feature PHO-12, tasks and processes feature, in detail considering various metrics, such as total number of tasks and processes,

Table 3 Phone usage features.

	Feature	Reference
PHO-1	Number of communication tool usage	[21], [35], [48], [50], [52], [53], [54], [57], [60], [62], [65], [67], [68], [71], [74], [76], [77]
PHO-2	Duration of calls	[50], [52], [53], [54], [58], [68], [71], [74]
PHO-3	Number of people interacted with via communication tools	[50], [52], [53], [59], [67], [74]
PHO-4	Duration of screen on	[32], [39], [43], [52], [54], [59], [67], [68], [71], [74]
PHO-5	Number of times screen on/off	[21], [32], [52], [54], [60], [71], [74]
PHO-6	Time stamp of call, SMS, screen on/off	[21], [52], [54], [59], [68]
PHO-7	Duration of app usage	[50], [52], [60], [68], [69], [71], [73], [74]
PHO-8	Number of times apps are launched	[60], [71], [73], [74]
PHO-9	Number of calendar events	[49], [67]
PHO-10	Keypress features (number of clicks, delay, autocorrect, etc.)	[74], [75]
PHO-11	Amount of network traffic	[68]
PHO-12	Tasks and processes	[53]
PHO-13	Notification reaction	[74]
PHO-14	External device attachment	[60]
PHO-15	Battery power	[60]
PHO-16	Storage used	[60]

[*3] Standard deviation.
[*4] Short Message Service.

Table 4 Context features.

	Feature	Reference
CON-1	Ambient illuminance	[43], [49], [53], [60], [69]
CON-2	Connectivity	[49], [77]
CON-3	Weather features	[53], [59], [60]
CON-4	Days of the week	[60], [69]

Table 5 Speech features.

	Feature	Reference
SPE-1	Amount of conversation or noise	[39], [43], [45], [46], [50], [51], [53], [55], [58]
SPE-2	Acoustic feature of voice	[43], [50], [55], [57], [62], [68], [72], [80]
SPE-3	Vocal cues	[50]

frequency of change in tasks and processes, and time between changes [53]. Mehrotra et al. considered various metrics to define feature PHO-13, notification reaction. For example, they considered the number of notifications clicked, percentage of notifications clicked out of total arrived, average seen time (time from notification arrival until the time the notification was seen by the user), average decision time (time from the moment a user saw a notification until the time they acted on it, e.g., by clicking), average response time of all notifications where response time is the sum of seen and decision times [74]. Yamamoto et al. defined the number of times each user attached a charger and earphone (see feature PHO-14), changes in the amount of battery power remaining (see feature PHO-15), and the amount of storage used (see feature PHO-16) [60].

3.5 Context Feature

Context features are generated using a smartphone sensor and an external source. **Table 4** lists the context features.

Feature CON-1, ambient illuminance, is created from data obtained from illuminance sensor values or screen proximities. Here, the minimum, maximum, median, average, and SD of ambient illuminance are used [53]. Farhan et al. calculated the total duration and number of times when a participant is in a dark environment in a day [43]. Fukazawa et al. categorized the obtained illuminance sensor values as dark, medium, and bright [69]. Exler et al. factorized the lux values into categories, such as pitch black and direct sunlight [49]. Buddi et al. defined feature CON-2, connectivity, to estimate the participants' degree of social interaction [77]. They used the total number of Bluetooth IDs and the number of times the most common Bluetooth ID was seen. Feature CON-3, weather features, are obtained from an external source or sensor data implemented on a smartphone. Doryab et al. collected temperature, cloudiness, humidity, precipitation, and events (rain, snow, and wind) from a weather database [53]. Jaques et al. collected weather data related to sunlight, temperature, wind, and barometric pressure [59]. Yamamoto et al. obtained barometric pressure data from a smartphone atmospheric pressure sensor [60]. They also created a feature CON-4, days of the week feature, comprising seven-dimensional one-hot vectors and a binary feature that represents either a weekday or weekend day [60].

3.6 Speech Feature

Speech features are generated using sound data captured by a smartphone microphone. **Table 5** lists speech features.

Feature SPE-1, amount of conversation or noise, is calculated using sound data recorded from a microphone. Some studies first segment the audio stream into 15-ms frames and apply an audio classifier to obtain the number of independent conversations and their duration [45], [46]. Ben-Zeev et al. activated the microphone every two minutes to capture ambient sound [51]. If the smartphone detected speech, it remained active for the duration of the conversation. Here, the speech duration was calculated as the total duration of the conversation in minutes. Other studies have detected noise using audio classifiers, and they calculated the amount of noise as a feature [43], [53]. Some studies have applied frequency analysis (feature SPE-2) and vocal cues analysis (feature (SPE-3)) to conversation data. Grünerbl et al. used the open-source openSmile toolbox [78] to extract acoustic features such, as root mean square (RMS) frame energy, mel-frequency cepstral coefficients, pitch frequency F0, harmonic-to-noise ratio, and zero crossing-rate [55]. Guidi et al. used Camacho's SWIPE algorithm [79] to estimate the fundamental frequency (F0) in each voiced segment as a feature [80]. Saeb et al. used fast Fourier transformation to extract the dominant frequency of an audio signal [62]. Place et al. asked participants to leave an audio diary entry in a smartphone app, and extract features of vocal cues such as speaking rate, pitch and vocal effort [50].

3.7 Sleep Feature

Many studies have shown a strong relationship between sleep duration and mental health [81], [82], [83]. A number of approaches have been proposed to measure sleep duration automatically using a smartphone. For example, Ben-Zeev et al. exploited smartphone use data, accelerometer, sound features, and light levels to approximate the amount of time each participant was sleeping [51]. Staples et al. assumed that subjects sleep with their phones resting near their bed [84]. This lack of phone movement is detected by the accelerometer. They adopted a supervised approach, and multivariate accelerometer data are fit to true sleep intervals. Some studies have implemented sleep classifiers that use four types of features, i.e., light features, phone usage features (including phone lock state), activity features (e.g., stationary), and sound features from the microphone [39], [45], [46]. They develop a supervised sleep duration estimation model that combines these features. Gjoreski et al. developed a simple sleep duration estimation rule [58]. First, they calculate three durations, e.g., duration of phone being in the dark using a light sensor, the duration the phone was charging, and the duration the phone was locked, using only data from the previous night (10pm until 10am). From the three durations, the maximum was taken to estimate sleep duration.

3.8 Feature Normalization

There are individual differences in the scale of the each previously described feature. To create a model that is robust against individual differences in scale, some studies have normalized the values of each individual's data. There are two primary normalization methods, i.e., the min-max and z-score normalization

Table 6 Correlation analysis studies. Abbreviations used in the columns are given in Table A·1.

