Université de Montréal

INTERPRETABLE PREDICTIONS OF FUTURE STRESS LEVELS BASED ON SMARTPHONE SENSING DATA

par

Thierry Jean, 20127778

Département de psychologie

Faculté des arts et des sciences

Travail présenté à Monsieur Antonio Zadra
dans le cadre du cours PSY 4002,
Projet de recherche Honor

3 mai 2021

INTERPRETABLE PREDICTIONS OF FUTURE STRESS LEVELS BASED ON SMARTPHONE SENSING DATA

Thierry Jean[1,2]
thierry.jean@umontreal.ca
[1]Université de Montréal, département de psychologie
[2]Centre de recherche de l'Institut en santé mental de l'Université de Montréal

Abstract

Contemporary clinical psychiatry shows difficulty accounting for interindividual and temporal variability of a person’s symptoms and deficits. To succeed in overcoming this issue, clinicians would need appropriate objective, longitudinal, and ecological data, and tools to make sense of such big data. As a proof-of-concept, a forecast model of stress for a non-clinical population was developed from data passively collected through participants’ smartphone. An open-source dataset containing 130 participants and a total of 12,315 days of valid data was used to train two eXtreme Gradient Boosting (XGBoost) machine learning models. A full model was trained using 187 input features and a reduced one using 33 input features handpicked for their higher actionability. Both had a significant predictive accuracy of about 60%, which is far above chance level. SHapley Additive Explanations (SHAP) was implemented to make the models’ prediction interpretable. Features related to mobility were found to be highly important, and those related to the social dimension to be of low importance. Results underline the potential of digital phenotyping tools for providing accurate predictions and actionable information. Future developments should improve predictive performance without sacrificing interpretability or actionability.

Introduction

A core aspect of psychiatric conditions is their dynamic influence on physiology, cognition, affect, and behaviour over time (Abdullah & Choudhury, 2018; Wright & Woods, 2020). Conditions such as major depression, bipolar disorder, and schizophrenia can be conceptualized as alternating periods of partial remission and of symptomatic dysfunction at various time scales (e.g., day, week, month; Wichers et al., 2015). The temporal dynamic of a person's condition is idiosyncratic (Kendler & Engstrom, 2017). Indeed, phenomena such as manic phases, psychosis, or depressive episodes follow heterogenous trajectories at the group level (Wright & Woods, 2020). The role of a psychiatrist is to investigate a patient's personal characteristics, environment, and their interaction to anticipate the evolution of a condition and mitigate potential repercussions. It is therefore critical for clinicians to be able to monitor the evolution of patients' mental state with enough precision to adequately predict the risk of relapse or experiencing a marked decrease in functioning or quality of life.

Nowadays, psychiatric care is primarily organized around short periodic appointments in hospital or private practice. Typically, health professionals rely on clinical interviews during which the patient reports symptoms and events since their previous appointment. This one-to-one evaluation and intervention paradigm of mental health services are not scalable, and the high demand is hardly matched by the limited number of available trained professionals (Abdullah & Choudhury, 2018). A study conducted in Ontario, Canada, revealed that people with mental health needs (14.4% of the population), on average, saw a general practitioner three times and a psychiatrist six times during 2014 (Chiu et al., 2018). Another study conducted with a multidisciplinary group of mental health practitioners reported that, on average, less than 15 minutes per encounter was spent on risk assessment (Cohen et al., 2019). Of course, these averages do not reflect the number of encounters or the time allocated to patients considered at higher risk. Under this paradigm, the low frequency of appointments makes it difficult to identify “warning signs” of future dysfunction at appropriate times (Wichers et al., 2019; Wright & Woods, 2020). Health professionals cannot properly appreciate the idiosyncratic trajectory of a person’s condition and integrate that information in the clinical decision-making process.

Additionally, current clinical assessment methods used in the medical office have inherent limitations. Clinical interviews rely on health professionals forming a subjective and qualitative judgment of the patient mental state by listening, observing, and interacting within the context of the interview. Using validated clinical scales can provide more objective and quantitative assessments of symptoms, functioning, and quality of life. Even though clinical scales are promoted by the American Psychiatric Association (2015), they remain underused. Nevertheless, both clinical interviews and quantitative scales are susceptible to biases (e.g., social desirability, recency effect) and constrained to the patient’s mnesic ability, which could be impaired (Rogler et al., 2001). More generally, the context of the medical cabinet limits the ecological validity and the generalizability of observations to the patient’s day-to-day life (Zulueta et al., 2020).

Clinical practice in mental health relies on the professionals’ ability to obtain relevant information to guide the clinical decision-making process. Measurements currently used are problematic due to the nature of the assessment methods and the low frequency of appointments. Healthcare professionals would benefit from objective, quantitative, accurate, continuous, and ecologically valid measures, acquired in the patient’s own environment. Thus, clinicians would be better equipped to consider the temporal dynamic of an individual’s psychiatric and provide personalized treatments (Abdullah & Choudhury, 2018; Insel, 2017).

Digital Phenotyping

The term digital phenoty-*ping* refers to the process of quantification of the human phenotype in its idiosyncrasy, moment-by-moment, and in the natural living environment, by means of the smartphone (Onnela & Rauch, 2016; Torous & Baker, 2016). Following this perspective, interactions with technology would be a valuable source of information to fully understand the human phenotype. For example, reduced mobility patterns observed in GPS and Wi-Fi signals could be related to a broken limb, a depressive episode, or societal factors such as access to transport (Birk & Samuel, 2020). Similarly, Jain et al. (2015) proposed the concept of digital phenoty-*pe* defined as a "continuously measured manifestation of biologic disease" (p. 463). This conception suggests a certain "digital signature" for biological phenomena (i.e., the genotype). A person with Parkinson's disease could show signs of tremor through distinctive phone typing patterns, or increased gyroscope activity while holding the phone for a call. The concepts of *digital phenotyping* and *digital phenotype* differ slightly, but they are often used interchangeably as they both strive for a similar purpose: developing novel methods for ecologically valid measurement of human activity outside the laboratory (Birk & Samuel, 2020). Notably, bipolar disorder, schizophrenia, and depression were studied using data from smartphones (Beiwinkel et al., 2016; Ben-Zeev et al., 2017; Grunerbl et al., 2015), smartwatches (Gershon et al., 2016), and social media usage (Birnbaum et al., 2019).

