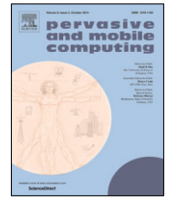




Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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

Predicting clinically relevant changes in bipolar disorder outside the clinic walls based on pervasive technology interactions via smartphone typing dynamics

Casey C. Bennett^{a,b,*}, Mindy K. Ross^c, EuGene Baek^a, Dohyeon Kim^a, Alex D. Leow^c

^a Department of Intelligence Computing, Hanyang University, Seoul, Republic of Korea

^b Department of Computing, DePaul University, Chicago, IL, USA

^c Department of Psychiatry, University of Illinois - Chicago, Chicago, IL, USA

ARTICLE INFO

Article history:

Received 23 July 2021

Received in revised form 20 March 2022

Accepted 22 April 2022

Available online 2 May 2022

Keywords:

Mental health

Human–computer interaction

Smartphones

mHealth

Ecological momentary assessment

Machine learning

ABSTRACT

Modeling smartphone keyboard dynamics as the foundation of an early warning system (EWS) for mood instability holds potential to expand the reach of healthcare beyond the traditional clinic wall's, which may lead to better ongoing care for chronic mental illnesses such as bipolar disorder. Here, we investigate the feasibility of such a system using a real-world open-science dataset. In particular, we are interested in whether passive technology interaction patterns in real-world datasets reflect findings from more controlled research trials, and the implications for clinical care. Data from 328 people who downloaded an open-science app was analyzed using a variety of machine learning methods, including different modeling methods (random forests, gradient boosting, neural networks), different types of class rebalancing, and pre-processing techniques. The aim was to predict fluctuations in PHQ scores in the weeks before the fluctuation occurred. Various feature selection methods were also employed to identify the top features driving the predictive patterns (out of total 54 starting features). Results showed predictive accuracy around ~90%, similar to controlled research trials, while revealing a number of interesting features (e.g. PTSD and mood instability) that suggest future research avenues. The findings from our analysis appear to indicate that real-world interaction data from smartphones can be utilized as an EWS monitoring tool for mood disorders like bipolar. We also discuss the broader applicability of ecological momentary assessment (EMA) approaches to connected systems combining different forms of pervasive technology interaction (smartphones, wearables, social robots) to track everyday health status.

© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Background

The defining feature of bipolar disorder from other affective disorders is reoccurring manic/hypomanic and depressive episodes [1]. Those meeting the lifetime criteria for bipolar disorder Type I can have an average of 20 episodes of mania

* Corresponding author at: Department of Intelligence Computing, Hanyang University, Seoul, Republic of Korea.

E-mail address: cabennet@hanyang.ac.kr (C.C. Bennett).

and depression lasting for months over the course of their lives [2]. As a result, it is one of the top 20 causes of disability worldwide [3] with a lifetime prevalence of 2.4% [4]. Individuals with this disorder experience repeated changes in mood, ranging between clinical definitions of mania and depression as well as subclinical fluctuations that occur on a more rapid basis.

As such, this mood instability is known to be a critical component for understanding the prognosis of bipolar disorder. This comes in two facets. First, as a symptom of the disease in patients suffering from it, which relates to their clinical outcomes and treatment success [5,6]. Second, as a predictive factor for diagnosing bipolar disorder in undiagnosed patients and for understanding the etiology of the disease [7]. Indeed, previous research has shown that mood instability in families is a trait marker for bipolar development in other relatives [8,9]. For those with bipolar disorder, this instability manifests not only as the characteristic fluctuations between syndromal depressive and manic states (as well as states with mixed features), but also as subsyndromal fluctuations during euthymia. Patients have also reported heightened responses when exposed to emotional stimuli in comparison to healthy controls, suggesting that there is an increase in affect intensity in addition to affective lability (i.e. increases in both strength and frequency) [10]. Greater mood instability has been associated with worse functional outcomes long-term, independent of episode severity [6]. Given all of this, modeling mood instability holds potential to enable better clinical care for bipolar patients by predicting changes in the disease over time before those changes occur, on multiple time scales, even prior to a formal diagnosis.

Previous research has found significantly higher mood instability in the 60 days leading up to a clinical event of depression or mania [11]. Thus, real-time monitoring of fluctuations in mood might be a target for preventative care prior to an episode, which would contribute to the feasibility of predictive and personalized medicine [12]. One approach to understanding mood instability in bipolar disorder is through technology interaction. Such interactions can theoretically be predictive of mood instability in terms of their pace, rhythm, frequency, circadian patterns, etc. A number of human-computer interaction (HCI) researchers have evaluated the predictive power of these interactions with a variety of technology, including smartphones, wearables, and other mobile devices [13]. For instance, there are many studies looking at the use of smartphones for monitoring of self-reported symptoms [9,14,15]. Most of the studies report mood instability as an important factor, though the effects can vary depending on the specific type of bipolar (Type I versus Type II) as well as when comparing different types of symptoms (e.g. depressive versus euthymic) [16].

More recent studies have begun to examine the interactions with smartphone technology directly (rather than self-report surveys) as a method to predict mental health status, such as keyboard typing dynamics [17]. For instance, Mastoras et al. [18] used typing dynamics to predict the existence of depression symptoms based on the PHQ with near 90% accuracy in a case-control study of 25 individuals, using common machine learning (ML) methods such as Random Forests and Gradient Boosting. Elsewhere, Cao et al. [19] evaluated a deep learning architecture based on gated-recurrent units (GRUs), a type of recurrent neural network) to predict depression scores in bipolar patients in a controlled trial of 20 individuals, again finding nearly 90% accuracy. Huang et al. [20] extended that work to incorporate the use of circadian rhythms as predictive factor. These approaches connect to the broader concept of digital phenotyping in terms of diagnosis, treatment, symptom monitoring, and so on [21,22]. However, challenges remain in applying digital phenotyping in the real world during pervasive technology interaction, given the variations seen in actual clinical populations, small sample sizes of controlled studies, and the current limits on explainability of AI predictions that are necessary to promote wide-scale adoption [23].

1.2. Current study approach

The study here differs from the work mentioned above, in that we are focused on a naturalistic dataset of users typical of what may be seen in real-world clinical settings [17]. The data contains users who both have and do not have bipolar disorder, as well as other mental health disorders, and is based on their pervasive interactions with a smartphone (i.e. typing dynamics). The data collection is not enforced as in a clinical trial or case-control study, but contains the types of messy data typically seen in real-world clinical datasets and electronic health records [24]. A primary goal of this paper is to evaluate how machine learning modeling of mood instability may work in real-world clinical care of bipolar patients, outside the scope of controlled clinical studies.