Reference	Feature	Task definition	Algorithm	Metric
Rabbi et al. [72]	ACT, SPE	Examine the correlation between the human speech and activity feature and the paper-based surveys for mental health like CES-D score, SF-36, and friendship scale.	Pearson's correlation	P-value
Osmani et al. [47]	LOC, ACT	Examine the correlation between physical activity levels and psychiatric evaluation scores (HAMD and YMRS scores).	Pearson's correlation	P-value
Doryab et al. [53]	LOC, ACT, PHO, CON, SPE	Examine the correlations between features and CES-D questions.	Tertius association rule mining [102]	Confirmation values [102]
Wang et al. [45]	LOC, ACT, SPE, SLE	Examine the correlation between smartphone data and several mental health scores (EMA, PHQ-9 and PSS).	Pearson's correlation	P-value
Faurholt-Jepsen et al. [76]	PHO	Examine the correlation between data collected using smartphones and depressive state (clinical assessments of depressive and manic symptoms using both the HDRS-17 and the YMRS, respectively).	Linear mixed-effect regression	P-value
Stütz et al. [68]	ACT, PHO, SPE	Examine the correlation between features and PSS scores from min=4 to max = 29 (high level of perceived stress).	Pearson's correlation	P-value
Guidi et al. [80]	SPE	Examine the correlation between features and the score obtained by QID and the YMRS.	Spearman's rank correlation coefficient	P-value
Sabatelli et al. [63]	LOC	Examine the correlation between patients' self-reported state (EMA, HAMD and YMRS) and their presence in an identified significant location.	Pearson's correlation	P-value
Canzian et al. [37]	LOC	Examine the correlation between each mobility metric and the PHQ score.	Pearson's correlation	P-value
Beiwinkel et al. [35]	LOC, PHO	Examine the relationship between smartphone data and depressive symptoms (YMRS and HAMD) by the between-patients analysis.	Standardized regression coefficients	Beta
Mehrotra et al. [74]	PHO	Examine the correlation between depression score (PHQ-8) and smartphone usage feature (notification metrics and their phone usage pattern).	Pearson's correlation	P-value
Saeb et al. [42]	LOC	Examine the relationship between smartphone features and depressive symptoms severity measured by the PHQ-9.	Pearson's correlation	P-value
Huang et al. [64]	LOC	Examine the correlation between sensed feature and social anxiety level measured by SIAS score rated from 0 to 4.	Pearson's correlation	P-value
Boukhechba et al. [48]	LOC, PHO	Examine the correlation between each mobility and communication feature and the SIAS score.	Pearson's correlation	P-value
Saeb et al. [62]	LOC, ACT, PHO, SPE	Examine the relationship between the time spent at different semantic locations and the level of depression and anxiety symptoms, measured by PHQ-9 and GAD-7.	Pearson's correlation	P-value
Tron et al. [70]	ACT	Examine the correlations between features of activity and measures derived from the psychiatrist's clinical assessment (PANSS).	Multiple correlation coefficient	Coefficient
Renn et al. [44]	LOC	Examine the relationship between daily mobility and frequency of depressed mood and anhedonia measured by PHQ-2.	Spearman's rank correlation coefficient	P-value
Boukhechba et al. [65]	LOC, ACT, PHO	Examine the relationship between communication patterns and students' mental states (SIAS, DASS and PANAS).	Pearson's correlation	P-value
Zulueta et al. [75]	PHO	Examine the relationship between mobile phone keyboard activity and manic and depressive signs and symptoms as measured via clinician-administered rating scales (HDRS-17 and YMRS).	Multiple linear regression	P-value

methods. Min-max normalization, which was used in Ref. [54], transforms value x by $(x - \min(x))/(\max(x) - \min(x))$. This produces values in a fixed range (0 to 1) but does not handle outliers well. Z-score normalization, which was used in Refs. [41], [46], [59], [60], [71], transforms value x by $(x-\mu)/\sigma$. This process does not produce fixed ranges between participants; however, it can handle outliers well. Palmius et al. normalized the Euclidean distance from home using z-score normalization [41]. Asselbergs et al. normalized the frequency and total duration of screen-on events of each participant using z-score normalization [71]. Differing from the above normalization techniques, Osmani et al. normalized activity levels by calculating the sum of all activity percentages on an hourly basis for each day [47].

3.9 Feature Calculation Time Interval

Most features were created on a daily basis; however, physical activity, context, and phone activity were not normal within a day. To capture the transition of feature values within a day, some studies have divided the day into intervals, and features were computed over these time intervals.

Most studies divided the time interval equally within a day. Some studies averaged the feature over six hour periods (12–6am, 6am–12pm, 12pm–6pm, and 6pm–12am) [47], [60]. Sano et al. divided the day into eight intervals (12am—3am, 3am–6am, 6am–9am, 9am–12pm, 12pm–3pm, 3pm–6pm, 6pm–9pm and 9pm–12am) and calculated screen-on/off related feature in each interval [52]. Fukazawa et al. divided the day into hourly periods (i.e., 24 time windows) and created a feature vector for each hourly interval [69]. Some studies considered different granularity during the night: 12am–3am and 6pm–12am [54], and 6am–12pm, 12pm–6pm, 6pm–9pm, 9pm–12am, and 12am–6am [21].

Some studies have designed time intervals in consideration of participant lifestyle. For example, Jaques et al. computed phone and physiological features over four time intervals per day (12am–3am, 3am–10am, 10am–5pm, and 5pm–11:59pm) based on an examination of density plots of the times students were most likely to be asleep (3am–10am) or in class (10am–5pm) [59]. Gjoreski et al. calculated features for three different epochs [58]. The first epoch is 7:30am–6pm (roughly from waking up until the end of the classes). The second epoch is 6pm–12am (period of the day when the students are studying, exercising, visiting friends, partying etc.). The third epoch is from 12am

Table 7 Regression task studies. Abbreviations used in the columns are listed in Table A·1.

Reference	Feature	Task definition	Algorithm	Metric	Cross validation
Ben-Zeev et al. [51]	LOC, ACT, SPE, SLE	Predict daily and pre/post changes in participants' mental health (UCLA Loneliness Scale, PSS and PHQ-9).	Mixed-effects linear regression	Coefficient, P-value	Test-retest reliability
Stütz et al. [68]	ACT, PHO, SPE	Predict PSS scores rated from min = 4 to max = 29 (high level of perceived stress)	Linear regression, Random forest.	MAE, P-value	5-fold and 10-fold cross validation
Saeb et al. [32]	LOC, PHO	Estimate PHQ-9 score obtained at the beginning of the study.	Linear regression	MAE, P-value	Leave-one-participant-out
Palmius et al. [41]	LOC	Estimate the depressive symptomatology obtained by participants reported weekly questionnaire (QIDS).	Linear regression, Generalized linear model (GLM)	MAE	Leave-one-participant-out
Asselbergs et al. [71]	ACT, PHO	Predict daily mood mean (range 1–10) measured by self-reported EMA with a personalized model.	Forward stepwise regression	MSE	Leave-one-participant-out
Huang et al. [64]	LOC	Predict social anxiety level measured by SIAS score rated from 0 to 4.	Least Square Error estimator, LASSO	MAE, P-value	10-fold cross validation
Wang et al. [46]	LOC, ACT, SPE, SLE	Predict participants' aggregated ecological momentary assessment (EMA) scores.	Gradient Boosted Regression Trees	MAE	Leave-one-subject-out
Staples et al. [84]	SLE	Predict concurrent and future PSQI scores ranging from 0-14.	Multiple linear regression	SE, P-value	Leave-one-participant-out
Place et al. [50]	LOC, PHO, SPE	Predict the presence of clinically assessed symptoms of depression and PTSD.	LASSO	AUC	10-fold cross validation
Jaques et al. [59]	LOC, PHO, CON	Predict continuous levels of tomorrow's reported mood and stress.	Deep neural network, Gaussian Process with Domain Adaptation	MAE	Held-out test (dividing dataset into training, validation, and testing sets using a 60/20/20% split)
Wang et al. [39]	LOC, ACT, PHO, SPE, SLE	Predict student self-reported PHQ-8 and PHQ-4 depression scores.	LASSO	MAE	10-fold cross validation
Lu et al. [61]	LOC	Predict self-reported QIDS scores and clinical assessment of depression severity.	Multi-task Learning	Coefficient of determination	Leave-one-week-out

until 7:30am (period of the day when the students are probably sleeping). This granularity is introduced in order to distinguish the students' behavior for the three different epochs of the day.

3.10 Combination of Multiple Feature Types

Tables 6, **7**, and **8** list the articles reviewed in this paper. In the tables, we show the types of features used in the reviewed papers. Location and mobility features (LOC), phone usage feature (PHO), activity feature (ACT), speech feature (SPE), context feature (CON), and sleep feature (SLE) appeared in 32, 26, 21, 14, 7, and 6 papers, respectively. Fifteen studies utilized a single feature type from six types of features. 29 studies utilized a combination of multiple feature types. The most popular combination of features is the phone usage feature (PHO) and location and mobility feature (LOC), which appeared in 18 papers. Combinations of greater than three types of feature appeared in 18 papers.

When multiple features are combined, most studies treated each feature independently; however, some produced a new feature by treating one set of features as conditions for other features. For example, Boukhechba et al. produced a new phone usage feature conditioned by the location and mobility feature, i.e., the distribution of communications (calls) at each location type (e.g., home, restaurant, campus area, and shops) [48]. Farhan et al. produced a speech feature conditioned by the location and mobility feature, e.g., the amount of social conversation of students using student location to determine if the student attended lectures, and removing conversations associated with lectures [43]. Fukazawa et al. explored more general combinations. They produced new features by combining the context, activity, and phone usage features as conditions for each other [69]. For example, they could represent common behaviors, such as walking in a dark place while using social network app.