The ubiquity of smartphones in the pockets of millions of users offers a unique window on individuals' behaviours, environment, and day-to-day mental state. Over the past decade, the rate of smartphone adoption (iPhone, Android, BlackBerry) in the province of Québec, Canada, rose from 13% in 2009 to 77% in 2019 according to the Centre for Facilitating Research and Innovation in Organizations (CEFRIO). The adoption rate climbs to 94% for the 18 to 34 age group. This trend of increased smartphone ownership is observable around the world as the entry cost for this technology decreased drastically (Ericsson Mobility Report, 2020). The smartphone is becoming a technological Swiss Army knife, replacing devices, namely music players, cameras, GPS navigators, watches, alarm clocks, calendars, voice recorders, calculators, wallets, etc. While there is a variety of usage profiles, people are generally motivated to carry their device and interact with it throughout the day, every day.

In addition to being ubiquitous, smartphones are unobtrusive devices, which promotes ecologically valid measurements and limits the risk of denaturing behaviours. Vaizman (2018) proposes four principles for the acquisition of authentic ecological data: 1) naturally used devices, 2) unconstrained device placement, 3) natural environment, and 4) natural behavioural content. Even though previous research using various specialized devices yielded conclusive results, most fail to meet principle 1), 2) and 4). Devices with the sole purpose of monitoring have considerable risks of losing the user's engagement (e.g., forgetting the device, deciding not to carry it), or have their saliency (e.g., placement, visibility) denatures behaviours (Farago, 2012; Onnela & Rauch, 2016; Vaizman, 2018). The applicability of deploying specialized devices at a large scale, outside of a research program, remains unknown. Being both ubiquitous and unobtrusive, smartphones are a promising platform to develop digital phenotyping tools for clinical purposes.

The digital phenotyping data acquired through dedicated smartphone applications can be divided as either active or passive (Onnela & Rauch, 2016). On one hand, active data requires the user's direct involvement, either by answering questionnaires, or completing cognitive tasks on the application. On the other hand, passive data exploits the device's numerous sensors (accelerometer, gyroscope, Wi-Fi, Bluetooth, GPS, microphone, camera, touch screen, etc.), and its usage metadata (call and texts logs, phone unlocks, contacts, battery status, installed apps, app use, etc.). Importantly, the amount and the type of active data required (e.g., a questionnaire) could interact with the phenomenon measured (e.g., stress), in addition to potentially becoming burdensome for the user (Torous et al., 2019). This highlights a trade-off between the quantity of active data and the behavioural authenticity according to the four principles from (Vaizman, 2018). Thus, finding a balance between active and passive data is essential to obtain continuous, and ecologically valid measurements that address existing limitations in conventional mental health clinical practice.

Machine Learning

Digital phenotyping leads to massive amounts of noisy data. Machine learning is the ideal tool to identify patterns amongst the high dimensionality data available, and generate meaningful information about behaviour, emotions, social activity, or cognition (Mohr et al., 2017). Broadly, machine learning refers to a system (the *model*) that receives large amounts of input data and uses an algorithm to learn a relationship (the *training*) between input variables (the *features*) and an output variable (the *label*; c.f. Appendix for machine learning glossary). Conceptually, the label is the information one wants to learn from the features. Once the training is done, the model can be given unseen features to generate a prediction about its label.

Indeed, machine learning has proven helpful to develop forecast models of future mental states (Dwyer et al., 2018; Garcia-Ceja et al., 2018; Mohr et al., 2017). Such forecast model could aid clinicians in the prognostic of an individual with a psychiatric condition. Spathis et al. (2019) created a forecast model of mood (valence and arousal) using mood levels 566 app users reported sparingly over 3 years. Their model uses mood self-reports from the previous 3 weeks to predict the sequence of mood levels for the next week. They highlighted the importance of considering personal factors such as mood variability, and external factors like the day of the week (highest variability on Mondays and lowest on Saturdays). Wang et al. (2020) have been working with 55 participants diagnosed with schizophrenia. Using tree ensemble models (RandomForest, ExtraTrees, and XGBoost) on digital phenotyping data, they predicted ratings on a social functioning clinical scale with a mean absolute error of less than 10% on each subscale. Umematsu et al. (2019a) used advanced deep learning models on smartphone and smartwatch data to predict if a person were "high stress" or "low stress" on the next day from the 7 days prior. Impressively, stress was correctly predicted in 83.6% of the test cases. One author of this research noted that simple model predicting the "same-day" stress levels could only achieve about 74% accuracy, highlighting the complexity of the task (Jaques, 2020). She also reports that the accuracy for predicting happiness or stress typically found in the literature ranges from 55 to 76% (Bogomolov et al., 2014; Bogomolov et al., 2014; Canzian & Musolesi, 2015; Grunerbl et al., 2015). Predicting a relapse or the worsening of a patient's mental state remains notoriously difficult (Cohen, 2019).

Interpretability and actionability

Currently, the opacity of machine learning models constitutes one of the main barriers to the use of predictive tools based in clinical settings. A model could provide excellent predictions, but those are not necessarily interpretable or actionable. The "black box" analogy is often mentioned to describe the opacity of the process transforming inputs into an output (Murdoch et al., 2019). Exploring the black box constitutes itself an emerging and complex field of research usually titled "interpretable", or "explicable" artificial intelligence. In this study, the term "interpretability" will refer to the ability to describe how inputs contributed towards a prediction, and this description will be designated as an "explanation." More vaguely, "actionability" refers to the practical usefulness of an interpretable model. In other words, an actionable model would provide meaningful prediction explanations (interpretable) that can be included by mental health professionals in their decision-making (actionable). Importantly, the General Data Protection Regulation implemented by the European government in 2018 defines the right of the person from whom the data originates to be unaffected by a decision based solely on automatic processing, thereby emphasizing the right to an explanation of the prediction. Interpretability and actionability are crucial to garner trust and to allow critical evaluations of the model, both from professionals and people receiving care.

Interpretable machine learning models. Promising approaches to make models interpretable rely on Shapley values. These values originate from the field of economy and game theory and are used to credit members of a group proportionally to their contribution towards an outcome. Analogically, for a given model prediction, a Shapley value can be attributed to each input feature to quantify their contribution towards this particular prediction (Lundberg & Lee, 2017). This approach provides a *local* explanation since it describes a single prediction at a time, and not how the full model functions. However, by sampling a sufficient number of predictions and summarizing the Shapley values obtained, it can offer rich insights into the model's global behaviour, including the most important variables, interaction effects, prevalence, outliers, etc. (Lundberg et al., 2019). The strong theoretical and mathematical foundations of Shapley values make them a reliable tool for interpretability.

Digital phenotyping data types and feature categories

When discussing interpretability and actionability of digital phenotyping, it is important recognize the different *data types* and *feature categories*. On one hand, data types refer to the type of signal collected by the device such as accelerometer, GPS, Wi-Fi, phone metadata, etc. A data type can provide various information. For example, call logs constitute one data type, but the contact calling, the duration of calls, and the number of missed calls can lead to different insights.