In particular, we are interested in whether it is possible to use such modeling of passively collected mobile technology data as a sort of early warning system (EWS) for symptoms of bipolar disorder, which could then be used to inform clinical treatment or trigger specific interventions [5]. As such, we focus below on predicting changes in depressive symptoms in bipolar disorder based on data *prior* to the change. We chose to focus on depressive symptoms as these have been found to dominate illness presentation and influence functional outcome more than manic symptoms [6,25]. Our primary research questions here are two-fold: (1) is it possible to make such predictions with real-world open datasets, and (2) does the accuracy of those predictions align with what has been observed in controlled trials in the existing literature.

1.3. Ecological momentary assessment in healthcare

There is a growing need in mood disorders, and healthcare more broadly, to utilize technology to track daily fluctuations in health status beyond the clinic walls. Healthcare is, of course, something that occurs every day, rather than the few times a year when people show up at a doctor's office or clinic. Indeed, a range of technology has shown promise in

Table 1-1
Data description — Demographics and phone dimension.

Data	Description
Age, Gender	User's age/gender
PhoneSize	Scale of phone size based on phone model

recent years in this regard, from smartphone interactions [26] to robotic pets [27]. Many studies related to these generally fall into the spectrum of ecological momentary assessment (EMA), which is form of ongoing, in-the-wild assessment based on the notion of randomly sampling each user's behavior and/or health status multiple times throughout the day over a period of time (days, weeks, months) to capture realistic human behavior when no one is directly observing them. However, it is important to note that EMA can be subjective, as the data typically is based on self-report assessments of the user's status [28]. Thus one approach to augmenting EMA is by incorporating ambulatory assessment, such as by using the technologies mentioned above, in order to monitor interaction behaviors directly in real-time [29]. Such information can then be tied back to EMA self-assessments as well as other clinical measures of health outcomes [30]. This EMA approach holds even more potential if one considers incorporating a combination of multiple monitoring technologies into a single internet-of-things type connected system: smartphones, wearables, smart home sensors, in-home robots, and other devices.

The study described here focuses on one particular type of technology interaction (smartphones), but there are many ongoing studies (both ours and others) using EMA with other types of pervasive technologies, such as socially-assistive robots (SARs). We return to this topic in the Discussion section and consider how some of our findings here might be applicable to future connected systems.

2. Methods

2.1. Data description

Our dataset was comprised of 328 individuals from an open-science dataset, who downloaded the BiAffect smartphone app from the Apple Playstore between Spring 2018 and Spring 2021. Upon downloading, the BiAffect app replaces the standard iOS keyboard with a cosmetically similar keyboard and records keystroke dynamics metadata regardless of whether the user is texting, writing an email, posting on social media, etc. [17]. This allows us to collect pervasive data on these passive technology interactions over an extended period of time. Additionally, users were pinged to complete a PHQ outcome scale weekly, which is widely used measure of depression symptoms [31]. Approximately 2/3 of the individuals reported having diagnosed bipolar, whereas others were undiagnosed (who may or may not have bipolar). Our dataset thus contains both diagnosed and potential cases of bipolar disorder in the general population. The dataset included both the variable we were trying to predict (target) and the variety of predictor variables (features), as described below.

The target variable in the dataset was the PHQChangeFlag, which indicated whether a significant change (defined as greater than 4 points) in the PHQ outcome scale occurred in the current week relative to previous weeks. Here, we used the PHQ8, which excludes the suicidality item in the PHQ9 scale. As the aim was to predict mood instability, the change could be either an increase or decrease. The value of 4 was chosen as it represents the clinical threshold for a significant change in depressive symptoms [31], whereas smaller changes may or may not represent clinically significant events and our focus here was on real-world clinical relevance.

The features are described in Table 1 (separated by feature type into Tables 1-1, 1-2, and 1-3). There were a total of 54 features, comprising a mix of clinical, demographic, and typing dynamics variables. On average, each individual had 9.6 weeks worth of data, with a PHQ change occurring roughly 18.7% of the time. Approximately half of individuals experienced at least one significant PHQ change (54.4%). The average age was 41.3. Approximately 62% reported having depression (with an average PHQ score ~ 9.3), and 24% reported having PTSD. The average number of weekly keypresses per person was roughly 5280 (total of 17.3 million overall), while the average weekly autocorrect rate was 0.013 (1.3% of presses) and the backspace rate was 0.089 (8.9% of presses). The average weekly median press duration (hold time of the keypress) was approximately 90 ms, while the average interkey delay between keypresses was 360 ms.

We note that average PHQ score here (~ 9.3) was on the lower side of what we expected, straddling the boundary of mild and moderate depression (with a score of 5–9 being mild depression and 10 being the threshold for moderate depression). This could be due to a variety of reasons. Since the bipolar diagnosis is self-reported, it could a past diagnosis from years before (or even a misdiagnosis). It could also be possibly due to the fact that bipolar individuals' mood fluctuates and thus they are not always in a depressed state. This may be something of note for future research.

2.2. Exploratory data analysis

Prior to modeling any data, we were concerned about potential collinearity in our feature set (see Table 1). In particular, since some of the typing dynamic features overlap in terms of domain (i.e. variations within the same general concept, such as median keypress duration), they may be strongly correlated. Such collinearity represents noise in the dataset,

Table 1-2

Data description – Clinical variables.

Data	Description
ADHD, Anxiety, BD_binary, Depression, OCD, PTSD, Schizophrenia, SeasonalAffectiveDisorder, SubstanceAddictionDisorder	Diagnosis of ADHD/anxiety/bipolar disorder/depression/OCD/PTSD/Schizophrenia/seasonal affective disorder/substance addiction disorder (binary, 0/1)
Baseline_PHQ	Absolute value of the PHQ8 score from weeks prior
Diag_PreferNotAnswer	Preferred not to respond to diagnosis questions (binary, 0/1)
MDQdiag	MDQ diagnosis (binary, 0/1)
NoneofTheseDiags	None of the listed diagnoses (binary, 0/1)
PHQChangeFlag	Whether there was a change of 4 or more in PHQ score since the previous week (binary, 0 means no change)

Table 1-3

Data description – Typing dynamics.