4. Machine Learning and Statistical Analysis

4.1 Task Definition

The objectives of previous studies can be categorized as discovering effective smartphone features that strongly correlate the participants' mental state or build a model to estimate the participants' mental state with high accuracy. To achieve the former, statistical correlation analysis, which examines the correlation between psychiatric evaluation scores and smartphone features, has been widely used. Relative to the latter goal, previous studies have defined the problem in the framework of supervised machine learning and solved it as a regression or classification task. The difference between regression and classification is whether the target variables are continuous or discrete. A regression task attempts to estimate a continuous score that represents the mental state. In contrast, a classification task attempts to classify the discretized scores or class that represents mental state. Lists of papers related to correlation analysis, regression tasks, and classification tasks are given in Tables 6, 7, and 8 respectively.

The target score is primarily collected in two ways, i.e., via self-reported questionnaires using smartphones and conducting psychological scale-tests performed by a clinical psychologist. Both approaches use questionnaires that suit the purpose of the study. For example, the Patient Health Questionnaire (PHQ-9) is used to assess the severity of depressive symptoms, the Hamilton Depression Scale (HAMD) is used to determine depression, and

Table 8 Classification task studies. Abbreviations used in the columns are shown in Table A·1.

Reference	Feature	Task definition	Algorithm	Metric	Cross validation
Sano et al. [21]	LOC, PHO	Classify the two groups based on self-reported perceived stress scale ratings (e.g. high PSS score (≥17) and low PSS score (≤12)).	PCA, SVM (Linear), SVM (RBF), *k*-NN	Accuracy	10-fold cross validation with 10 times iteration
Grünerbl et al. [55]	LOC, ACT, SPE	Classify participants psychological state grades between −3 for severe depression and +3 for severe mania with intermediate steps of depression, slight depression, normal (0) slight mania and mania based on HAMD and YMRS.	*k*-NN, C4.5, conjunctive rule learner and Naïve Bayes	Accuracy, Precision, Recall	3-fold cross validation by random test/training splits with 500 times iteration
Grünerbl et al. [56]	LOC, ACT	Classify patients states measured by psychological scale-tests performed by clinical psychologist (scales between −3 (heavily depressed) to 3 (heavily manic)).	Naïve Bayes, *k*-NN, C4.5, conjunctive rule learner	Accuracy, Precision, Recall	3-fold cross-validation by random test/training splits with 500 times iteration
Sano et al. [52]	LOC, PHO	Classify high/low PSS, PSQI and MCS groups.	SVM (Linear), SVM (RBF)	Accuracy	Leave-one-participant-out
Saeb et al. [32]	LOC, PHO	Classify participants who had symptoms of depression (PHQ-9 ≥5) versus the ones with no symptoms (PHQ-9 <5).	Logistic regression classifier	Accuracy, Sensitivity, Specificity	Leave-one-participant-out
Canzian et al. [37]	LOC	Classify personalized binary depressed mood (1 if the PHQ score is larger than the average PHQ score of that user plus one standard deviation, otherwise the label is equal to 0).	SVM (RBF)	Sensitivity, Specificity	Leave-one-participant-out
Gjoreski et al. [58]	LOC, ACT, PHO, SPE, SLE	Classify the student's perceived level of stress (Stressed > Slightly stressed > Not stressed).	EM, SVM, C4.5, Bagging, Random forest	Accuracy	Leave-one-student-out
Ferdous et al. [73]	PHO	Classify the subjective stress levels of the participants collected by a question ("what is your stress level?") answered on a 5-point scale.	SVM	Accuracy, Precision, Recall	10-fold cross validation
Wahle et al. [67]	LOC, ACT, PHO	Classify samples with a PHQ-9≥11 and PHQ-9≤10.	SVM, and Random forest	Accuracy, Sensitivity, Specificity	Leave-one-participant-out
Palmius et al. [41]	LOC	Classify whether the participant is depressed (QIDS score ≥ 11).	Generalized linear model (GLM)	ROC curve, AUC	Leave-one-participant-out, 10-fold, 5-fold and 3-fold cross-validation
Farhan et al. [43]	LOC, ACT, PHO, CON, SPE	Classify individuals into the correct subgroups correlated with depression measures such as patient health questionnaire (PHQ-9).	PCA, Multi-view Clustering, SVM (Linear)	Confusion matrix	10-fold cross validation
Abdullah et al. [57]	LOC, ACT, PHO, SPE	Classify unstable (SRM score < 3.5) or stable state (SRM score ≥ 3.5).	SVM	Feature importance analysis	10-fold cross-validation with 10 times iteration
Exler et al. [49]	LOC, ACT, PHO, CON	Classify the mood into low (0 to 1.5), neutral (2 to 4), and high (4.5 to 6) measured by Multidimensional Mood Questionnaire.	C4.5, LADTree	Accuracy	Train model with data of the first three weeks and test the model with the data of the fourth
Boukhechba et al. [48]	LOC, PHO	Classify SIAS scores (low (SIAS < 34) or high (SIAS ≥ 34)) by using the GPS location and SMS texts and call features.	C4.5	Accuracy	10-fold and 3-fold cross validation
Sano et al. [54]	LOC, PHO	Classify high stress group (PSS ≥ 16) or low stress group (PSS<16) and high mental health group (MCS ≥ 50) or low mental health group (MCS ≤ 29.4).	LASSO, SVM (Linear), SVM (RBF)	Accuracy	Leave-one-participant-out, and 10-fold cross validation
Yamamoto et al. [60]	LOC, ACT, PHO, CON	Classify samples whose average daily LF/HF values were higher or lower than the average of each participants' LF/HF values.	*t*-SNE, *k*-means, *k*-NN, SVM, Random forest	Accuracy, Sensitivity, Specificity	Leave-one-participant-out
Buddi et al. [77]	PHO, CON	Classify participants into two groups: high stress (PSS score> 14) and low stress (PSS score≤13).	Naïve Bayes, Decision Trees	Sensitivity, Specificity, Precision, Accuracy, P-value	Not clearly stated
Mehrotra et al. [66]	LOC	Classify the daily depressive states; absence or presence of depressed mood.	Autoencoders, Random forest, XGBoost	Sensitivity, Specificity, DOR	Held-out test (splitting each users' data into the portions of 80% and 20%)
Lu et al. [61]	LOC	Classify subjects into the stable and unstable classes.	Multi-task Learning	F1-score	Leave-one-week-out
Fukazawa et al. [69]	ACT, PHO, CON	Classify the binary anxiety score of next day measured by STAI (1 if the STAI score increases the subsequent day and −1 if it drops).	LASSO, Random forest, XGBoost	Accuracy, Sensitivity	10-fold cross validation with 10 times iteration

the Young Mania Rating Scale (YMRS) is used to determine mania. Most studies used the score collected daily as the target variable. However, some studies used the differential of the score between two consecutive days as the target variable [51], [69]. Unlike the above target score creation, Yamamoto et al. used physiological measures (daily LF/HF values) as the target score.

Most studies attempted to estimate the score of the same day; however, some predicted the score of the next day [59], [69].

Both estimation and prediction tasks can be solved by shifting the ground truth data.

4.2 Selection of Machine Learning Algorithm

Machine learning algorithms are employed because they demonstrate high accuracy and it is easy to interpret their results. Here, we summarize the machine learning algorithms and statistical analysis methods shown in Tables 6, 7, and 8. For correlation analysis, Pearson's correlation [85] is widely used and appeared in 12 of 19 papers. Spearman's rank correlation coefficient [86] and linear regression were also used in several studies. Relative to regression tasks, 5 papers out of 12 adopted linear regression [87] or variations thereof, such as mixed-effects linear regression, the generalized linear model (GLM), and multiple linear regression. For the classification task, the most popular method was support vector machines (SVM), which appeared in 10 of 20 papers. C4.5 [88], Random forest [89], *k*-NN [90], and Naïve Bayes [87] were used in 5, 5, 4, and 3 papers, respectively. The least absolute shrinkage and selection operator (LASSO) [91] was popular for both regression (3 studies) and classification (2 studies) task.