On the other hand, feature categories form a framework to distinguish the type of insight data can provide. Rohani et al. (2018) reviewed 46 digital phenotyping studies and classified 85 features into 7 categories: *social*, *physical activity*, *location* (renamed to *mobility* for this study), *device*, *subject* (i.e., personal characteristics like voice pitch), *environment* (e.g., ambient), and *biological* (e.g., heart rate). Across studies, they assessed the correlations between features and mood symptoms. To summarize, many features had split results; mobility and physical activity were generally significant; calls and SMS were generally nonsignificant.

Objectives

In the context of this Honor research project, I joined a laboratory currently developing predictive tools based on digital phenotyping data from clinical populations. An objective of the research program is forecasting potential symptoms and deficits related to a person's clinical condition. Data acquisition is currently undergoing but was halted by the COVID-19 epidemic. For this reason, my objectives shifted to be able to contribute to the research program in the meantime. Then, I decided to leverage open-source datasets to create proof-of-concept models able to forecast stress levels in a non-clinical population.

The first objective is to develop an interpretable machine-learning forecast model of self-assessed daily stress levels using digital phenotyping data. Stress is a particularly relevant target considering its temporal fluctuations both in clinical and non-clinical populations. This objective would be met by achieving significant predictive performance and successfully implementing SHAP. The implementation of Shapley values to explain predictions will be a new contribution to the literature on digital phenotyping and the progress towards clinical tools. Positive results would support the value of digital phenotyping approaches and its potential for in psychiatric applications (e.g., prognostic, symptoms forecast).

The second objective is to develop a reduced forecast model of stress with increased actionability. Certain features from the original “full model” may contribute to its performance but are seemingly unactionable for clinicians. Accordingly, obtaining a prediction explanation attributing large Shapley values to unactionable features (e.g., phone battery temperature, relative location x coordinates) seems of low practical value. Therefore, training a reduced model exclusively on features judged actionable and handpicked *a priori* to training could yield a model with prediction explanations of higher practical value. The reduced selection of features should cover the most data types and feature categories possible. This objective would be met by achieving significant predictive performance, noted that performance is expected to be equal or lower to the full model.

Methodology

Dataset

To meet the objectives and to create a robust proof-of-concept relevant for future use, a digital phenotyping dataset including data with the following characteristics was needed: (1) in situ (2) longitudinal (3) passive sensor, and (4) self-assessment of stress. After exploring the literature, a selection of 7 open-source datasets was acquired and inspected (*Reality Mining* [Eagle & Pentland, 2006], *Social Evolution* [Madan et al., 2012], *Friends and Family* [Aharony et al., 2011], *StudentLife* [Wang et al., 2014], Asselbergs et al. [2016], *Copenhagen Network Study* [Sapiezynski et al., 2019], *ExtraSensory* [Vaizman et al., 2017]). Considering that it met all five criteria, the *Friends and Family* (http://realitycommons.media.mit.edu/friendsdataset.html) was selected as it had the longest span and the most participants. The dataset was acquired after signing a web form agreeing to preserve the participants’ confidentiality and respect the terms of use.

The *Friends and Family* study was conducted in two phases: Phase I included 55 participants, started in March 2010, and lasted 6 months; Phase II included 130 participants (including a portion from Phase I), started in September 2010, and lasted 12 months. Participants all came from the same residential community of about 400 members organized around a research university in North America. Every participant is in a relationship, some with another participant, and some have children. For the duration of the study, they were provided with an Android smartphone running the *Funf* platform, an open-source software to acquire passive data from phone sensors. In full, the study included a launch questionnaire with demographic questions, passive sensing, fitness indicators, daily self-assessments, and weekly and monthly surveys.

For this study, only the passive sensing data was used to train the stress forecast models. This way, the models would only rely on readily available data and minimal active contribution from the user. The following passive data sources were available: Wi-Fi, Bluetooth, GPS, accelerometer, phone app metadata, battery metadata, call and SMS logs (c.f. Table 1). None was excluded as it would be arbitrary at this stage of development. The stress level for the day was self-assessed on a scale from 1 to 7 with the anchors *1: Very unstressed, 4: Nor stressed, nor unstressed, 7: Very stressed*. The dataset also included daily self-assessments of *happiness, productivity, eating healthy, night sleep duration*, and *time spent doing social activities*, but they were discarded for this study. To note, assessment questions and answers were modified during the original study. Only the data following December 2010 was of the aforementioned format of interest. Finally, the data from all participants, from days with a valid self-assessment of stress, was included for a total of 12,315 days of data.

Table 1
Passive data

Data type	Raw data	Sampling rate*
Wi-Fi	Wi-Fi MAC address Received signal strength indicator (RSSI)	irregular
Bluetooth	scan for nearby Bluetooth MAC address	every 5 min
Location	relative location (x, y) coordinates estimation accuracy	every 30 min
Accelerometer	accelerometer (x, y, z) reading counts of periods with important accelerometer activity	4Hz every 2 min
Phone applications	applications installed on device applications in-use	every 10 min to 3 h every 30 sec when phone is in use
Battery metadata	temperature, voltage, currently plugged (yes/no)	Event: when a value changes
SMS logs	incoming or outgoing deidentified phone number	Event: SMS
Call logs	incoming, outgoing, or missed duration deidentified phone number	Event: Call

Note. Data with an "Event-based" sample rate is recorded each time the Event occurs.

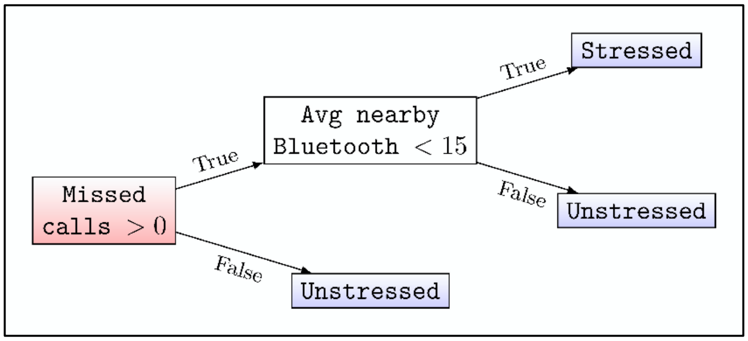
Model development

Operationalization of the model's objective. After examining the available data in the *Friends & Family* dataset, the objective of the model was defined as: "Given passive data for a certain day, will the person be stressed or unstressed on the next day." This formulation poses the problem as a binary classification task; the person is either stressed or unstressed. This binary formulation is in line with previous research on daily stress, mood, and happiness (Umematsu et al., 2019a; Umematsu et al., 2019b). Accordingly, self-reported stress levels were binarized with scores above 4 representing "stressed" (36.4% of the sample), and those equal to or below 4 signifying "unstressed" (63.6% of the sample).