Data	Description
autocorrectRate	Fraction of autocorrect events per total keypresses for each week
autocorrectRate_wkSD	Standard deviation of autocorrect rates for each day over the week
Avg_medPressDuration	Median press duration for each session averaged over the week (in seconds)
Avg_nAlphanum, Avg_nAutocorrect, Avg_nBackspace	Number of alphanumeric/backspace/autocorrect keypress types per session averaged over each week
Avg25PercentileAA, Avg75PercentileAA, Avg90PercentileAA, Avg95PercentileAA	25th/75th/90th/95th percentile of interkey delays for alphanumeric–alphanumeric transitions for each typing session averaged over the week (in seconds)
AvgMAD_AA	Median absolute deviation of interkey delays for alphanumeric–alphanumeric transitions for each typing session averaged over the week (in seconds)
AvgMedAA, AvgMedAB, AvgMedBA, AvgMedBB	Median interkey delay for alphanumeric–alphanumeric (AA)/backspace–backspace (BB)/alphanumeric–backspace (AB)/backspace–alphanumeric (BA) transitions for each typing session averaged over the week (in seconds).
AvgVarAA, AvgVarAB, AvgVarBA, AvgVarBB	Variance of interkey delays for alphanumeric–alphanumeric (AA)/backspace–backspace (BB)/alphanumeric–backspace (AB)/backspace–alphanumeric (BA) transitions for each typing session averaged over the week (in seconds)
backspaceRate	Fraction of backspaces per total keypresses for each week
bkspRate_wkSD	Standard deviation of backspace rates for each day over the week
distToCenterPrevRatioAA	Median ratio of distance to center of key and distance to previous key for alphanumeric–alphanumeric transitions for each week
med_sessDuration	Median session duration for each week (in seconds)
medianDistCenter	Median distance from touch to center of key for each week
medianDistCenterAA, medianDistCenterAB, medianDistCenterBA	Median distance from touch to center of key for each week for alphanumeric–alphanumeric (AA)/alphanumeric–backspace (AB)/backspace–alphanumeric (BA) transitions (in mm)
medianIKD	Median IKD for each week (interkey delay), i.e. time between consecutive keypresses in seconds
medianIKD_AA, medianIKD_AB, medianIKD_BA	Median IKD for alphanumeric–alphanumeric (AA)/alphanumeric–backspace (AB)/backspace–alphanumeric (BA) transitions for each week (in seconds)
MedianPressDur	Median time a key was pressed for each week (in seconds)
medianPressDurAA, medianPressDurAB, medianPressDurBA	Median keypress duration for alphanumeric–alphanumeric (AA)/alphanumeric–backspace (AB)/backspace–alphanumeric (BA) transitions for each week (in seconds)
medIKD_wkSD	Standard deviation of median IKDs for each day over the week
nKeypresses	Number of keypresses per week

which can distract models from the true underlying signal by obscuring it. As such, we produced a color-coded correlation matrix in Python using the Seaborn library to investigate this, which can be seen in Fig. 1.

Given the large number of typing dynamic features in our dataset, the individual numbers can be difficult to see, so we direct the readers' attention to overall color patterns. In particular the rectangular blocks of lighter shade colors along the diagonal indicate significant pockets of collinearity in some of our feature domains (i.e. correlations in the range of 0.75 to (1). To deal with this, we elected to use only a single feature for each domain: interkey delay, distance to center, and press duration (medianIKD, medianDistCenter, medianPressDur). Additionally, given the strong overlap between interkey delay and some of the alpha–alpha percentile fields, we also chose to use only the Avg90PercentileAA

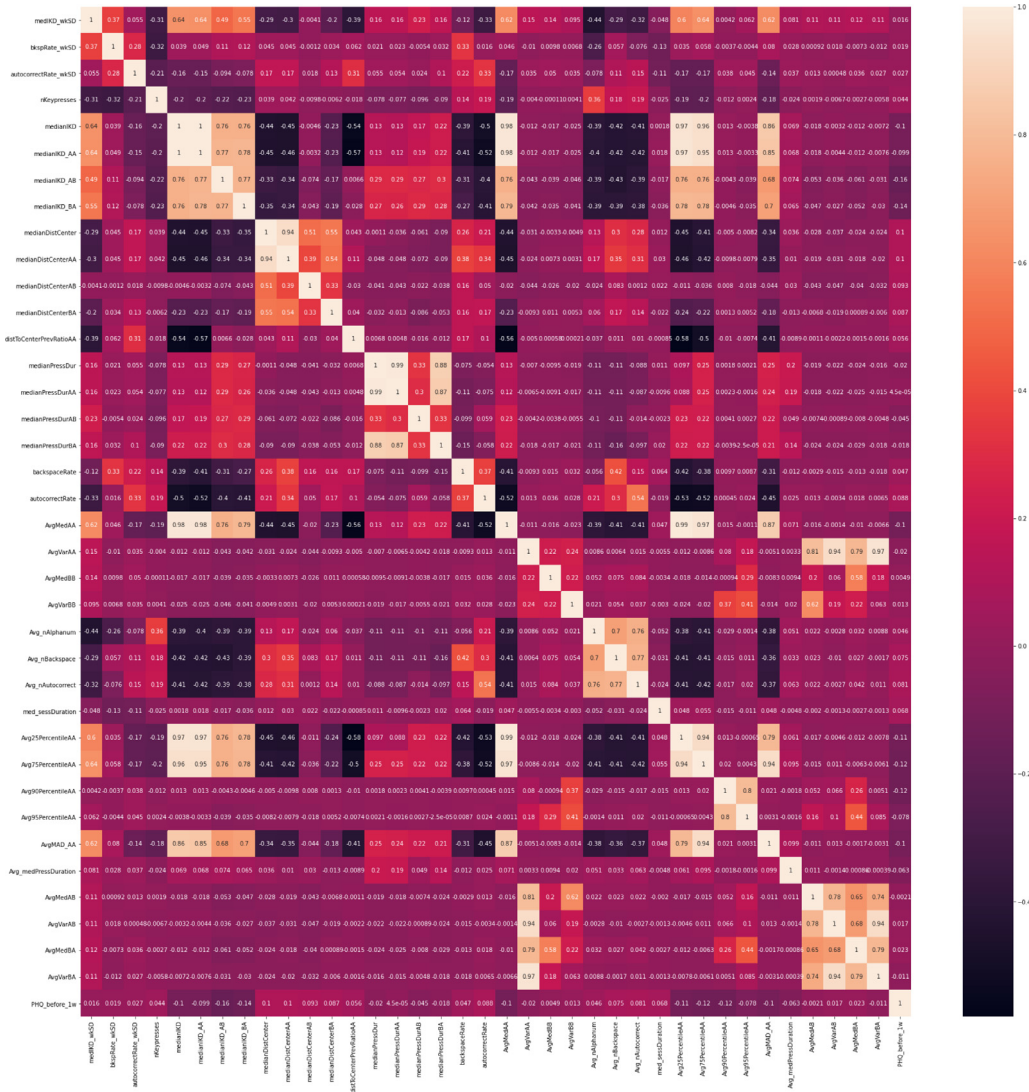


Fig. 1. Feature correlation matrix.

field. It is thought that the alpha-alpha fields likely capture more of the pauses between typing episodes, while medianIKD captures more of the actual typing speed. However, the lower percentiles (25th, 75th) seem to not capture that distinction as clearly, based on the correlations to medianIKD (similar to that reported in Vesel et al. [17]). Finally, we elected to only use the alphanumeric-backspace transitions (AvgMedAB, AvgVarAB), but to drop the reverse backspace to alphanumeric transitions. This reduced our dataset to 23 typing dynamic features, and 38 features total (from the original 54). Those 38 features can be seen in Section 3.2 in the results below (see Table 4 for reference)

2.3. Analysis approach

Data here was modeled using Python's Scikit-Learn package (<https://py.org/project/scikit-learn/>). Multiple modeling methods were attempted: Random Forest, Gradient Boosting, Neural Networks, and Support Vector Machines (SVM). Models were generally run using the default parameters in Scikit, though some experimentation was performed (similar to [32]). For instance, for SVMs the RBF kernel was used. For the neural networks, those were run using the python package Keras (<https://keras.io/>, version 2.5), which is a deep learning library based on TensorFlow, using the same default parameters as Scikit (a single Dense hidden layer of 10 units). We did attempt hyperparameter tuning with one of the better performing models (random forests), but in general that had minimal effect and is outside the scope of this paper (as we chose to focus our efforts on more promising avenues of analysis). As such, models are using set parameters. Beyond that, we explored various types of preprocessing, e.g. normalizing the features, target class rebalancing. We also

Table 2
Main result.