Most previous studies created a general model that is trained from all the participants. To investigate individual differences, some studies developed personalized models learned from individual datasets [37], [59]. Some studies developed clustering based models, which first create a cluster of participants with similar behaviors, and then build a prediction or estimation model using a datasets of people in each cluster [43], [58], [60]. Relative to clustering algorithms, the expectation maximization algorithm [92], *k*-means [93], and multi-view clustering [94] were used in Refs. [43], [58], [60].

Several machine learning tools are used, such as WEKA (Machine Learning Software in Java) [95], R [96], Python [97], SPSS [98], and Matlab [99]. SPSS and Matlab are commercial services, on the other hand, WEKA, R, and Python are open-source projects and were widely used in Refs. [46], [51], [56], [58], [68], [84]. R and Python have advantages in terms of low introduction cost and the variety of libraries; however, they require statistical knowledge and programming experience because they do not include graphical user interfaces. SPSS has advantage in terms of its graphical user interface, but it is slow to adapt new machine learning algorithms. A detailed comparison of these tools is described in Refs. [100], [101].

5. Evaluation Design

To assess how the results of machine learning will generalize to an independent dataset, quantitative evaluations were performed. In the following, we summarize the evaluation metrics and the cross-validation method used to measure how well the model estimates the mental state.

5.1 Evaluation Metrics

Evaluation metrics vary depending on the target task. For correlation analysis, the P-value is the most popular metric associated with Pearson's correlation analysis, which is used in 16 out of 19 papers. For the regression task, most studies measured the difference between cross-validated predicted and observed scores. There are several variations that measure such differences, such as mean absolute error (MAE), root mean square error (RMSE), and standard error (SE). Among these, the MAE is the most popular evaluation metric (7 of 12 papers). For the classification task, accuracy, precision, recall, sensitivity, and specificity were widely used (14, 4, 3, 7, and 6 papers, respectively). Some studies have adopted multiple evaluation metrics from the above five metrics (10 of 20 papers). The definitions of accuracy, precision, and recall are shown in Ref. [103], and those of sensitivity and specificity are shown in Ref. [87].

5.2 Cross-validation

The cross-validation technique is commonly adopted for classification and regression tasks. There are two cross-validation approaches. One is *N*-fold cross-validation (14 of 29 papers), and the other is the leave-one-participant-out approach (11 of 29 papers). Both methods split a dataset into training and validation data; however, the splitting process differs. In *N*-fold cross-validation, the dataset is split regardless of the participants. For example, Grünerbl et al. performed 3-fold cross-validation, where they divided the dataset into 2/3 training and 1/3 test samples [56]. This split and evaluation process was repeated 500 times with random test/training splits. With the leave-one-participant-out approach, the model is fit on all participants except one, and the excluded participant is used to test the estimation performance of the model trained on all other participants [84].

N-fold cross validation has the advantage that many samples can be taken; however, it is necessary to consider data leakage. For example, data from the same user can be mixed into both test and training data. In addition, if we ignore the time-series relationship when we split the data, there is a possibility of predicting past data using future data. The leave-one-participant-out approach can test for new users but has the disadvantage that the number of samples is reduced when the number of participants is small.

Although most papers did not describe computation of final performance measure over multiple splits of cross-validation in detail, two common approaches are evident, i.e., macro-averaging and micro-averaging[104]. With macro-averaging, the performance measure is computed separately for each split, and the final performance measure is calculated by the mean of the performance measure over all splits [67]. With micro-averaging, the results of all splits are aggregated, and the final performance measure is calculated using the aggregated results.

6. Recommendations and Limitations

6.1 Feature Design

As described in Section 3, many mental health related features have been proposed; however, not all of these feature need to be implemented. Here, we discuss the differences in feature type selection between different mental categories. **Table 9** shows number of articles that adopted 6 feature types for each mental category. We divided metrics for diagnosis into two categories. The stress and anxiety category includes studies that used PSS, GAD-7, SIAS, and STAI. The depression category includes studies that used PHQ-2, PHQ-8, PHQ-9, CES-D, HDRS-17, HAMD, QID,

Table 9 Number of articles that adopted 6 feature types for each mental category. Numbers in bold font represent the most popular feature types.

Category	LOC	ACT	PHO	CON	SPE	SLE	Total
Stress and anxiety	**9** (28.1%) [21], [45], [48], [51], [52], [54], [62], [64], [65]	6 (18.8%) [45], [51], [62], [65], [68], [69]	**9** (28.1%) [21], [48], [52], [54], [62], [65], [68], [69], [77]	2 (6.3%) [69], [77]	4 (12.5%) [45], [51], [62], [68]	2 (6.3%) [45], [51]	32 (100%)
Depression	**16** (32.0%) [32], [35], [39], [41], [42], [43], [44], [45], [47], [51], [53], [55], [61], [62], [63], [67]	10 (20.0%) [39], [43], [45], [47], [51], [53], [55], [62], [67], [72]	10 (20.0%) [32], [35], [39], [43], [53], [62], [67], [74], [75], [76]	2 (4.0%) [43], [53]	9 (18.0%) [39], [43], [45], [51], [53], [55], [62], [72], [80]	3 (6.0%) [39], [45], [51]	50 (100%)

Table 10 Number of articles that solved classification, regression, and correlation task for each mental health category.

Category	Correlation	Regression	Classification	Total
Stress and anxiety	6 (40.0%) [45], [48], [62], [64], [65], [68]	3 (20.0%) [51], [64], [68]	6 (40.0%) [21], [48], [52], [54], [69], [77]	15 (100%)
Depression	13 (56.5%) [35], [42], [44], [45], [47], [53], [62], [63], [72], [74], [75], [76], [80]	5 (21.7%) [32], [39], [41], [51], [61]	5 (21.7%) [32], [41], [43], [55], [67]	23 (100%)

and QIDS. We investigated the appearance ratio of the feature types in the papers for each mental health category. As can be seen, different feature types tend to be selected in different mental health category. Both location and mobility feature and phone usage feature were prominently used to estimate the degree of stress or anxiety. On the other hand, the location and mobility feature was prominently used, but the activity feature, phone usage feature, and speech feature were evenly used to estimate the degree of depression. This is due to the difference in hypothesis regarding the relationship between smartphone data and the characteristic behavior among different mental health categories. Therefore, we recommend designing features according to the following process. First, we understand characteristic behaviors taken by people with mental illness or mental health issues to be studied by referring to previous literature in the domain or metrics used for diagnosis. Next, we establish hypotheses about the relationship between smartphone data and characteristic behaviors. Finally, we select features that can verify the hypothesis by referring to the features discussed in this paper.

To design features, we must consider smartphone application trends. Previous studies have developed many features related to calls and SMSs as communication tools. However, people are increasingly using other messaging tools, such as Facebook, WhatsApp, and Line. We assume it is necessary to redesign features related to the communication tool.

In this survey, we primarily focused on handcrafted features. Note that an autoencoder [105] has been proposed to extract features automatically. The autoencoder is a neural network trained to attempt to copy its input to its output [106]. Mehrotra et al. used autoencoders to automatically extract features from raw location data [66]. They demonstrated the effectiveness of autoencoder-based features in predicting the depressive states of individuals compared to manual ones. We expect that such automatic feature extraction technology will be used for other feature types, such as phone usage or activity features. In addition, to capture time-series changes in human behavior, some studies have decomposed features according to different time intervals. However, human activity changes in continuous rather than discrete manners. Recently, time-series analysis of multimodal data has been developed, such as time-series clustering [107], time-series prediction [108], and deep learning-based approaches [109], [110]. Using such methods, we can capture transitions in human activities without time interval decomposition.

6.2 Task Definition

As described in Section 4.1, there are three types of task definition, i.e., correlation analysis, classification tasks, and regression tasks. We recommend using correlation analysis to validate the presence or absence of correlation between specific smartphone-based feature and mental health. Both classification and regression tasks are used to measure the performance of designed set of features to estimate the mental state. We investigated the difference in the number of articles that solved classification and regression for each mental health category (**Table 10**). Both classification and regression are used almost as well in two mental health categories. To satisfy various user needs, we recommend evaluating proposed method in both tasks.