Data preprocessing. Since the dataset was gathered for a different purpose, a significant amount of preprocessing was required. First, the dataset included one file per feature, each containing all participants. It was converted to a format with one file per participant, containing all features. Next, all features had to be resampled to a uniform rate. Finding a common denominator proved to be challenging as certain features are sampled at a fixed rate and others are event-based (c.f. Table 1). The data was aggregated in periods of 30 minutes, then 48 periods were aggregated to summarize an entire 24-hour day. This two-step approach allowed for a better representation of fluctuation of features during the day. Indeed, while aggregating the 48 periods into one, the mean, the standard deviation, the median, the minimum, the maximum, and in some cases the sum, could be computed for each feature. In the end, 187 features were computed across 12,315 days of data.

Model selection. An eXtreme Gradient Boosting (XGBoost) model was created using the XGBoost Python package (https://xgboost.readthedocs.io/en/latest). XGBoost models use an ensemble of *decisions trees* to fit the data. A single decision tree would consider the passive data (features) from all available days (examples) and try to determine the best criteria to correctly separate the “stressed” and the “unstressed” (label; see Figure 1 for an example). Then, an ensemble of trees pools the predictions from multiple different decision trees to formulate a final prediction. XGBoost models use this ensemble approach in conjunction with sophisticated computation techniques to optimize performance while avoiding overfitting and underfitting. This type of model generally performs very well and is often used in data mining competitions. Wang et al. (2020) also used XGBoost and other tree ensembles to predict clinical subscale ratings of participants diagnosed with schizophrenia.

Figure 1
Decision Tree Example



Model training. With 36.4% of the available data being "stressed" days and 73.6% "unstressed", the *class imbalance* has to be considered. Class imbalance refers to the fact that examples are not evenly distributed the two categories (i.e., a 50-50 split). The lower representation of "stressed" days influences the model's training since there are fewer examples to learn from. If this imbalance is not accounted for, a model always predicting "unstressed" would correctly predict 73.6% of the test cases. To balance the two categories, all "stress" and "unstress" days are separated into two pools, then an equal number of days from each category is randomly sampled without repetition to create a *balanced* training set containing 50% of each category. Then, from the remaining days, an equal number of "stress" and "unstress" days are picked to create a balanced test set. This resulted in a training set composed of 6000 days and a test set of size 2914. Therefore, the balanced training should increase the model's ability to identify "stress" days, and correctly predicting above 50% of the test cases would indicate the model learned patterns within the data to distinguish "stressed" and "unstressed" days.

Implementing interpretability mechanisms. Once the XGBoost model was trained, the Python package SHAP (Shapley Additive exPlanations; https://shap.readthedocs.io/en/latest/) was implemented to make it interpretable. To generate an explanation for a prediction, SHAP gives the passive data from one day (i.e., features from one example) to the trained model, then the model outputs a prediction, and SHAP computes an explanation for this prediction. The explanation is constituted of a Shapley value per input feature, which is proportional to how much each one contributed towards a “stressed” or “unstressed” prediction. By computing Shapley values for a sufficient sample of days, trends can be observed across the many explanations, and reveal the features with the largest effect on predictions globally.

Increasing actionability. Following the same procedure, the second XGBoost model was developed from a subset of 33 features (the “reduced model”). Certain features included in the full model (the one with 187 features) The inclusion of this feature also prevents more actionable features from being attributed higher Shapley values. Additionally, reducing the number of features would simplify the mental representation needed to understand the model, and facilitate integrating predictions in clinical decision-making. Since the 187 full features are the result of several aggregations of the original data presented in Table 1, a prediction explanation may contain multiple aggregations of the same feature (e.g., the mean and the minimum phone call duration). Different aggregations carry different information, but duplicate features could have diminishing actionability, and potentially make explanations less clear. Consequently, to increase the actionability of the model’s explanation, a subset of features was made by discarding features that appeared unactionable and keeping the aggregations that intuitively made sense (c.f. Table 2).

Table 2
Reduced model features

Feature type (*n* features)	Data type	Data collected
Physical (9)	Accelerometer	accelerometer (x, y, z) readings
Device use (5)	Applications, Battery	applications in use battery level battery currently plugged
Mobility (5)	GPS, Wi-Fi	location estimation accuracy Wi-Fi MAC addresses
Social (14)	Call logs, SMS logs, Bluetooth	incoming, outgoing, or missed (calls, SMS) call duration deidentified phone number (calls, SMS)

Results

Model performance

Full model. Once the full XGBoost model trained, it was presented with a test set composed of unseen examples (i.e., days of passive data) and it generated a prediction for each one. The *accuracy* of the model, or the percentage of days correctly classified as "stressed" or "unstressed", is 61.5%. The confusion matrix (Figure 2) displays the correct and incorrect classifications for each category, and presents the number of true positives, false positives, true negatives, and false negatives. The model was trained with equal weighting for both types of false results. Accordingly, the model did not display biased predictions on the test set. Based on statistical simulations from an independent peer-reviewed article (Combrisson & Jerbi, 2015), a binary classification accuracy above 58.2% for a sample size of 500 would be equivalent to a *p* value below 10^{-4}. With the test set containing 2914 days of data, a classification accuracy of 61.5% is significant. This performance is not high, but it still highly differs from chance level, which indicates the model learned predictive patterns within the data.

Figure 2 *Confusion matrix—Full model*

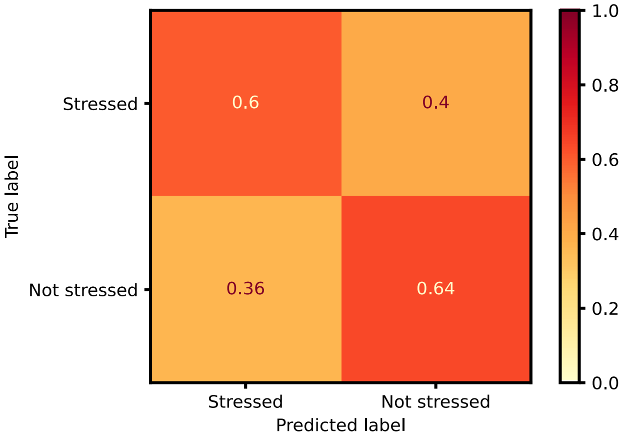
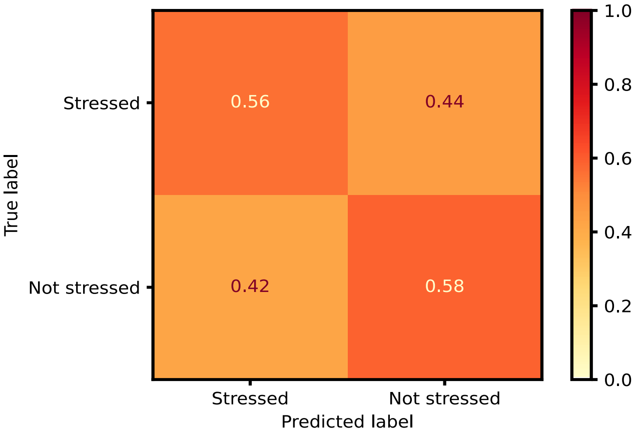