Classifier	Accuracy	AUC	Sensitivity
Random Forest	0.90	0.97	0.8545
Neural Network	0.87	0.93	0.9134
Gradient Boost	0.89	0.95	0.8335
SVM	0.82	0.91	0.8884

explored various types of feature selection to identify which features were driving patterns observed in the data. Those results are presented in Section 3.

Performance was estimated using multiple performance metrics, based on 5-fold cross validation, following standard machine learning guidelines [24]. Given that various types of analysis were performed, more specific details are provided in the Results (Section 3), where appropriate. Additionally, given the wide range of analysis, some results unrelated to our main conclusions are not shown, for brevity. We note that models were trained to be person-specific in order to make predictions for specific individuals, based on each unique characteristics (i.e. features).

Due to imbalances in the target variable (approximately 18.7% of samples had a change, 81% did not), we used Python's imblearn package (<https://pypi.org/project/imblearn/>) to evaluate several different methods of class rebalancing on the dataset: undersampling the majority class, SMOTE [33], and a hybrid approach (combining undersampling with SMOTE). We also employed different types of feature selection methods, including both filter-based and wrapper-based [34]. The filter-based method utilized information gain to rank each feature (univariate approach) which could then be used to select some top k features (k -count). The wrapper-based method used a Random Forest model to evaluate different sets of features across hundreds of trials, identifying the best set of features based on the predictive performance of the resulting model. To further evaluate the directionality of features relative to the target variable, we also calculated odds ratios. This allowed us to determine, for instance, whether a higher number or lower number of keypresses predicted future fluctuations of depressive symptoms.

3. Results

3.1. Main results

Our primary goal focused on whether we could predict whether a participant's PHQ score significantly changed in the coming week, based on their phone typing interaction data in the past weeks leading up to it. Our main results (henceforth referred to as our "baseline") are shown in Table 2. We used all available features as well as hybrid approach to re-balance the data (described in Methods, Section 2.2). Several types of rebalancing were explored, but the hybrid approach produced the most stable performance. As such, the subsequent tables in these results all use the same hybrid approach as Table 2, for fair comparison. SMOTE alone is known to sometimes produce overly optimistic estimates of performance in highly imbalanced datasets when a large number of synthetic samples are created [35], and likewise our data size was too limited to allow for undersampling by itself to perform well.

As can be seen in Table 2, the tree-based models (Random Forest and Gradient Boosting) were generally the best performing models, without the need for any significant hyperparameter tuning or other more sophisticated methods. These results (~90% accuracy) are similar to what has been reported elsewhere in the literature for predicting bipolar outcomes using smartphone keypress dynamics [18,19], and as such may represent the high-water benchmark for current machine learning approaches. We do note that for Neural Network approaches, more complex deep learning methodologies (beyond the single dense hidden layer used here) have pushed the result up to near 90% accuracy as well [19]. However in contrast to previous research, our prediction target here is slightly different — specifically predicting significant changes in depressive symptoms based on the keypress dynamics before the change. Thus, the research question being asked is different, with a focus on those depressive symptoms, rather than predicting bipolar in general. The main takeaway here is that pervasive smartphone interactions appear to be predictive of multiple facets of bipolar symptomology, which may have important implications for clinical care in the future.

3.2. Feature selection

In order to understand which features were driving the patterns seen in our main results, we undertook several types of feature selection approaches, including both wrapper-based and filter-based approaches (described in Methods section) [34]. For brevity, we present a summary below. Table 3 shows the results using a wrapper-based method.

As can be seen Table 3, the results are very similar to our main baseline results in the previous section, except in this case using only 16 out of the 38 total features in the dataset. Similarly, we explored filter-based method, iteratively reducing the k -count (see Methods section) to detect the threshold where ML model performance dropped. In other words, we started with all features (setting k = total number of features) and then gradually reduced that number with each iteration, building ML models on each iteration using that feature subset and measuring model performance. This

Table 3
Wrapper-based results.

Classifier	Accuracy	AUC	Sensitivity
Random Forest	0.90	0.97	0.8683
Neural Network	0.84	0.90	0.9064
Gradient Boost	0.89	0.96	0.8417
SVM	0.82	0.91	0.8783

Table 4
Filter-based results.

Ranking	Feature name	Information gain	Odds ratio
1	Age	0.34240	0.86
2	Baseline_PHQ	0.21307	2.01
3	PhoneSize	0.19388	0.67
4	Gender	0.18613	1.74
5	Depression	0.18577	1.12
6	MedianPressDur	0.18349	0.28
7	BD_Binary	0.17448	2.43
8	Anxiety	0.16636	1.43
9	nKeypresses	0.15585	2.22
10	MDQdiag	0.14184	2.00
11	PTSD	0.14063	1.89
12	ADHD	0.11527	1.24
13	medianDistCenter	0.10431	1.03
14	med_sessDuration	0.08647	1.11
15	AvgVarAA	0.08537	0.99
16	Avg_nAutocorrect	0.07778	1.20
17	AutocorrectRate	0.07588	2.10
18	AvgVarAB	0.07260	1.28
19	AvgMedBB	0.07234	0.92
20	bkspRate_wkSD	0.07229	0.99
21	Avg90PercentileAA	0.07203	0.94
22	NoneofTheseDiags	0.07109	0.54
23	Avg_medPressDuration	0.06773	1.34
24	autocorrectRate_wkSD	0.06638	1.16
25	Avg_nBackspace	0.06343	1.21
26	Avg_nAlphanum	0.06277	1.04
27	AvgMedAB	0.06210	1.03
28	AvgVarBB	0.06182	1.32
29	medIKD_wkSD	0.06169	1.52
30	SeasonalAffectiveDisorder	0.06017	1.51
31	medianIKD	0.05590	0.72
32	AvgMedAA	0.05519	0.77
33	OCD	0.05364	1.57
34	AvgMedAA	0.04973	0.97
35	distToCenterPrevRatioAA	0.04549	1.20
36	BackspaceRate	0.04537	2.06
37	SubstanceAddictionDisorder	0.04180	1.49
38	Diag_PreferNotAnswer	0.00394	1.51

approach is akin to backwards stepwise selection in a traditional statistical sense, though it differs slightly in that we are not using coefficients based on regression nor running statistical tests to compare features, but rather comparing ML model performance directly based on different feature subsets. Using this approach optimal k-count was found to be approximately the top 10–11 features across different ML classification methods. The feature ranking from the filter-based approach is shown in [Table 4](#), with the blue line indicating the approximate cutoff of important features (around 10).