6.3 Machine Learning Algorithm

As discussed in Section 4.2, many studies have adopted traditional machine learning algorithms because their results are easy to interpret. In other domains, cutting edge machine learning algorithms, such as ensemble [111] and deep learning [106] methods, are widely used. For the research targeted in this survey, it is important to consider accuracy improvement and interpretability of the results. Recently, explainable AI has been examined, and many methods that can explain the results of machine learning

have been proposed, such as LIME (local interpretable model-agnostic explanations) [112] and SHAP (Shapley additive explanations) [113]. To achieve both result interpretability and high accuracy, we recommend using explainable AI with state-of-the-art machine learning methods. For example, Chen et al. proposed a platform for remote and unobtrusive monitoring of symptoms related to cognitive impairment using several consumer-grade smart devices, such as the iPhone and Apple Watch [114]. They collected data of individuals with and without cognitive impairment and tested whether these data can be used to differentiate between them by XGBoost. They analyzed characteristic features of the trained model using SHAP values and found that symptomatic participants tended to demonstrate slower typing and exhibited less routine behavior than the healthy controls.

6.4 Evaluation

As discussed in Section 5.1, evaluation metrics are diverse, especially for classification tasks (e.g., accuracy, precision, recall, sensitivity, and specificity). Evaluation metrics should be selected according to the purpose of the target application. If the application relates to medical examinations, sensitivity and specificity should be selected because we must avoid detection omission. For actual service, we recommend using precision and accuracy because it is important to maintain the system reliability from the user. In addition, as discussed in Section 5.2, there are advantages and disadvantages for each cross-validation method. We recommend using the leave-one-participant-out method to measure the performance of new user estimation tasks, and we recommend using N-fold cross-validation to increase the number of trials. However, we must consider the time-series order relation and data leakage from the same users into both training and test sets.

6.5 Limitations

Finally, we discuss several limitations. In this review, we did not consider privacy issues. Especially, location and speech features demand privacy considerations; thus, it is necessary to pay careful attention to how these features are handled. In addition, we did not apply any prioritization among features. Some studies have demonstrated important feature values; however, we assume that it is difficult to generalize feature importance results because evaluations are dependent on the target problem in each study. We did not cover features using wearable devices but did cover features using smartphones. Smartphone is not perfect to acquire behavioral data constantly. For example, users do not always carry smartphone, but sometimes leave it on the desk or leave it in their bags. On the other hand, wearable devices have an advantage that physiological data or more detailed behavioral data can be obtained constantly. Also, some wearable devices such as Apple watch and Fitbit are becoming widely used. As a future research direction, we recommend considering a combination of smartphone and wearable devices for mental state estimation. For example, Wang et al. [39] and Lu et al. [61] used Fitbit to collect heart rate in addition to data collection from smartphone. Note that several surveys have examined mental healthcare and wearable devices [10], [11], [12].

7. Conclusion

In this review, we have summarized research activities on smartphone-based mental state estimation from a machine learning perspective. Forty-four studies were reviewed and summarized from the four perspectives: list of designed features, task definition, machine learning algorithm selection, and evaluation method. By utilizing the research reviewed, it is possible to grasp the mental state of a user without any load on the mental monitoring service. To further develop mental monitoring services, we expect to consider the following topics in future. First, we should pursue more accurate mental state estimation using cutting edge machine learning technology. We should also follow changes in user behavior and mental state over time by adopting online learning frameworks. Finally we should explore effective feedback based on the estimated mental state to promote effective self-care.

References

[1] Hunt, J. and Eisenberg, D.: Mental health problems and help-seeking behavior among college students, *Journal of Adolescent Health*, Vol.46, No.1, pp.3–10 (2010).

[2] Hofmann, S.G., Newman, M.G., Ehlers, A. and Roth, W.T.: Psychophysiological differences between subgroups of social phobia, *Journal of Abnormal Psychology*, Vol.104, No.1, pp.224–231 (1995).

[3] Turner, S.M., Beidel, D.C. and Larkin, K.T.: Situational determinants of social anxiety in clinic and nonclinic samples: Physiological and cognitive correlates, *Journal of Consulting and Clinical Psychology*, Vol.54, No.4, pp.523–527 (1986).

[4] Editorial: Mind matters, *Nature*, Vol.532, p.6 (2016).

[5] World Health Organization: *Preventing suicide: A global imperative* (2014).

[6] Kawakami, N. and Tsutsumi, A.: The stress check program: A new national policy for monitoring and screening psychosocial stress in the workplace in Japan, *Journal of Occupational Health*, Vol.58, No.1, pp.1–6 (2016).

[7] Pew Research Center: The Smartphone Difference (2015).

[8] Cornet, V.P. and Holden, R.J.: Systematic review of smartphone-based passive sensing for health and wellbeing, *Journal of Biomedical Informatics*, Vol.77, pp.120–132 (2018).

[9] World Health Organization: *Mental health action plan 2013-2020* (2013).

[10] Drissi, N., Ouhbi, S., Abdou Janati Idrissi, M., El Koutbi, M. and Ghogho, M.: On the use of sensors in mental healthcare, *International Workshop on Intelligent Environments Supporting Healthcare and Well-being* (2018).

[11] Thapliyal, H., Khalus, V. and Labrado, C.: Stress detection and management: A survey of wearable smart health devices, *IEEE Consumer Electronics Magazine*, Vol.6, No.4, pp.64–69 (2017).

[12] Can, Y., Arnrich, B. and Ersoy, C.: Stress Detection in Daily Life Scenarios Using Smart Phones and Wearable Sensors: A Survey, *Journal of Biomedical Informatics*, Vol.92, p.103139 (2019).

[13] Vineetha, R., Pai, K.M., Vengal, M., Gopalakrishna, K. and Narayanakurup, D.: Usefulness of salivary alpha amylase as a biomarker of chronic stress and stress related oral mucosal changes - a pilot study, *Journal of Clinical and Experimental Dentistry*, Vol.6, No.2, pp.132–137 (2014).

[14] Dickerson, S. and Kemeny, M.: Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research, *Psychological bulletin*, Vol.130, No.3, pp.355–391 (2004).

[15] van Eck, M., Berkhof, H., Nicolson, N. and Sulon, J.: The effects of perceived stress, traits, mood states, and stressful daily events on salivary cortisol, *Psychosomatic Medicine*, Vol.58, No.5, pp.447–458 (1996).

[16] Beidel, D.C., Turner, S.M. and Dancu, C.V.: Physiological, cognitive and behavioral aspects of social anxiety, *Behaviour Research and Therapy*, Vol.23, No.2, pp.109–117 (1985).

[17] Vrijkotte, T.G.M., van Doornen, L.J.P. and de Geus, E.J.C.: Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability, *Hypertension*, Vol.35, No.4, pp.880–886 (2000).

[18] Muaremi, A., Arnrich, B. and Tröster, G.: Towards measuring stress with smartphones and wearable devices during workday and sleep,