Figure 3 *Confusion matrix—Reduced model*

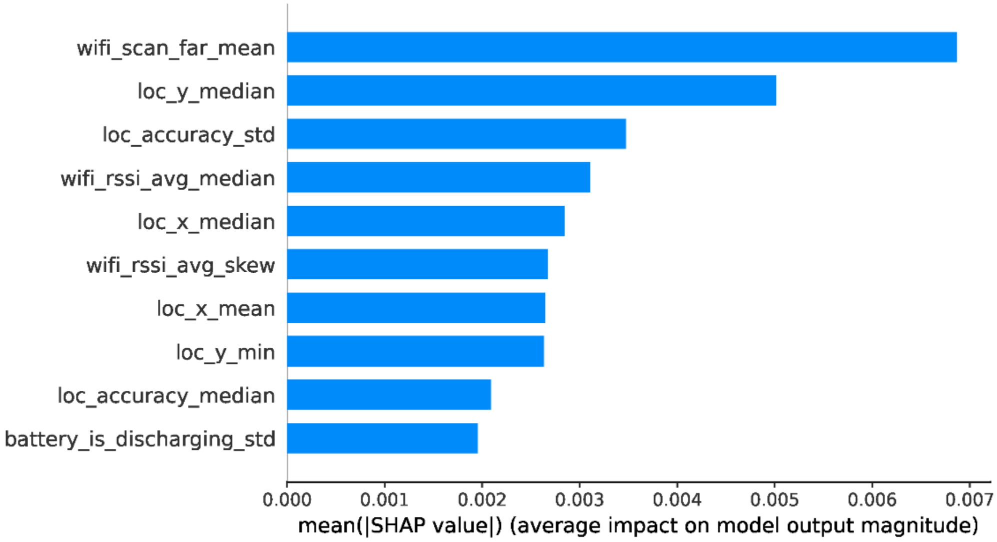


Reduced model. The reduced XGBoost model has an accuracy of 59.3%. The confusion matrix is presented in Figure 3. The reduced model did not present biased predictions either. Following the same reasoning used for the full model, a binary classification accuracy of 59.3% on a test set of 2914 days is significant, and far from chance level. The accuracy of the full and the reduced models are to be considered similar.

Model explanations

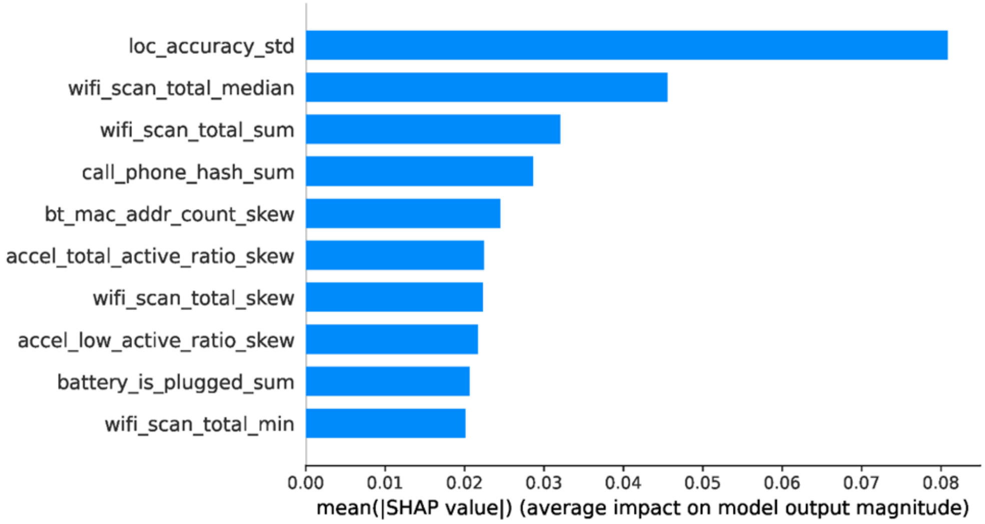
Full model. Using SHAP, prediction explanations were computed for the full model (c.f. Figure 4) The most important features for explanations are displayed in descending order. Out of the ten, the first nine features are related to mobility (GPS location x y coordinates and estimation accuracy, and Wi-Fi number of scanned devices, their distance, and the received signal strength). The tenth is device-related and indicates the rate of discharge of the battery, which increases when the device is used, and becomes negative when charging (i.e., it no longer discharges).

Figure 4 *Top explanations features—Full model*



Reduced model. The top ten features for prediction explanations of the reduced model are shown in Figure 5. Importantly, out of the full model's top feature, only the standard deviation of location accuracy (*loc_accuracy_std*) was included in the 33 features for the reduced model. This feature is ranked third in the full model and first in the reduced one. Otherwise, none of the 32 remaining handpicked actionable features from the reduced model are in the full model's top features. Five of the ten are related to mobility (GPS location estimation accuracy, and number of Wi-Fi devices scanned). While these are not all exactly the same from the full model, they constitute the entirety of the GPS and Wi-Fi data available to the reduced model. Two features are related to physical activity (ratios from accelerometer signal), and two others are social (number of phone calls with different contacts, and number of Bluetooth devices scanned). The feature for the number of Bluetooth devices scanned was the skewness aggregation. It most likely contributes to establishing mobility (or immobility) patterns as the number of surrounding devices would change from place to place. Also, one feature was device related (device being plugged). This feature is almost equivalent to the feature *battery discharging* included in the full model's top ten when its value is negative (i.e., the battery is charging). Having a device charging for a certain amount of time is probably indicative of mobility patterns too.

Figure 5 *Top explanations features—Reduced model*



Discussion

Using digital phenotyping data passively acquired from participants' smartphone, two forecast models of participants' next-day stress levels were created. Both are of type XGBoost and followed the same development procedure. The full model used 187 features, which were the product of indiscriminate aggregations of the original data. The reduced model was trained on a subset of 33 handpicked features to increase actionability. Higher actionability would be achieved by removing features that appeared irrelevant for clinical decision-making and resulted from redundant or unactionable aggregations. Both models showed significant predictive performance. Subsequently, Shapley values were used to formulate prediction explanations and identify the most important features of each model.

