Predicting Subjective Measures of Social Anxiety from Sparsely Collected Mobile Sensor Data

HAROON RASHID, University of Virginia, USA
SANJANA MENDU, University of Virginia, USA
KATHARINE E. DANIEL, University of Virginia, USA
MIRANDA L. BELTZER, University of Virginia, USA
BETHANY A. TEACHMAN, University of Virginia, USA
MEHDI BOUKHECHBA, University of Virginia, USA
LAURA E. BARNES, University of Virginia, USA

Exploiting the capabilities of smartphones for monitoring social anxiety shows promise for advancing our ability to both identify indicators of and treat social anxiety in natural settings. Smart devices allow researchers to collect passive data unobtrusively through built-in sensors and active data using subjective, self-report measures with Ecological Momentary Assessment (EMA) studies. Prior work has established the potential to predict subjective measures from passive data. However, the majority of the past work on social anxiety has focused on a limited subset of self-reported measures. Furthermore, the data collected in real-world studies often results in numerous missing values in one or more data streams, which ultimately reduces the usable data for analysis and limits the potential of machine learning algorithms. We explore several approaches for addressing these problems in a smartphone based monitoring and intervention study of eighty socially anxious participants over a five week period. Our work complements and extends prior work in two directions: (i) we show the predictability of seven different self-reported dimensions of social anxiety, and (ii) we explore four imputation methods to handle missing data and evaluate their effectiveness in the prediction of subjective measures from the passive data. Our evaluation shows imputation of missing data reduces prediction error by as much as 22%. We discuss the implications of these results for future research.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

Additional Key Words and Phrases: Social anxiety, mental health, mobile sensing, data imputation, ecological momentary assessment

ACM Reference Format:
Haroon Rashid, Sanjana Mendu, Katharine E. Daniel, Miranda L. Beltzer, Bethany A. Teachman, Mehdi Boukhechba, and Laura E. Barnes. 2020. Predicting Subjective Measures of Social Anxiety from Sparsely Collected Mobile Sensor Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 109 (September 2020), 24 pages. https://doi.org/10.1145/3411823

Authors' addresses: Haroon Rashid, University of Virginia, Charlottesville, VA-22904, USA, hl7ck@virginia.edu; Sanjana Mendu, University of Virginia, Charlottesville, USA, sm7gc@virginia.edu; Katharine E. Daniel, University of Virginia, Charlottesville, USA, ked4fd@virginia.edu; Miranda L. Beltzer, University of Virginia, Charlottesville, USA, beltzer@virginia.edu; Bethany A. Teachman, University of Virginia, Charlottesville, USA, bteachman@virginia.edu; Mehdi Boukhechba, University of Virginia, Charlottesville, USA, mob3f@virginia.edu; Laura E. Barnes, University of Virginia, Charlottesville, USA, lb3dp@virginia.edu.

2474-9567/2020/9-ART109 $15.00
https://doi.org/10.1145/3411823

1 INTRODUCTION

Social anxiety disorder (SAD) affects approximately 15 million American adults at any given time [1], around 7% of the adult population in the U.S. [37]. SAD is characterized by intense fear and avoidance of socially evaluative situations, and is associated with both academic challenges and negative developmental trajectories [40, 71, 74]. Currently, clinically validated methods for SAD detection heavily depend upon retrospective self-reports and questionnaires conducted in a laboratory setting. These approaches require patients to be motivated to seek out assessment opportunities and the approach is prone to recall bias. Further, the approach typically only provides a single snapshot of the behavior of interest, rather than permitting dynamic tracking and numerous samples of behavior. It is thus important to investigate alternative approaches to identify, monitor, and treat social anxiety.

The ubiquity of mobile technology affords promising opportunities to unobtrusively monitor SAD symptoms. For example, using mobile health (mHealth) technology has enabled researchers and clinicians to observe the manifestation of SAD in natural settings, rather than only in clinical or lab-based settings [24, 33]. Apart from passive data, mHealth technologies can be used to understand real-time, naturalistic experiences via Ecological Momentary Assessment (EMA) [68]. As an important step towards understanding real-time experiences of social anxiety in a diagnosed sample, we use data from a sample with elevated symptoms of social anxiety to test relationships between passive and self-reported social anxiety measures via smartphone technologies. Specifically, our first research question aims to address whether passive data be used to predict daily self-reported experiences related to social anxiety (e.g., negative affect and avoidance of social situations)?