BioNanoScience, Vol.3, No.2, pp.172–183 (2013).
[19] Dishman, R.K., Nakamura, Y., Garcia, M.E., Thompson, R., Dunn, A. and Blair, S.N.: Heart rate variability, trait anxiety, and perceived stress among physically fit men and women, *International Journal of Psychophysiology*, Vol.37, No.2, pp.121–133 (2000).
[20] Ogino, K., Ito, T., Eguchi, E. and Nagaoka, K.: Association of arginase I or nitric oxide-related factors with job strain in healthy workers, *PLOS ONE*, Vol.12, No.4, p.e0175696 (2017).
[21] Sano, A. and Picard, R.: Stress recognition using wearable sensors and mobile phones, pp.671–676 (2013).
[22] Hernandez, J., Morris, R.R. and Picard, R.W.: Call center stress recognition with person-specific models, *Affective Computing and Intelligent Interaction*, D'Mello, S., Graesser, A., Schuller, B. and Martin, J.-C. (Eds.), pp.125–134, Springer Berlin Heidelberg (2011).
[23] Wang, K., Varma, D. and Prosperi, M.: A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders, *Journal of Psychiatric Research*, Vol.107, pp.73–78 (2018).
[24] Anthes, E.: Pocket psychiatry: Mobile mental-health apps have exploded onto the market, but few have been thoroughly tested, *Nature*, Vol.532, No.7597, pp.20–23 (2016).
[25] Þórarinsdóttir, H., Kessing, L. and Faurholt-Jepsen, M.: Smartphone-based self-assessment of stress in healthy adult individuals: A systematic review, *Journal of Medical Internet Research*, Vol.19, No.2, p.e41 (2017).
[26] Khan, W., Xiang, Y., Aalsalem, M. and Arshad, Q.: Mobile phone sensing systems: A survey, *IEEE Communications Surveys & Tutorials*, Vol.15, pp.402–427 (2013).
[27] Steinhubl, S.R., Muse, E.D. and Topol, E.J.: Can mobile health technologies transform health care?, *The Journal of the American Medical Association*, Vol.310, No.22, pp.2395–2396 (2013).
[28] Luxton, D.A., McCann, R., Bush, N.C., Mishkind, M. and Reger, G.: mHealth for mental health: Integrating smartphone technology in behavioral healthcare, *Professional Psychology: Research and Practice*, Vol.42, pp.505–512 (2011).
[29] Seppälä, J., De Vita, I., Jämsä, T., Miettunen, J., Isohanni, M., Rubinstein, K., Feldman, Y., Grasa, E., Corripio, I., Berdun, J., D'Amico, E. and Bulgheroni, M.: Mobile phone and wearable sensor-based mHealth approaches for psychiatric disorders and symptoms: Systematic review, *JMIR Mental Health*, Vol.6, No.2, p.e9819 (2019).
[30] Prigerson, H.G., Monk, T.H., Reynolds III, C.F., Begley, A., Houck, P.R., Bierhals, A.J. and Kupfer, D.J.: Lifestyle regularity and activity level as protective factors against bereavement-related depression in late-life, *Depression*, Vol.3, No.6, pp.297–302 (1995).
[31] Vallée, J., Cadot, E., Roustit, C., Parizot, I. and Chauvin, P.: The role of daily mobility in mental health inequalities: The interactive influence of activity space and neighbourhood of residence on depression, *Social Science & Medicine*, Vol.73, pp.1133–1144 (2011).
[32] Saeb, S., Zhang, M., J Karr, C., Schueller, S., Corden, M., Kording, K. and Mohr, D.: Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study, *Journal of Medical Internet Research*, Vol.17, No.7, p.e175 (2015).
[33] Mitchell, P.B. and Malhi, G.S.: Bipolar depression: Phenomenological overview and clinical characteristics, *Bipolar Disorders*, Vol.6, No.6, pp.530–539 (2004).
[34] Weinstock, L.M. and Miller, I.W.: Functional impairment as a predictor of short-term symptom course in bipolar I disorder, *Bipolar Disorders*, Vol.10, No.3, pp.437–442 (2008).
[35] Beiwinkel, T., Kindermann, S., Maier, A., Kerl, C., Moock, J., Barbian, G. and Rössler, W.: Using smartphones to monitor bipolar disorder symptoms: A pilot study, *JMIR Mental Health*, Vol.3, p.e2 (2016).
[36] Kroenke, K., Spitzer, R., Williams, J. and Löwe, B.: The patient health questionnaire somatic, anxiety, and depressive symptom scales: A systematic review, *General hospital psychiatry*, Vol.32, No.4, pp.345–359 (2010).
[37] Canzian, L. and Musolesi, M.: Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis, *Proc. 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp*), pp.1293–1304 (2015).
[38] Roshanaei-Moghaddam, B., Katon, W. and Russo, J.: The longitudinal effects of depression on physical activity, *General hospital psychiatry*, Vol.31, No.4, pp.306–315 (2009).
[39] Wang, R., Wang, W., daSilva, A., Huckins, J., Kelley, W., Heatherton, T. and Campbell, A.: Tracking depression dynamics in college students using mobile phone and wearable sensing, *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol.2, pp.1–26 (2018).
[40] Association, A.P.: *Diagnostic and statistical manual of mental disorders* (*DSM-5*), American Psychiatric Association Publishing (2013).
[41] Palmius, N., Tsanas, A., Saunders, K., Bilderbeck, A., Geddes, J., Goodwin, G. and de Vos, M.: Detecting bipolar depression from geographic location data, *IEEE Trans. Biomedical Engineering*, Vol.64, No.8, pp.1761–1771 (2016).
[42] Saeb, S., Lattie, E.G., Schueller, S.M., Kording, K.P. and Mohr, D.C.: The relationship between mobile phone location sensor data and depressive symptom severity, *PeerJ*, Vol.4, p.e2537 (2016).
[43] Farhan, A.A., Lu, J., Bi, J., Russell, A., Wang, B. and Bamis, A.: Multi-view bi-clustering to identify smartphone sensing features indicative of depression, *IEEE 1st International Conference on Connected Health: Applications, Systems and Engineering Technologies* (*CHASE*), pp.264–273 (2016).
[44] Renn, B., Pratap, A., Atkins, D.D., Mooney, S. and Areán, P.A.: Smartphone-based passive assessment of mobility in depression: Challenges and opportunities, *Mental Health and Physical Activity*, Vol.14, pp.136–139 (2018).
[45] Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D. and Campbell, A.T.: StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones, *Proc. 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp*), pp.3–14 (2014).
[46] Wang, R.A., Scherer, E., Tseng, V., Ben-Zeev, D.S.H., Aung, M., Abdullah, S., Brian, R.T., Campbell, A., Choudhury, T., Hauser, M., Kane, J. and Merrill, M.: CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia, *Proc. 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp*), pp.886–897 (2016).
[47] Osmani, V., Maxhuni, A., Grünerbl, A., Lukowicz, P., Haring, C. and Mayora, O.: Monitoring activity of patients with bipolar disorder using smart phones, *Proc. International Conference on Advances in Mobile Computing & Multimedia* (*MoMM*), Vol.85, pp.85–92 (2013).
[48] Boukhechba, M., Huang, Y., Chow, P., Fua, K., Teachman, B. and Barnes, L.: Monitoring social anxiety from mobility and communication patterns, *Proc. 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp*) *and Proc. 2017 ACM International Symposium on Wearable Computers* (*ISWC*)*: Adjunct*, pp.749–753 (2017).
[49] Exler, A., Schankin, A., Klebsattel, C. and Beigl, M.: A wearable system for mood assessment considering smartphone features and data from mobile ECGs, *Proc. 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp*)*: Adjunct*, pp.1153–1161 (2016).
[50] Place, S., Blanch-Hartigan, D., Rubin, C., Gorrostieta, C., Mead, C., Kane, J., Marx, B., Feast, J., Deckersbach, T., "Sandy" Pentland, A., Nierenberg, A. and Azarbayejani, A.: Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders, *Journal of Medical Internet Research*, Vol.19, No.3, p.e75 (2017).
[51] Ben-Zeev, D., A Scherer, E., Wang, R., Xie, H. and T Campbell, A.: Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health, *Psychiatric Rehabilitation Journal*, Vol.38, No.3, pp.218–226 (2015).
[52] Sano, A.J., Phillips, A.Z., Yu, A., Mchill, A., Taylor, S., Jaques, N., Czeisler, C., Klerman, E. and Picard, R.: Recognizing academic performance,sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones, *Proc. IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks* (BSN), pp.1–6 (2015).
[53] Doryab, A., Min, J.-K., Wiese, J., Zimmerman, J. and Hong, J.I.: Detection of behavior change in people with depression, *Proc. AAAI Workshop: Modern Artificial Intelligence for Health Analytics* (2014).
[54] Sano, A., Taylor, S., Mchill, A., JK Phillips, A., K Barger, L., Klerman, E. and Picard, R.: Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones, *Journal of Medical Internet Research*, Vol.20, No.6, p.e210 (2017).
[55] Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Ohler, S., Troester, G., Mayora, O., Haring, C. and Lukowicz, P.: Smartphone-based recognition of states and state changes in bipolar disorder patients, *IEEE Journal of Biomedical and Health Informatics*, Vol.19, No.1, pp.140–148 (2014).
[56] Grünerbl, A., Osmani, V., Bahle, G., Carrasco, J.C., Oehler, S., Mayora, O., Haring, C. and Lukowicz, P.: Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients, *Proc. 5th Augmented Human International Conference* (*AH*), pp.1–8 (2014).
[57] Abdullah, S., Matthews, M., Frank, E., Doherty, G., Gay, G. and Choudhury, T.: Automatic detection of social rhythms in bipolar