The predictive performances of the two models are situated in the low-to-mid range of accuracy (from 55 to 75%) for similar tasks reported by Jaques (2020). An accuracy of 60% remains low but reaffirms the viability and potential of digital phenotyping data as a basis for predicting dynamic constructs like stress. Important performance improvement will be needed for such tools to be used in clinical settings. Nonetheless, the predictive performance is not everything. One must consider how a clinician would incorporate this new information in decision-making, and for what decisions. For example, using a prediction to guide clinical interviews, versus to pose a diagnostic, can lead to consequences of different magnitudes. Pragmatically, future models value should be assessed by comparing its predictions to previous predictive methods (Paulus, 2017). By observing how many predictions differ between the two, and assessing if these changes are globally beneficial, the model can be judged on the incremental benefits of using it, in addition to or instead of previous methods.

Valuable insights can be derived from the top explanations of the two models. While the full model almost exclusively identified mobility features (GPS and Wi-Fi) as the most important, the reduced model included the five mobility features (also GPS and Wi-Fi) out of 33 in its top ten. Out of the five remaining, two (Bluetooth and battery metadata probably) had a mobility component to them. These results are coherent with the significant correlations between mood symptoms and mobility patterns highlighted by Rohani et al. (2018). From the 33 features of the reduced model, 14 were classified as social, but only two were part of the top ten. This is surprising considering the theoretical link between stress and social activity (Cacioppo, 1994), but remains in line with the non-significant correlations between mood symptoms and phone and SMS data identified by Rohani et al. (2018). Additionally, it suggests that the same data types and feature categories remain important to predictions even after reducing the number of features from 187 to 33. This probably indicates that more distinctive mobility patterns are linked to “stressed” and “unstressed” days than for other feature categories. The GPS and Wi-Fi data types may provide particularly adapted resolutions to capture those patterns as opposed to call and SMS logs for the social dimension. Investigating explanatory frameworks relating daily mobility patterns to stress would contribute to the actionability of predictive models using those input features.

Limitations and future Work

In the context of this Honor research, XGBoost models were developed as they display top performance for a variety of tasks and often serve as a baseline when building more complicated machine learning models. Nonetheless, a known clear limitation of XGBoost models is their inability to fully exploit the data’s temporal dimension. Indeed, this type of model learns a mapping between the passive data of a day (features) and the next-day stress level (label) from numerous examples. Each day is considered independent, but in practice temporal patterns are to be expected both in the passive data (e.g., a day with missed call might be followed by a day with an increased number of calls) and self-assessments (e.g., a person's stress might increase as they approach a deadline). Relying on distinct "learning mechanisms", recurrent neural networks (RNN) are particularly proficient at capturing the temporal dependencies from such densely longitudinal data (Durstewitz et al., 2019; Koppe et al., 2019). Consequently, the next logical step is to develop RNN models, as they are expected to outperform XGBoost models for forecasting based on digital phenotyping data. Some of the current best performance was produced by RNNs (Umematsu et al., 2019a; Umematsu et al., 2019b)

The actionability of the model's explanations appears to be primarily limited by the input features (i.e., the digital phenotyping data used). Therefore, there is a need for *high-level* input features. Instead of providing the phone's x-axis accelerometer signal, or its battery temperature to the model, new features integrating multiple type of data can be engineered (Mohr et al., 2017). For example, previous research established that sleep duration and schedule can be inferred from location, ambient lighting, phone movement, and phone usage (Chen et al., 2013). Such insight would not be available from inspecting each sensor individually. Since a prediction explanation is constituted of a Shapley value per input feature, models built from high-level features would provide more actionable explanations. Nonetheless, a known trade-off exists between performance and the "meaningfulness" (i.e., actionability) of machine learning models (Bologna & Hayashi, 2017; Jin & Sendhoff, 2008). The people engineering high-level features cannot fully comprehend or anticipate the patterns machine learn algorithms find within big data. Then, any data transformation applied to increase actionability is likely to diminish the richness of the data from which models can learn. However, independently of predictive endeavours, high-level features themselves can be valuable sources of information. In particular, Fulford et al. (2020) showed that indicators built exclusively from passive data could serve as statistically significant proxies for clinical scale used for schizophrenia.

By implicating clinicians, care receivers, and other stakeholders, high-level features could be designed to carry information they judged relevant. In two separate studies, the majority of patients with an anxiety or mood disorder expressed interest in using a mobile digital phenotyping application for clinical use in the mental health field. They were unsurprisingly more cautious about collecting passive sensor data and usage metadata from the smartphone (Di Matteo et al., 2018; Nicholas et al., 2019). The different forms of sensors are also not equal: a third of patients said they were not open to allowing the recording of GPS data or motion sensors, while just over half were opposed to allowing access to audio data or the content of text messages (Di Matteo et al., 2018). Patients are thus more comfortable sharing data related to health information (physical activity, sleep, emotion, etc.) than data deemed more personal (communication, social activity, etc.; Nicholas et al., 2019). The main concern being related to data security and confidentiality, it can be argued that a higher proportion of patients would be inclined to use digital phenotyping tools if it is guaranteed that the most sensitive raw data are not shared as is, by integrating the transformation of raw sensor measurements into higher-level features on the phone itself (Di Matteo et al., 2018).

Conclusion

This study supports the feasibility of interpretable forecast models based on data passively acquired from smartphones. Using the open-source *Friends and Family* dataset, two eXtreme Gradient Boosting (XGBoost) models were developed for predicting next-day stress. A full model included 187 features, and a reduced one included 33 features handpicked for their actionability. Both showed a test accuracy significantly above chance at around 60%. The implementation of

SHapley Additive exPlanations (SHAP) highlighted the potential of this tool for providing explanations for predicted outcomes. Mobility features showed to be particularly important to explain predictions, and social features of low importance. Increasing the interpretability and actionability of predictive models will allow mental health practitioners, care receivers, and other stakeholders to be critical both of individual predictions and broader model behaviour. This is an important step forward for clinical applications of digital phenotyping tools. Future work should consider training more powerful deep learning models to learn from potential temporal patterns in the data. Also, training models exclusively on actionable high-level features would increase actionability and help understand potential trade-offs with predictive performance.