We note that there are a mix of features selected here in the top 10–12 features, including several interesting features right around the cutoff line such as nKeypresses and PTSD. Curiously, PTSD was amongst the features with the highest odds ratio, with participants having PTSD being nearly twice as likely to exhibit mood instability than others. Combined with the typing dynamic features, this suggests that there may be some interplay between PTSD and bipolar symptoms that impacts their interaction patterns with technology, perhaps related to trauma-induced patterns of behavior (see Discussion) [36]. Such interaction patterns may serve as useful digital markers of symptoms for future remote monitoring of bipolar disorder. Finally, we summarize the selected features across methods in [Table 5](#) (only features selected by at least one method are shown). Features in bold/italic were selected in both wrapper-based and filter-based approaches, though we chose only the top 10 for the filter-based method in [Table 5](#) to be as conservative as possible. We note there is a large amount of overlap in the selected features between both approaches.

Outside of phone size dimensions, the selected features fall into three basic categories: typing dynamics, Age/Gender demographics, and mental disorder status. The presence of typing-related features indicate that typing dynamics and

Table 5
Selected features.

Feature selection type (# features selected)		
Features	Wrapper method (16)	Filter method (10)
Age	1	1
Gender		1
MDQdiag	1	1
phoneSize	1	1
BD_binary	1	1
Anxiety	1	1
ADHD	1	
Depression	1	1
PTSD	1	
nKeypresses	1	1
bkspRate_wkSD	1	
medianPressDur	1	1
AvgMedAA	1	
med_sessDuration	1	
Avg90PercentileAA	1	
Avg_medPressDuration	1	
Baseline PHQ	1	1

mood instability are related, concurring with previous research [18,19]. In our case, a higher number of total keypresses (7200 weekly average) and more typing errors – e.g. more backspace presses (11.7% of time vs. 8.7%) and a lower Avg90PercentileAA value (0.75 vs. 1.05) – predicted greater future fluctuations of depressive symptoms, based on the odds ratios. The results also showed that when previous weeks PHQ scores (Baseline PHQ) were higher to begin with, individuals exhibited a greater likelihood for future fluctuations in PHQ scores, averse to those with a lower Baseline PHQ (average of 11.6 vs. 9.1). Likewise, individuals who were younger in age and who identified as female also exhibited more fluctuations, which fits with existing medical knowledge. Individuals who experienced significant fluctuations had higher rates of self-reported bipolar (75% vs 50%), as well as depression, anxiety, and ADHD. Of particular note, those who experienced fluctuations also had much higher rates of PTSD (32.3% vs 17.3%).

3.3. Other analyses

We also performed a number of other analyses to further evaluate the sensitivity of model performance to slight perturbations in model parameters and features, akin to a sensitivity analysis approach in agent-based modeling in ML [37]. For brevity, the results are summarized in this section but not shown here in table form, as for the most part they had minimal effect.

We assessed the effects of normalizing the data in various ways, including normalization of the target and features. We also conducted hyperparameter tuning on our best performing model (Random Forest) using the Grid Search functionality in Python's Scikit-Learn package. However, neither normalization nor grid search had any substantive effect on model performance.

Additionally, we attempted several variations of data slicing on our feature set, to assess the importance of some feature groupings. In the feature selection analysis above, phone size dimensions, demographics, and baseline PHQ scores appear to impact the influence of typing dynamics on predicting changes in mental health status. For instance, for Baseline PHQ (prior to the change), people with higher absolute starting PHQ values appear to have different probabilities of subsequent fluctuations in depression symptoms. Likewise, in most of the research literature (including this paper), many of the keyboard dynamic features are based on the speed and rhythm with which people typed, which means that size of the phone might affect those features because on a larger phone the distance from one key to another is typically larger and the screen larger. In order to isolate and quantify the effects of other features without the influence phone/screen size or baseline PHQ scores, we conducted additional analyses where we sliced out those feature groupings a priori one-group-at-a-time, and then analyzed only the remaining ones. However, like normalization and grid search, this data slicing had minimal effect. There were some changes to selected features. For instance, when slicing out the phone size dimensions, we found that additional typing dynamic features (medianIKD and medianDistCenter) became selected by wrapper-based and filter-based methods. In other words, not surprisingly, differences in phone size impact typing speed and rhythm. Similar effects were seen with demographics and baseline PHQ scores.

However, overall model performance in terms of accuracy and AUC remained unchanged regardless of the data slicing. Some of this may though explain why different research studies are reporting slightly different sets of predictive features based on traditional statistical analyses. As has been shown elsewhere, in many domains there are potentially multiple feature sets that can produce comparably good results [32].

4. Discussion

4.1. Summary of results

The findings from our analysis here appear to indicate that real-world data from passive technology interactions with smartphones can be utilized as an EWS monitoring tool for mood disorders like bipolar. We were able to predict fluctuations in depressive symptoms (i.e. mood instability) with approximately 90% accuracy. This is similar to what has been reported elsewhere in the literature using controlled study data with a well-characterized sample [18,19]. Our study here adds to those prior results, showing that such patterns are detectable in a real-world open dataset. This holds relevance for future clinical practice given that we can predict changes in depressive symptoms before they occur, which may enable us to develop more proactive strategies for managing mood instability in patients.

Beyond those primary results, we also explored what variables were driving those patterns through feature selection and data slicing. To summarize, there were a mix of features including both keypress typing dynamics and clinical variables that were important for accurate prediction, and a subset of approximately 10–15 features was able to produce ML models of sufficient performance. We also found that data slicing out the phone dimensions caused more of the keypress typing features to appear as relevant, i.e. differences in phone dimensions can obscure the importance of the typing dynamics. Finally, we also noted some interesting individual features in the selected list, such as PTSD. We discuss the implications of that more in the next section.

4.2. Trauma & technology interaction patterns in mental health

One of the more interesting findings from the feature selection analysis in this study was the appearance of PTSD as a variable of high relevance for predicting clinically-relevant fluctuations in depression symptoms. Indeed, individuals with PTSD were twice as likely to experience a fluctuation as those who did not have PTSD. Elsewhere in the literature, interactions between PTSD and bipolar symptoms have been reported [38,39], although it is not exactly clear whether people with bipolar are more likely to be exposed to a trauma event due to bipolarity, or whether bipolarity makes them more vulnerable when exposed to trauma events, thus leading to higher rates of PTSD [40]. Regardless of the exact etiology, past trauma appears to exacerbate mood instability in people with bipolar, leading to overall worse outcomes in the long-term [41–43].