disorder, *Journal of the American Medical Informatics Association*, Vol.23, No.3, pp.538–543 (2016).
[58] Gjoreski, M., Gjoreski, H., Lustrek, M. and Gams, M.: Automatic detection of perceived stress in campus students using smartphones, *Proc. 11th International Conference on Intelligent Environments* (2015).
[59] Jaques, N., Rudovic, O.O., Taylor, S., Sano, A. and Picard, R.: Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation, *Proc. IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, Vol.66, pp.17–33 (2017).
[60] Yamamoto, N., Ochiai, K., Inagaki, A., Fukazawa, Y., Kimoto, M., Kiriu, K., Kaminishi, K., Ota, J., Okimura, T., Terasawa, Y. and Maeda, T.: Physiological stress level estimation based on smartphone logs, *Proc. 2018 11th International Conference on Mobile Computing and Ubiquitous Network* (*ICMU*), pp.1–6 (2018).
[61] Lu, J., Bi, J., Shang, C., Yue, C., Morillo, R., Ware, S., Kamath, J., Bamis, A. and Russell, A.: Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning, *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol.2, pp.1–21 (2018).
[62] Saeb, S., Lattie, E., Kording, K. and Mohr, D.: Mobile phone detection of semantic location and its relationship to depression and anxiety, *JMIR mHealth and uHealth*, Vol.5, No.8, p.e112 (2017).
[63] Sabatelli, M., Osmani, V., Mayora, O., Grünerbl, A. and Lukowicz, P.: Correlation of significant places with self-reported state of bipolar disorder patients, *Proc. 4th International Conference on Wireless Mobile Communication and Healthcare* (*MobiHealth*), pp.116–119 (2015).
[64] Huang, Y., Xiong, H., Leach, K., Zhang, Y. and Barnes, L.: Assessing social anxiety using GPS trajectories and point-of-interest data, *Proc. 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp*)*: Adjunct* (2016).
[65] Boukhechba, M., Daros, A., Fua, K., Chow, P., Teachman, B. and Barnes, L.: DemonicSalmon: Monitoring mental health and social interactions of college students using smartphones, *Smart Health, CHASE 2018 Special Issue*, Vol.9-10, pp.192–203 (2018).
[66] Mehrotra, A. and Musolesi, M.: Using autoencoders to automatically extract mobility features for predicting depressive states, *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol.2, pp.1–20 (2018).
[67] Wahle, F., Kowatsch, T., Fleisch, E., Rufer, M. and Weidt, S.: Mobile sensing and support for people with depression: A pilot trial in the wild, *JMIR Mhealth Uhealth*, Vol.4, p.e111 (2016).
[68] Stütz, T., Kowar, T., Kager, M., Tiefengrabner, M., Stuppner, M., Blechert, J., Wilhelm, F. and Ginzinger, S.: Smartphone based stress prediction, *Proc. International Conference on User Modeling, Adaptation, and Personalization*, pp.240–251 (2015).
[69] Fukazawa, Y., Itoh, T., Okimura, T., Yamashita, Y., Maeda, T. and Ota, J.: Predicting anxiety state using smartphone-based passive sensing, *Journal of Biomedical Informatics*, Vol.93, p.103151 (2019).
[70] Tron, T., Resheff, Y., Bazhmin, M., Peled, A. and Weinshall, D.: Real-time schizophrenia monitoring using wearable motion sensitive devices, *Proc. 7th EAI International Conference on Wireless Mobile Communication and Healthcare* (*MobiHealth*) (2017).
[71] Asselbergs, J., Ruwaard, J., Ejdys, M., Schrader, N., Sijbrandij, M. and Riper, H.: Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: An explorative study, *Journal of Medical Internet Research*, Vol.18, No.3, p.e72 (2016).
[72] Rabbi, M., Ali, S., Choudhury, T. and Berke, E.: Passive and In-Situ assessment of mental and physical well-being using mobile sensors, *Proc. 2011 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp*), pp.385–394 (2011).
[73] Ferdous, R., Osmani, V. and Mayora-Ibarra, O.: Smartphone app usage as a predictor of perceived stress levels at workplace, *Proc. 9th International Conference on Pervasive Computing Technologies for Healthcare* (*PervasiveHealth*), pp.225–228 (2015).
[74] Mehrotra, A., Hendley, R. and Musolesi, M.: Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction, *Proc. 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp*)*: Adjunct*, pp.1132–1138 (2016).
[75] Zulueta, J., Piscitello, A., Rasic, M., Easter, R., Babu, P., Langenecker, S., McInnis, M., Ajilore, O., Nelson, P., Ryan, K. and Leow, A.: Predicting mood disturbance severity with mobile phone keystroke metadata: The biaffect digital phenotyping study, *Journal of Medical Internet Research*, Vol.20, No.7, p.e241 (2018).
[76] Faurholt-Jepsen, M., Vinberg, M., Frost, M., Christensen, E., Bardram, J. and Kessing, L.: Smartphone data as an electronic biomarker of illness activity in bipolar disorder, *European Psychiatry*, Vol.17, No.7, pp.28–31 (2015).
[77] Buddi, P., Prasad, V.V.R., Sunitha, K.V.N., Reddy, N.C.S. and Anil, C.H.: DetectStress: A novel stress detection system based on smartphone and wireless physical activity tracker, *Proc. 1st International Conference on Artificial Intelligence and Cognitive Computing* (2018).
[78] Eyben, F., Wöllmer, M. and Schuller, B.W.: Opensmile: The munich versatile and fast open-source audio feature extractor, *Proc. ACM Multimedia* (2010).
[79] Tacconi, D., Mayora, O., Lukowicz, P., Arnrich, B., Setz, C., Tröster, G. and Haring, C.: Activity and emotion recognition to support early diagnosis of psychiatric diseases, *Proc. 2nd International Conference on Pervasive Computing Technologies for Healthcare* (*PervasiveHealth*), pp.100–102 (2008).
[80] Guidi, A., Salvi, S., Ottaviano, M., Gentili, C., Bertschy, G., De Rossi, D., Scilingo, E. and Vanello, N.: Smartphone application for the analysis of prosodic features in running speech with a focus on bipolar disorders: System performance evaluation and case study, *Sensors*, Vol.15, pp.28070–28087 (2015).
[81] Tanaka, H., phd, Taira, K., Arakawa, M., Masuda, A., Yamamoto, Y., Komoda, Y., Kadegaru, H. and Shirakawa, S.: An examination of sleep health, lifestyle and mental health in junor high school students, *Psychiatry and Clinical Neurosciences*, Vol.56, pp.235–236 (2002).
[82] Abdullah, S., Matthews, M., Murnane, E., Gay, G. and Choudhury, T.: Towards circadian computing: "Early to bed and early to rise" makes some of us unhealthy and sleep deprived, *Proc. 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp*), pp.673–684 (2014).
[83] Cho, K.: Chronic 'jet lag' produces temporal lobe atrophy and spatial cognitive deficits, *Nature Neuroscience*, Vol.4, pp.567–568 (2001).
[84] Staples, P., Torous, J., Barnett, I., Carlson, K., Sandoval, L., Keshavan, M. and Onnela, J.-P.: A comparison of passive and active estimates of sleep in a cohort with schizophrenia, *npj Schizophrenia*, Vol.3, No.37 (2017).
[85] Freedman, D., Pisani, R. and Purves, R.: *Statistics*, W.W. Norton (1998).
[86] Marden, J.: *Analyzing and modeling rank data*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis (1996).
[87] Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer series in statistics, Springer (2009).
[88] Salzberg, S.L.: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., *Machine Learning*, Vol.16, No.3, pp.235–240 (1994).
[89] Breiman, L.: Random forests, *Machine Learning*, Vol.45, No.1, pp.5–32 (2001).
[90] Cover, T. and Hart, P.: Nearest neighbor pattern classification, *IEEE Trans. Information Theory*, Vol.13, pp.21–27 (1967).
[91] Tibshirani, R.: Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* (*Methodological*), Vol.58, No.1, pp.267–288 (1996).
[92] McLachlan, G. and Krishnan, T.: *The EM algorithm and extensions*, Wiley series in probability and statistics, Wiley-Interscience (2008).
[93] Duda, R., Hart, P. and Stork, D.: *Pattern classification*, Wiley (2012).
[94] Sun, J., Lu, J., Xu, T. and Bi, J.: Multi-view sparse co-clustering via proximal alternating linearized minimization, *Proc. 32nd International Conference on Machine Learning* (*ICML*), pp.757–766 (2015).
[95] University of Waikato: Weka: Data mining software in Java (2017), available from ⟨https://www.cs.waikato.ac.nz/ml/weka/⟩.
[96] The R Foundation: R programming language (2019), available from ⟨https://www.r-project.org/⟩.
[97] Python Software Foundation: Python (2019), available from ⟨https://www.python.org/⟩.
[98] IBM Corporation: IBM SPSS software (2019), available from ⟨https://www.ibm.com/analytics/spss-statistics-software⟩.
[99] MathWorks: Matlab machine learning toolbox (2019), available from ⟨https://www.mathworks.com/solutions/machine-learning.html⟩.
[100] Ozgur, C., Colliau, T., Rogers, G., Hughes, Z. and Bennie, E.: MatLab vs. Python vs. R, *Journal of Data Science*, Vol.15, pp.355–372 (2017).
[101] Kromme, J.: Python & R vs. SPSS & SAS (2017), available from ⟨https://www.r-bloggers.com/python-r-vs-spss-sas/⟩.
[102] Flach, P.A. and Lachiche, N.: Confirmation-Guided Discovery of First-Order Rules with Tertius, *Machine Learning*, Vol.42, No.1, pp.61–95 (2001).
[103] Liu, B.: *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Data-centric systems and applications, Springer (2007).
[104] Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proc. 14th International Joint Con-*