References

Abdullah, S., & Choudhury, T. (2018). Sensing Technologies for Monitoring Serious Mental Illnesses. *IEEE MultiMedia*, *25*(1), 61–75. https://doi.org/10.1109/MMUL.2018.011921236

Aharony, N., Pan, W., Ip, C., Khayal, I., & Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing, 7*(6), 643–659. https://doi.org/10.1016/j.pmcj.2011.09.004

American Psychiatric Association. (2015). *The American Psychiatric Association Practice Guidelines for the Psychiatric Evaluation of Adults* (Third Edition). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890426760

Asselbergs, J., Ruwaard, J., Ejdys, M., Schrader, N., Sijbrandij, M., & Riper, H. (2016). Mobile Phone-Based Unobtrusive Ecological Momentary Assessment of Day-to-Day Mood: An Explorative Study. *Journal of Medical Internet Research*, *18*(3), e72. https://doi.org/10.2196/jmir.5505

Beiwinkel, T., Kindermann, S., Maier, A., Kerl, C., Moock, J., Barbian, G., & Rössler, W. (2016). Using Smartphones to Monitor Bipolar Disorder Symptoms: A Pilot Study. *JMIR Mental Health*, *3*(1), e2. https://doi.org/10.2196/mental.4560

Ben-Zeev, D., Brian, R., Wang, R., Wang, W., Campbell, A. T., Aung, M. S. H., Merrill, M., Tseng, V. W. S., Choudhury, T., Hauser, M., Kane, J. M., & Scherer, E. A. (2017). CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric Rehabilitation Journal*, *40*(3), 266–275. https://doi.org/10.1037/prj0000243

Birk, R., & Samuel, G. (2020). Can digital data diagnose mental health problems? A sociological exploration of "digital phenotyping." *Sociology of Health & Illness*, *42*(8), 1873—1887. https://doi.org/10.1111/1467-9566.13175

Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., Arenare, E., R. Van Meter, A., De Choudhury, M., & Kane, J. M. (2019). Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. *Npj Schizophrenia*, *5*(1), 1—9. https://doi.org/10.1038/s41537-019-0085-9

Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., Alex, & Pentland. (2014). Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. *MM 2014—Proceedings of the 2014 ACM Conference on Multimedia*. https://doi.org/10.1145/2647868.2654933

Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., Pentland, S., Fondazione, B., & Kessler. (2014, March 24). *Pervasive Stress Recognition for Sustainable Living*. 2014 IEEE International

Conference on Pervasive Computing and Communication Workshops, PERCOM WORKSHOPS 2014. https://doi.org/10.1109/PerComW.2014.6815230

Bologna, G., & Hayashi, Y. (2017). Characterization of Symbolic Rules Embedded in Deep DIMLP Networks: A Challenge to Transparency of Deep Learning. *Journal of Artificial Intelligence and Soft Computing Research*, *7*(4), 265–286. https://doi.org/10.1515/jaiscr-2017-0019

Cacioppo, J. T. (1994). Social neuroscience: Autonomic, neuroendocrine, and immune responses to stress. *Psychophysiology, 31*(2), 113–128. https://doi.org/10.1111/j.1469-8986.1994.tb01032.x

Canzian, L., & Musolesi, M. (2015). Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing—UbiComp '15*, 1293–1304. https://doi.org/10.1145/2750858.2805845

Chen, Z., Lin, M., Chen, F., Lane, N., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T., & Cambell, A. (2013, May 23). *Unobtrusive Sleep Monitoring using Smartphones*. 7th International Conference on Pervasive Computing Technologies for Healthcare. https://eudl.eu/doi/10.4108/icst.pervasivehealth.2013.252148

Chiu, M., Gatov, E., Vigod, S. N., Amartey, A., Saunders, N. R., Yao, Z., Pequeno, P., & Kurdyak, P. (2018). Temporal Trends in Mental Health Service Utilization across Outpatient and Acute Care Sectors: A Population-Based Study from 2006 to 2014. *The Canadian Journal of Psychiatry*, *63*(2), 94–102. https://doi.org/10.1177/0706743717748926

Cohen, A. S. (2019). Advancing ambulatory biobehavioral technologies beyond "proof of concept": Introduction to the special section. *Psychological Assessment*, *31*(3), 277—284. https://doi.org/10.1037/pas0000694

Cohen, A. S., Fedechko, T., Schwartz, E. K., Le, T. P., Foltz, P. W., Bernstein, J., Cheng, J., Rosenfeld, E., & Elvevåg, B. (2019). Psychiatric Risk Assessment from the Clinician's Perspective: Lessons for the Future. *Community Mental Health Journal, 55*(7), 1165–1172. https://doi.org/10.1007/s10597-019-00411-x

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods, 250,* 126–136. https://doi.org/10.1016/j.jneumeth.2015.01.010

Di Matteo, D., Fine, A., Fotinos, K., Rose, J., & Katzman, M. (2018). Patient Willingness to Consent to Mobile Phone Data Collection for Mental Health Apps: Structured Questionnaire. *JMIR Mental Health*, *5*(3), e9539. https://doi.org/10.2196/mental.9539

Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry*, *24*(11), 1583–1598. https://doi.org/10.1038/s41380-019-0365-9

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Eagle, N., & (Sandy) Pentland, A. (2006). Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, *10*(4), 255–268. https://doi.org/10.1007/s00779-005-0046-3

Ericsson Mobility Report November 2020. (2020). 36.

Farago, P. (2012, October 22). *App Engagement: The Matrix Reloaded | Flurry*. https://www.flurry.com/blog/app-engagement-the-matrix-reloaded/

Fulford, D., Mote, J., Gonzalez, R., Abplanalp, S., Zhang, Y., Luckenbaugh, J., Onnela, J.-P., Busso, C., & Gard, D. E. (2020). Smartphone sensing of social interactions in people with and without schizophrenia. *Journal of Psychiatric Research*, S002239562031058X. https://doi.org/10.1016/j.jpsychires.2020.11.002

Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., & Tørresen, J. (2018). Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, *51*, 1—26. https://doi.org/10.1016/j.pmcj.2018.09.003

Gershon, A., Ram, N., Johnson, S. L., Harvey, A. G., & Zeitzer, J. M. (2016). Daily Actigraphy Profiles Distinguish Depressive and Interepisode States in Bipolar Disorder. *Clinical Psychological Science*, *4*(4), 641—650. https://doi.org/10.1177/2167702615604613

Grunerbl, A., Muaremi, A., Osmani, V., Bahle, G., Ohler, S., Troster, G., Mayora, O., Haring, C., & Lukowicz, P. (2015). Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients. *IEEE Journal of Biomedical and Health Informatics*, *19*(1), 140–148. https://doi.org/10.1109/JBHI.2014.2343154

Insel, T. R. (2017). Digital Phenotyping: Technology for a New Science of Behavior. *JAMA*, *318* (13), 1215. https://doi.org/10.1001/jama.2017.11295

Jain, S. H., Powers, B. W., Hawkins, J. B., & Brownstein, J. S. (2015). The digital phenotype. *Nature Biotechnology*, *33*(5), 462—463. https://doi.org/10.1038/nbt.3223

Jaques, N. (2020). *Towards Social and Affective Machine Learning,*. https://www.youtube.com/watch?v=P4-8wd-t9mg

Jin, Y., & Sendhoff, B. (2008). Pareto-Based Multiobjective Machine Learning: An Overview and Case Studies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(3), 397–415. https://doi.org/10.1109/TSMCC.2008.919172

Kendler, K. S., & Engstrom, E. J. (2017). Criticisms of Kraepelin's Psychiatric Nosology: 1896–1927. *American Journal of Psychiatry*, *175* (4), 316—326. https://doi.org/10.1176/appi.ajp.2017.17070730

Koppe, G., Guloksuz, S., Reininghaus, U., & Durstewitz, D. (2019). Recurrent Neural Networks in Mobile Sampling and Intervention. *Schizophrenia Bulletin*, *45*(2), 272—276. https://doi.org/10.1093/schbul/sby171

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). Explainable AI for Trees: From Local Explanations to Global Understanding. *ArXiv:1905.04610 [Cs, Stat]*. http://arxiv.org/abs/1905.04610

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. 10.