Curiously, these trauma-induced patterns of behavior seem to also mediate the technology use patterns in people with co-morbid PTSD and bipolar. In our study here, this appears to impact keyboard typing dynamics during smartphone use. Beyond this, a growing body of research has reported similar effects on a variety of pervasive technology interactions, from smartphone addiction [36] to internet use [44]. Other studies have even shown how these technology interactions even lead to poor sleep quality in patients [45]. Expanding these investigations to a broader array of interaction types and technologies, e.g. wearables, social robots (see Section 4.3), holds potential to further shed light on the aspects of bipolar that relate to quality-of-life issues and day-to-day functioning, beyond simply measuring symptoms themselves. In other words, understanding the role of trauma history in bipolar sufferers' technology use patterns may serve as useful digital markers for better understanding the course of illness in the future.

4.3. EMA-based connected ecosystems in healthcare

The results of the study here are indicative of the potential real-time monitoring via EMA of people's everyday health and health-related behaviors in combination with pervasive interactive devices capable of direct monitoring (i.e. ambulatory assessment) that are increasingly becoming embedded into the spaces where we live, work, and play. Such approaches are being applied to everything from bipolar disorder [46] to dementia and neurodegenerative diseases [47] to cardiovascular disease [48]. The advantage of this EMA approach is that it can be integrated with various types of technology, thus taking advantage of devices that have direct benefits to the end users (smartphones, wearables, social robots) without requiring the burden of installing additional technology solely for the purpose of monitoring [26,27,29]. There is further potential to use multiple types of data from each device. For instance, beyond the smartphone typing dynamics utilized here, one could envision in the future incorporating data from self-report apps, accelerometers, GPS, and other sensors embedded in modern smartphones. Moreover, such devices, when appropriately designed, can offer direct benefits to the psychological and well-being of the users, as evidenced by social robot research [49,50]. Or to put this in simpler terms, we can create a window into people's everyday health simply by giving them something to play with.

As mentioned in Section 1.3, the true power of EMA may lie in combining it with ecosystems of interconnected technology into a single internet-of-things type system: smartphones, wearables, in-home robots, and other devices. Challenges exist such as dealing with issues like sensor fusion across different devices [51]. However despite those challenges, there is burgeoning evidence that EMA utilizing technology ecosystems may create shift in healthcare technology beyond the clinic walls much akin to the advances resulting from the development of systems biology [52,53]. Patterns that are only vague signatures in single-device data may become vastly more apparent in multiple-device data. This is an area of untapped potential for future research.

4.4. Limitations

There appears to be a clear connection between the typing dynamics and mental health status in bipolar patients, and the potential ability to predict changes in depression status prior to their occurrence using smartphone data. However, we note that as the sample was drawn from the general population, a majority but not all participants self-reported a formal diagnosis of bipolar disorder, which was not possible to independently verify. While many individuals in the general population may be self-diagnosed, or simply suffer from the spectrum of mood disorders without formal diagnosing, others may also be simply mis-diagnosed. These are real-world issues that have no clear solution (i.e. the challenge of differential diagnosis). In general, we think this makes our dataset here more reflective of actual clinical practice outside research settings, but it does represent a significant limitation that warrants caution. We also detected a potential interaction between self-reported PTSD and bipolar symptoms that impacts their interaction patterns with technology, but the significance is not entirely clear given our dataset. This may be of interest for future research.

Beyond the clinical limitations, we note that the standard deviation of performance metrics across cross-validation folds was very large, which was at least partly due to the relatively small size of the dataset. Although our sample size would be considered sizable for many healthcare datasets, the differences in typing dynamics from one person to the next is often measured in fractions of second, so the differences between target classes (e.g. depression change vs. no depression change) are very small in absolute terms. This is a limitation of the current research in this area. Therefore, we advise caution in interpreting these modeling results, as there was significant instability in the observed patterns. More research is likely needed here with additional datasets in this domain before coming to any firm conclusions. Furthermore, it is possible that other modeling methods not explored here may produce better results, including both machine learning and other types of feature selection (such as feature importance from random forest models). There is an endless array of possible alternatives, which deserve further exploration. We leave those questions for future research.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alex Leow is a co-founder of KeyWise AI, currently serves as a consultant for Otsuka USA, and is currently on the medical board of Buoy health. The other authors have no conflicts to report.

Acknowledgments

This work was supported by the research fund of Hanyang University in South Korea (HY-2020), as well as the Heinz C. Prechter Bipolar Research Fund at the University of Michigan Depression Center and the Richard Tam Foundation in the United States. Grant funding support in the US came from National Center for Advancing Translational Sciences of the National Institutes of Health to the Michigan Institute for Clinical and Health Research (UL1TR002240) under Award Number 2KL2TR000434. Additional partial funding support for this research was through the “Mood Challenge for Research Kit” and 1R01MH120168 (to A. Leow). We would also like to thank our various collaborators over the years who have contributed to different aspects of this work.

References

- [1] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, Arlington, VA, 2013.
- [2] A.T. Peters, A.E. West, L. Eisner, J.H. Baek, T. Deckersbach, The burden of repeated mood episodes in bipolar I disorder: Results from the national epidemiological survey on alcohol and related conditions (NESARC), *J. Nerv. Ment. Dis.* 204 (2) (2016) 87–94, <http://dx.doi.org/10.1097/NMD.0000000000000425>.
- [3] D. Vigo, G. Thornicroft, R. Atun, Estimating the true global burden of mental illness, *Lancet Psychiatry* 3 (2) (2016) 171–178, [http://dx.doi.org/10.1016/S2215-0366\(15\)00505-2](http://dx.doi.org/10.1016/S2215-0366(15)00505-2).
- [4] K.R. Merikangas, R. Jin, J.P. He, R.C. Kessler, S. Lee, N.A. Sampson, et al., Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative, *Arch. Gen. Psychiatry* 68 (3) (2011) 241–251, <http://dx.doi.org/10.1001/archgenpsychiatry.2011.12>.
- [5] J.P. Stange, L.G. Sylvia, P.V. da Silva Magalhaes, D.J. Miklowitz, M.W. Otto, E. Frank, et al., Affective instability and the course of bipolar depression: Results from the STEP-BD randomised controlled trial of psychosocial treatment, *Br. J. Psychiatry* 208 (4) (2016) 352–358, <http://dx.doi.org/10.1016/j.jad.2017.07.029>.
- [6] S.A. Strejilevich, D.J. Martino, A. Murru, J. Teitelbaum, G. Fassi, E. Marengo, et al., Mood instability and functional recovery in bipolar disorders, *Acta. Psychiatr. Scand.* 128 (3) (2013) 194–202, <http://dx.doi.org/10.1111/acps.12065>.
- [7] M.R. Broome, K.E.A. Saunders, P.J. Harrison, S. Marwaha, Mood instability: Significance, definition and measurement, *Br. J. Psychiatry* 207 (4) (2015) 283–285, <http://dx.doi.org/10.1192/bjp.bp.114.158543>.
- [8] S. Stanislaus, M. Faurholt-Jepsen, M. Vinberg, K. Coello, H.L. Kjaerstad, S. Melbye, et al., Mood instability in patients with newly diagnosed bipolar disorder, unaffected relatives, and healthy control individuals measured daily using smartphones, *J. Affect. Disord.* 27 (2020) 336–344, <http://dx.doi.org/10.1016/j.jad.2020.03.049>.
- [9] V. Anand, B. Hu, A. Anand, Smartphone monitoring of mood instability in young depressed patients: A latent-class analyses, *AMIA Annu. Symp. Proc.* (2019) 220–227, PMID:32308814.
- [10] C. Henry, D. Van den Bulke, F. Bellivier, I. Roy, J. Swendsen, K. M'Bailara, et al., Affective lability and affect intensity as core dimensions of bipolar disorders during euthymic period, *Psychiatry Res.* 159 (1) (2008) 1–6, <http://dx.doi.org/10.1016/j.psychres.2005.11.016>.
- [11] T. Glenn, P.C. Whybrow, N. Rasgon, P. Grof, M. Alda, C. Baethge, M. Bauer, Approximate entropy of self-reported mood prior to episodes in bipolar disorder, *Bipolar Disord.* 8 (2006) 424–429, <http://dx.doi.org/10.1111/j.1399-5618.2006.00373.x>.