ference on Artificial Intelligence, Morgan Kaufmann, pp.1137–1143 (1995).
[105] Hinton, G. and Salakhutdinov, R.: Reducing the Dimensionality of Data with Neural Networks, *Science*, Vol.313, pp.504–507 (2006).
[106] Goodfellow, I.J., Bengio, Y. and Courville, A.: *Deep Learning*, MIT Press (2016).
[107] Hallac, D., Vare, S., Boyd, S. and Leskovec, J.: Toeplitz inverse covariance-based clustering of multivariate time series data, *Proc. 27th International Joint Conference on Artificial Intelligence* (*IJCAI*), pp.5254–5258 (2018).
[108] Kurashima, T., Althoff, T. and Leskovec, J.: Modeling interdependent and periodic real-world action sequences, *Proc. 2018 World Wide Web Conference* (*WWW*), pp.803–812 (2018).
[109] Yao, S., Hu, S., Zhao, Y., Zhang, A. and Abdelzaher, T.: DeepSense: A unified deep learning framework for time-series mobile sensing data processing, *Proc. 26th International Conference on World Wide Web* (*WWW*), pp.351–360 (2016).
[110] Ochiai, K., Senkawa, K., Yamamoto, N., Tanaka, Y. and Fukazawa, Y.: Real-time on-device troubleshooting recommendation for smartphones, *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*), pp.2783–2791, ACM (2019).
[111] Sagi, O. and Rokach, L.: Ensemble learning: A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol.8, No.5, p.e1249 (2018).
[112] Ribeiro, M.T., Singh, S. and Guestrin, C.: "Why should I trust you?": Explaining the predictions of Any Classifier, *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*), pp.1135–1144, ACM (2016).
[113] Lundberg, S.M. and Lee, S.-I.: A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems* (*NIPS*), pp.4765–4774 (2017).
[114] Chen, R., Jankovic, F., Marinsek, N., Foschini, L., Kourtis, L., Signorini, A., Pugh, M., Shen, J., Yaari, R., Maljkovic, V., Sunga, M., Song, H.H., Jung, H.J., Tseng, B. and Trister, A.: Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams, *Proc. International Conference on Knowledge Discovery & Data Mining* (*KDD*), pp.2145–2155 (2019).

Appendix

List of the abbreviations used in the Tables 6, 7, and 8 are shown in **Table A·1**.

Table A·1 Abbreviations used in Tables 6, 7, and 8.

Column	Abbreviation	Name
Task definition	CES-D	Center for Epidemiologic Studies-Depression scale.
	SF-36	Short Form-36 Health Survey.
	HAMD	Hamilton Depression Rating Scale.
	YMRS	Young Mania Rating Scale.
	HDRS	Hamilton Depression Rating Scale.
	PSS	Perceived Stress Scale.
	MCS	Mental Component Summary.
	QID	Quick Depression Inventory.
	QIDS	Quick Inventory of Depressive Symptomatology.
	EMA	Ecological Momentary Assessment.
	PHQ	Patient Health Questionnaire.
	GAD	Generalized Anxiety Disorder.
	PANSS	Positive and Negative Syndrome Scale.
	DASS	Depression, Anxiety, and Stress Scale.
	SIAS	Social Interaction Anxiety Scale.
	UCLA	University of California, Los Angeles.
	PSQI	Pittsburgh Sleep Quality Index.
	PTSD	Post Traumatic Stress Disorder.
	SRM	Social Rhythm Metric.
	LF/HF	Low Frequency / Hi Frequency.
	STAI	State Trait Anxiety Inventory.
Feature	LOC	Location and mobility feature.
	PHO	Phone usage feature.
	ACT	Activity feature.
	SPE	Speech feature.
	CON	Context feature.
	SLE	Sleep feature.
Algorithm	PCA	Principal Component Analysis.
	SVM	Support Vector Machine.
	RBF	Radial Basis Function.
	k-NN	*k*-Nearest Neighbors.
	EM	Expectation maximization.
	t-SNE	*t*-Stochastic Neighbor Embedding.
	LASSO	Least Absolute Shrinkage and Selection Operator.
Metric	MAE	Mean Absolute Error.
	RMSE	Root Mean Square Error.
	SE	Standard Error.
	ROC curve	Receiver Operating Characteristic curve.
	AUC	Area under the Curve.
	DOR	Diagnostic Odds Ratio.

Yusuke Fukazawa received his B. Eng. and M. Eng. degrees from the University of Tokyo in 2002 and 2004, respectively. He joined NTT DOCOMO, Inc. in 2004. He received his Ph.D. at the University of Tokyo in 2011. He joined RACE (Research into Artifacts, Center for Engineering), the University of Tokyo as a collaborative researcher between 2011–2016 and a visiting researcher between 2016–2019. His research interests include human behavior understanding and content recommendation. He is a member of IEEE, IEICE, JSAI and IPSJ.

Naoki Yamamoto received M. Eng. degree in Kanazawa University in 2016. He joined NTT DOCOMO, Inc. in 2016. His research interests are in mobile and wearable computing, mobile healthcare and human behavior understanding.

Takashi Hamatani received M.E. Degree and Ph.D. in Information and Computer Science from Osaka University in 2015 and 2018, respectively. He was a research fellow of the Japan Society for the Promotion of Science (JSPS) from 2017 to 2018. He joined NTT DOCOMO, Inc. in 2018. His research interests are in mobile and wearable computing, context recognition and mobile healthcare. He is a member of the Information Processing Society of Japan.

Keiichi Ochiai received his B. Eng. and M. Eng. degrees from Chiba University in 2006 and 2008 respectively. He has joined NTT DOCOMO, Inc. since 2008. He received Ph.D. at the University of Tokyo in 2017. His research interest includes social media and location data analysis, mobile computing, and mobile healthcare. He is a member of the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, the Database Society of Japan, and Association for Computing Machinery.

Akira Uchiyama received his M.E. and Ph.D. degrees in Information and Computer Science from Osaka University in 2005 and 2008, respectively. He is an Assistant Professor at Graduate School of Information Science and Technology, Osaka University. He was a visiting scholar in University of Illinois at Urbana-Champaign in 2008 and a research fellow of the Japan Society for the Promotion of Science from 2007 to 2009. His current research interests include mobile sensing and applications in pervasive and ubiquitous computing. He is a member of IEEE, ACM, IEICE and IPSJ.

Ken Ohta received his B.E., M.E., and D.E. degrees from Shizuoka University, Japan in 1994, 1996, and 1998, respectively. In 1999, he joined NTT Mobile Communications Network Inc. (NTT DOCOMO). His research interests include mobile computing, distributed systems, and system security. He is a member of the Information Processing Society of Japan and of the Institute of Electronics, Information and Communication Engineers.