Madan, A., Cebrian, M., Moturu, S., Farrahi, K., & Pentland, A. "Sandy." (2012). Sensing the "Health State" of a Community. *IEEE Pervasive Computing*, *11*(4), 36–45. https://doi.org/10.1109/MPRV.2011.79

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology*, *13*(1), 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116* (44), 22,071—22080. https://doi.org/10.1073/pnas.1900654116

Nicholas, J., Shilton, K., Schueller, S. M., Gray, E. L., Kwasny, M. J., & Mohr, D. C. (2019). The Role of Data Type and Recipient in Individuals' Perspectives on Sharing Passively Collected Smartphone Data for Mental Health: Cross-Sectional Questionnaire Study. *JMIR MHealth and UHealth*, *7*(4), e12578. https://doi.org/10.2196/12578

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology*, *41*(7), 1691–1696. https://doi.org/10.1038/npp.2016.7

Paulus, M. P. (2017). Evidence-Based Pragmatic Psychiatry-A Call to Action. *JAMA Psychiatry*, *74*(12), 1185—1186. https://doi.org/10.1001/jamapsychiatry.2017.2439

Rogler, L. H., Mroczek, D. K., Fellows, M., & Loftus, S. T. (2001). The Neglect of Response Bias in Mental Health Research. *The Journal of Nervous and Mental Disease*, *189* (3), 182–187.

Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2018). Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients With Affective Disorders: Systematic Review. *JMIR MHealth and UHealth*, *6*(8), e9691. https://doi.org/10.2196/mhealth.9691

Sapiezynski, P., Stopczynski, A., Lassen, D. D., & Lehmann, S. (2019). Interaction data from the Copenhagen Networks Study. *Scientific Data*, *6*(1), 315. https://doi.org/10.1038/s41597-019-0325-x

Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., & Rentfrow, J. (2019). Sequence Multi-task Learning to Forecast Mental Wellbeing from Sparse Self-reported Data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2886–2894. https://doi.org/10.1145/3292500.3330730

Torous, J., & Baker, J. T. (2016). Why Psychiatry Needs Data Science and Data Science Needs Psychiatry: Connecting With Technology. *JAMA Psychiatry*, *73*(1), 3. https://doi.org/10.1001/jamapsychiatry.2015.2622

Torous, J., Gershon, A., Hays, R., Onnela, J.-P., & Baker, J. T. (2019). Digital Phenotyping for the Busy Psychiatrist: Clinical Implications and Relevance. *Psychiatric Annals*, *49*(5), 196–201. https://doi.org/10.3928/00485713-20190417-01

Umematsu, T., Sano, A., & Picard, R. W. (2019). Daytime Data and LSTM can Forecast Tomorrow’s Stress, Health, and Happiness. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2186––2190. https://doi.org/10.1109/EMBC.2019.8856862

Umematsu, T., Sano, A., Taylor, S., & Picard, R. W. (2019). Improving Students’ Daily Life Stress Forecasting using LSTM Neural Networks. *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 1–4. https://doi.org/10.1109/BHI.2019.8834624

Vaizman, Y. (2018). *Behavioral Context Recognition In the Wild* [University of California]. https://escholarship.org/uc/item/200910xx

Vaizman, Y., Ellis, K., & Lanckriet, G. (2017). Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches. *IEEE Pervasive Computing*, *16*(4), 62–74. https://doi.org/10.1109/MPRV.2017.3971131

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 3–14. https://doi.org/10.1145/2632048.2632054

Wang, W., Mirjafari, S., Harari, G., Ben-Zeev, D., Brian, R., Choudhury, T., Hauser, M., Kane, J., Masaba, K., Nepal, S., Sano, A., Scherer, E., Tseng, V., Wang, R., Wen, H., Wu, J., & Campbell, A. (2020). Social Sensing: Assessing Social Functioning of Patients Living with Schizophrenia using Mobile Phone Sensing. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3313831.3376855

Wichers, M., Wigman, J. T. W., & Myin-Germeys, I. (2015). Micro-Level Affect Dynamics in Psychopathology Viewed From Complex Dynamical System Theory. *Emotion Review*, *7*(4), 362—367. https://doi.org/10.1177/1754073915590623

Wichers, Marieke, Schreuder, M. J., Goekoop, R., & Groen, R. N. (2019). Can we predict the direction of sudden shifts in symptoms? Transdiagnostic implications from a complex systems perspective on psychopathology. *Psychological Medicine*, *49*(3), 380–387. https://doi.org/10.1017/S0033291718002064

Wright, A. G. C., & Woods, W. C. (2020). *Personalized Models of Psychopathology*. 29.

Zulueta, J., Leow, A. D., & Ajilore, O. (2020). Real-Time Monitoring: A Key Element in Personalized Health and Precision Health. *FOCUS*, *18*(2), 175–180. https://doi.org/10.1176/appi.focus.20190042

APPENDIX 1

Machine Learning Glossary

Term	**Definition**
Machine learning	A program that builds a predictive model
Model	A system that learns from training data. A trained model can receive *never-seen-before* data to formulate predictions about it
Prediction	The output of a model for a given input
Example	One row of a dataset. It contains multiple features and a label. (e.g., each day of data is an example)
Feature	An input variable given to a model to make a prediction (e.g., each passive data measurement is a feature)
Label	The correct “answer” or prediction to make for a given example. Each example in the dataset has a label.
Training	The process done by the program to learn the parameters of the model resulting in the best performance
Training set	Subset of examples from a dataset used to train a model
Test set	Never-seen-before subset of examples from a dataset used to test the performance of a trained model
Loss	A measure of how far a prediction is from its label
Objective	A metric the program is trying to optimize. It formalizes what a “performant” model is
Parameter	A variable the model learns on its own during training (e.g., a regression coefficient)
Hyperparameter	A value that can be “tuned” to modify how the model learns or generally functions. Different hyperparameter values must be tried to find the best performing model. This tuning can be done manually or programmatically

Note. The glossary is adapted from Google’s *Machine Learning Glossary.*
Retrieved from https://developers.google.com/machine-learning/glossary