- [12] M. Flores, G. Glusman, K. Brogaard, N.D. Price, L. Hood, P4 medicine: How systems medicine will transform the healthcare sector and society, *Pers. Med.* 10 (6) (2013) 565–576, PMID:25342952.
- [13] T. Glenn, S. Monteith, New measures of mental state and behavior based on data collected from sensors, smartphones, and the internet, *Curr. Psychiatry Rep.* 16 (12) (2014) 523, <http://dx.doi.org/10.1007/s11920-014-0523-3>.
- [14] J.M. Bopp, D.J. Miklowitz, G.M. Goodwin, W. Stevens, J.M. Rendell, J.R. Geddes, The longitudinal course of bipolar disorder as revealed through weekly text messaging: A feasibility study, *Bipolar Disord.* 12 (3) (2010) 327–334, <http://dx.doi.org/10.1111/j.1399-5618.2010.00807.x>.
- [15] R.F. McKnight, A.C. Bilderbeck, D.J. Miklowitz, C. Hinds, G.M. Goodwin, J.R. Geddes, Longitudinal mood monitoring in bipolar disorder: Course of illness as revealed through a short messaging service, *J. Affect. Disord.* 223 (2017) 139–145, <http://dx.doi.org/10.1016/j.jad.2017.07.029>.
- [16] M. Faurholt-Jepsen, M. Frost, J. Busk, E.M. Christensen, J.E. Bardram, M. Vinberg, L.V. Kessing, Differences in mood instability in patients with bipolar disorder type I and II: A smartphone-based study, *Int. J. Bipolar Disord.* 7 (1) (2019) 5, <http://dx.doi.org/10.1186/s40345-019-0141-4>.
- [17] C. Vesel, H. Rashidisabet, J. Zulueta, J.P. Stange, J. Duffecy, F. Hussain, et al., Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A BiAffect iOS study, *J. Am. Med. Inform. Assoc.* 27 (7) (2020) 1007–1018, <http://dx.doi.org/10.1093/jamia/ocaa057>.
- [18] R.E. Mastoras, D. Iakovakis, S. Hadjimiditriou, V. Charisis, S. Kassie, T. Alsaadi, et al., Touchscreen typing pattern analysis for remote detection of the depressive tendency, *Sci. Rep.* 9 (1) (2019) 13414, <http://dx.doi.org/10.1038/s41598-019-50002-9>.
- [19] B. Cao, L. Zheng, C. Zhang, P.S. Yu, A. Piscitello, J. Zulueta, et al., DeepMood: MModeling mobile phone typing dynamics for mood detection, in: 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, 2017, pp. 747–755, <http://dx.doi.org/10.1145/3097983.3098086>.
- [20] H. Huang, B. Cao, P.S. Yu, C. Wang, A.D. Leow, Dpmood: Exploiting local and periodic typing dynamics for personalized mood prediction, in: IEEE International Conference on Data Mining, ICDM, 2018, pp. 157–166, <http://dx.doi.org/10.1109/ICDM.2018.00031>.
- [21] A.G. Mondéjar, G.F.M. Silva-Calpa, A.B. Raposo, D.C. Mograbi, Brapolar: An M-health application for remote monitoring of people with bipolar disorder, in: IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH), 2019, pp. 1–8, <http://dx.doi.org/10.1109/SeGAH.2019.8882469>.
- [22] A.G. Mondéjar, G.F.M. Silva-Calpa, A. Barbosa Raposo, D.C. Mograbi, An m-health application for remote monitoring of people with bipolar disorder through digital phenotyping and smartphone dependency, in: IEEE 33rd International Symposium on Computer-Based Medical Systems, CBMS, 2020, pp. 388–391, <http://dx.doi.org/10.1109/CBMS49503.2020.00080>.
- [23] Y. Liang, X. Zheng, D.D. Zeng, A survey on big data-driven digital phenotyping of mental health, *Inf. Fusion* 52 (2019) 290–307, <http://dx.doi.org/10.1016/j.inffus.2019.04.001>.
- [24] C.C. Bennett, T.W. Doub, R. Selove, Ehers connect research and practice: Where predictive modeling, artificial intelligence, and clinical decision support intersect, *Health Policy Technol.* 1 (2) (2012) 105–114, <http://dx.doi.org/10.1016/j.hlpt.2012.03.001>.
- [25] L.L. Judd, H.S. Akiskal, Depressive episodes and symptoms dominate the longitudinal course of bipolar disorder, *Curr. Psychiatry Rep.* 5 (6) (2003) 417–418, <http://dx.doi.org/10.1007/s11920-003-0077-2>.
- [26] J. Zulueta, A. Piscitello, M. Rasic, R. Easte, P. Babu, S.A. Langenecker, et al., Predicting mood disturbance severity with mobile phone keystroke metadata: A BiAffect digital phenotyping study, *J. Med. Internet Res.* 20 (7) (2018) PMID:30030209.
- [27] C.C. Bennett, S. Sabanovic, J.A. Piatt, S. Nagata, L. Eldridge, N. Randall, A robot a day keeps the blues away, in: IEEE International Conference on Healthcare Informatics, ICHI, 2017, pp. 536–540, <http://dx.doi.org/10.1109/ICHI.2017.43>.
- [28] S. Shiffman, A.A. Stone, M.R. Hufford, Ecological momentary assessment, *Annu. Rev. Clin. Psychol.* 4 (2008) 1–32, <http://dx.doi.org/10.1146/annurev.clinpsy.3.022806.091415>.
- [29] C.C. Bennett, C. Stanojević, S. Šabanović, J.A. Piatt, S. Kim, When no one is watching: Ecological momentary assessment to understand situated social robot use in healthcare, in: 9th ACM International Conference on Human-Agent Interaction, HAI, 2021, pp. 245–251, <http://dx.doi.org/10.1145/3472307.3484670>.
- [30] S. Intille, C. Haynes, D. Maniar, A. Ponnada, J. Manjourides, μ EMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch, in: ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2016, pp. 1124–1128, <http://dx.doi.org/10.1145/2971648.2971717>.
- [31] K. Kroenke, R.L. Spitzer, The PHQ-9: A new depression diagnostic and severity measure, *Psychiatric Ann.* 32 (9) (2002) 509–515, <http://dx.doi.org/10.3928/0048-5713-20020901-06>.
- [32] C.C. Bennett, T.W. Doub, Data mining and electronic health records: Selecting optimal clinical treatments in practice, in: 6th International Conference on Data Mining, 2010, pp. 313–318, arXiv:1112.1668.
- [33] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *J. Artif. Intell. Res.* 16 (2002) 321–357.
- [34] V. Bolón-Canedo, N. Sánchez-Maróño, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, A review of microarray datasets and applied feature selection methods, *Inform. Sci.* 282 (2014) 111–135, <http://dx.doi.org/10.1016/j.ins.2014.05.042>.
- [35] R. Blagus, L. Lusa, Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models, *BMC Bioinformatics* 16 (1) (2015) 363, <http://dx.doi.org/10.1186/s12859-015-0784-9>.
- [36] A.A. Contractor, N.H. Weiss, J.D. Elhai, Examination of the relation between PTSD symptoms, smartphone feature uses, and problematic smartphone use, *Soc. Sci. Comput. Rev.* 37 (3) (2019) 385–403, <http://dx.doi.org/10.1177/0894439318770745>.
- [37] A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. 1st Ed, Wiley, New York, ISBN: 978-0-470-87093-8, 2004.
- [38] M.W. Otto, C.A. Perlman, R. Wernicke, H.E. Reese, M.S. Bauer, M.H. Pollack, Posttraumatic stress disorder in patients with bipolar disorder: A review of prevalence, correlates, and treatment strategies, *Bipolar Disord.* 6 (6) (2004) 470–479, <http://dx.doi.org/10.1111/j.1399-5618.2004.00151.x>.
- [39] J.M. Hernandez, M.J. Cordova, J. Ruzek, R. Reiser, I.S. Gwizdowski, T. Suppes, M.J. Ostacher, Presentation and prevalence of PTSD in a bipolar disorder population: A STEP-BD examination, *J. Affect. Disord.* 150 (2) (2013) 450–455, <http://dx.doi.org/10.1016/j.jad.2013.04.038>.
- [40] J.M. Cerimele, A.M. Bauer, J.C. Fortney, M.S. Bauer, Patients with co-occurring bipolar disorder and posttraumatic stress disorder: A rapid review of the literature, *J. Clin. Psychiatry*. 78 (5) (2017) e506–e514, PMID:28570791.
- [41] L.C. Quarantini, A. Miranda-Scippa, F. Nery-Fernandes, M. Andrade-Nascimento, A. Galvão-de Almeida, J.L. Guimarães, et al., The impact of comorbid posttraumatic stress disorder on bipolar disorder patients, *J. Affect. Disord.* 123 (1) (2010) 71–76, <http://dx.doi.org/10.1016/j.jad.2009.08.005>.
- [42] J. Fawcett, What we have learned from the systematic treatment enhancement program for bipolar disorder (STEP-BD) study, *Psychiatric Ann.* 38 (7) (2008) 450–456.
- [43] N.C. Jacobson, Current evolutionary adaptiveness of psychiatric disorders: Fertility rates, parent-child relationship quality, and psychiatric disorders across the lifespan, *J. Abnorm. Psychol.* 125 (6) (2016) 824, <http://dx.doi.org/10.1037/abn0000185>.
- [44] C. Carmassi, C.A. Bertelloni, A. Cordone, V. Dell'Oste, V. Pedrinelli, F.M. Barberi, et al., Problematic use of the internet in subjects with bipolar disorder: Relationship with posttraumatic stress symptoms, *Front. Psychiatry* 12 (2021) 646385, PMID:33981258.