Using smartphone technology, we can capture fine-grained measures related to human behavior, allowing researchers to passively monitor behavioral markers that correlate with individuals' mental health state, reducing the burden of repeated subjective measures. However, in practice, we rarely get continuous uninterrupted data from mobile devices due to a variety of factors (e.g., user switches off some data streams, some data streams are not collected in the background, diverse hardware-contingent policies). As a result, mobile sensor data often suffers from "missingness", thus reducing the effectiveness of many popular machine learning algorithms that researchers would otherwise like to apply to those sparse data. Thus, our second research question concerns how can we leverage data to predict self-reported experiences of social anxiety despite of missingness?

In this paper, we address our research questions by demonstrating the performance of machine learning algorithms in predicting self-reported measures from the passive data using diverse imputation methods. Our specific research questions are as follows:

RQ1: Can passive data be leveraged to predict daily self-reported experiences related to social anxiety (e.g., negative affect and avoidance of social situations)?

RQ2: How can we leverage data to predict self-reported experiences of social anxiety despite missingness?

First, we will provide important framing for the problem of predicting subjective EMA measures using passive behavioral measures. Then, we will outline common imputation approaches used across mobile sensing and other fields and evaluating models on both imputed and non-imputed data to assess the performance when using specific methods. Finally, we will discuss the implications of predicting subjective measures of social anxiety by using the passively sensed data. We will evaluate our work using data on 80 socially anxious participants using both passive data and subjective measures across a five-week period.

2 RELATED WORK

2.1 Mobile Sensing for Mental Health: From Biomarkers to Behaviors

Mobile sensing has become a promising avenue for collecting real-time, objective assessments in the context of mental health due to the cost efficiency and increasing ubiquity of smart devices (e.g. smartphones, wearables) [30, 39, 48]. Embedded sensors in technologies like smartphones and wearable devices can be harnessed to passively capture information related to users' personal and environmental factors (e.g., current location as

indicated by GPS) and behaviors (e.g., movements as indicated by accelerometer metrics) [10], reducing the burden of self-report measures on participants. Studies have shown that behavioral data captured with mobile sensing techniques are associated with psychiatric symptoms among individuals with mental illness and behavioral disorders [48], including depression [2, 3, 15, 17, 19, 20, 55, 57, 60, 65, 79, 81], schizophrenia [9, 76], bipolar disorder [7, 21], and post-traumatic stress disorder [55, 57].

The majority of existing work on mobile sensing for anxiety has focused on predicting clinically validated measures for assessing levels of anxiety at a single time point (e.g., assessing anxiety reported on one occasion on a trait questionnaire)[12–14, 19, 20, 26, 47]. However, these measures are limited because they are collected infrequently (usually only once) and do not reveal temporal variations in individuals' experience of anxiety symptoms. These retrospective trait measures are also subject to recall bias [23, 68], social desirability bias [23, 75], and limited self-knowledge [52]. EMA has emerged as a popular method for conducting *in-situ* experience sampling to monitor and understand mental health status in real time in the real world. Existing work on mobile sensing towards predicting momentary mental health states has shown promising results in the context of mood instability [50, 61], depressive moods [17, 46, 79], general mood [4, 34], and affect [77, 82]. By prompting individuals to assess disorder-relevant behaviors in daily life, EMA allows researchers to collect a more complete depiction of an individual's symptom changes throughout their lived experience, compared to relying on retrospective questionnaires alone. Further, EMA allows us to collect information on multiple aspects of a person's subjective experience of anxiety, painting a more complete, personalized picture of an individual's experience. For example, one socially anxious person might be willing to enter into social experiences but not enjoy them when they do, whereas another socially anxious person might avoid social experiences altogether. As such, assessing and predicting multiple subjective facets of social anxiety, both within and across individuals, holds great promise for researchers and clinicians hoping to better understand and treat SAD [22, 35, 49].