- [45] K. Demirci, M. Akgönül, A. Akpınar, Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students, *J. Behav. Addict.* 4 (2) (2015) 85–92, <http://dx.doi.org/10.1556/2006.4.2015.010>.
- [46] S. Schwartz, S. Schultz, A. Reider, E.F.H. Saunders, Daily mood monitoring of symptoms using smartphones in bipolar disorder: A pilot study assessing the feasibility of ecological momentary assessment, *J. Affect. Disord.* 19 (2016) 88–93, <http://dx.doi.org/10.1016/j.jad.2015.11.013>.
- [47] J. Habets, M. Heijmans, C. Herff, C. Simons, A.F. Leentjens, Y. Temel, et al., Mobile health daily life monitoring for Parkinson disease: Development and validation of ecological momentary assessments, *JMIR MHealth UHealth J.* 8 (5) (2020) e15628, <http://dx.doi.org/10.2196/15628>.
- [48] F.A.M. Jean, I. Sibon, M. Husky, T. Couffinhal, J. Swendsen, Feasibility and validity of ecological momentary assessment in patients with acute coronary syndrome, *BMC Cardiovasc. Disord.* 20 (1) (2020) 499, <http://dx.doi.org/10.1186/s12872-020-01774-w>.
- [49] A.A. Scoglio, E.D. Reilly, J.A. Gorman, C.E. Drebing, Use of social robots in mental health and well-being research: Systematic review, *J. Med. Internet Res.* 21 (7) (2019) e13322, <http://dx.doi.org/10.2196/13322>.
- [50] S. S. Šabanović, C.C. Bennett, W.L. Chang, L. Huber, Paro robot affects diverse interaction modalities in group sensory therapy for older adults with dementia, in: *IEEE 13th International Conference on Rehabilitation Robotics, ICORR, 2013*, pp. 1–6, <http://dx.doi.org/10.1109/ICORR.2013.6650427>.
- [51] R. Gravina, P. Alinia, H. Ghasemzadeh, G. Fortino, Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges, *Inf. Fusion* 35 (2017) 68–80, <http://dx.doi.org/10.1016/j.inffus.2016.09.005>.
- [52] A.L.S. de Lima, T. Hahn, L.J.W. Evers, N.M. de Vries, E. E. Cohen, M. Afek, et al., Feasibility of large-scale deployment of multiple wearable sensors in Parkinson's disease, *PLoS One* 12 (12) (2017) e0189161, <http://dx.doi.org/10.1371/journal.pone.0189161>.
- [53] H. Rashidisabet, P.J. Thomas, O. Ajilore, J. Zulueta, R.C. Moore, A.D. Leow, A systems biology approach to the digital behaviorome, *Curr. Opin. Syst. Biol.* 20 (2020) 8–16, <http://dx.doi.org/10.1016/j.coisb.2020.07.003>.