Although EMA affords several advantages for collecting in-situ data, it also introduces a number of disadvantages [67]. Due to the frequent nature of EMA prompts, participants are likely to experience high levels of response burden [72]. This poses a trade-off between the level of granularity of collected data and the amount of data burden placed on participants. EMA-based studies are further limited by participants' low compliance rates in response to potentially overwhelming data burden. Heron et al. conducted a survey of EMA-based studies and found that the survey completion rate was only 76% [32]. Even when response rates are good, this does not imply that the collected responses are of good quality [18]. Thus, there is a need to capitalize on the benefits of EMA and passive data collection while exploring new methods to reduce the disadvantages.

2.2 Imputation Methods for Passive Sensing Applications

While mobile sensors are a promising technology for continuous behavior assessment, these devices might not function as expected in the presence of many unpredictable factors, including: (1) participants turning off sensors due to excessive battery drainage or privacy concerns; (2) participants forgetting to charge or wear sensors; (3) sensors breaking or the signals becoming noisy; and (4) mobile phone connectivity, hardware sensor functionality, and mobile software updates, which can interfere with data integrity [64]. The resulting incomplete values have to be processed and approximated in order for further data processing to be more reliable and valid.

In the context of mental health applications, researchers have used a variety of approaches to deal with incomplete data. Wang et al. used a time-based threshold to mitigate the effects of missing data on resulting time series features [78]. Chow et al. attributed missing data to users turning off their phone or shutting down the app and used the last observation carried forward method to impute missing observations [20]. Sano et al. used an automated classifier to separate clean and noisy data epochs for further analysis. [64]. However, few studies have assessed the impact of their chosen imputation method on the resulting findings and model performance.

Researchers across diverse fields have explored measures to mitigate the effects of missing mobile sensor data. Past work can largely be categorized as Matrix Completion-based, interpolation-based, or regression-based methods [83]. Matrix Completion-based methods attempt to derive a low-PCA matrix from a small number of samples [16, 36, 73]. In general, these methods operate by under-sampling high-dimensional signals and accurately reconstructing them by exploiting hidden structures in the underlying data. Mazmuder et al. proposed a Matrix Completion-based method that leverages convex relaxation techniques to solve large-scale Matrix Completion problems [44]. This simple and efficient algorithm uses an iterative approach to replace missing entries in the data with those obtained from a soft-thresholded singular vector decomposition. By using the nuclear norm as a regularizer, the algorithm minimizes the matrix reconstruction error subject to a bound on the nuclear norm.

Alternatively, interpolation-based methods aim to find a smooth way to fill missing data points between values. Instead of removing rows and columns containing missing values, interpolation-based methods retain all the information in the original dataset, and replace missing values with a designated placeholder value (e.g., "0"s or the mean of other values). Reza et al. proposed an interpolation-based imputation algorithm which searches the sequential dataset to find data segments that have a prior and posterior segment that matches those of the missing data to use as a substitute [58]. Finally, regression-based methods are designed to predict observed values of a target variable based on other variables available in the dataset. The fitted regression model is then used to impute values in cases where the value of the target variable is missing. Specifically in the context of mobile sensing data, many researchers have leveraged regression models for imputation for hierarchical data [42], participatory sensing data [38], and mixed-attribute data (i.e., both continuous and discrete data streams) [84].

Researchers have also explored multiple-step imputation based methods [69, 70], first imputing missing values multiple times (m times) to generate m datasets, and then averaging the imputed values at the same data point of all generated datasets to get a final integrated value. Multiple Imputations by Chained Equations (MICE) [6] is a well known statistical method that has previously been used in activity recognition [54] and clinical research applications [59]. MICE is a particularly promising imputation method in the context of mobile sensing due to its ability to handle diverse variable types (e.g., continuous, binary, categorical) due to the independence of variable imputation models [80]. MICE has also been shown to scale well to larger datasets, with hundreds and thousands of observations [27, 66].

In this paper, we evaluate imputation methods from each of the mentioned categories (Matrix Completion-based, interpolation-based, or regression-based) on a dataset collected during a five-week study monitoring a high social anxiety sample. We compare a variety of modeling approaches using both imputed and non-imputed versions of the same dataset to examine the impact of these methods on our ability to predict time-variable measures of mental health.

Table 1. Demographic information

(a) Gender identity reported by sample

Gender	#
Male	20
Female	60

(b) Race reported by sample

Race	#
Caucasian	61
Asian	14
African American	5

(c) Status distribution

Status	#
Undergraduate students	62
Graduate students	8
Local community	10

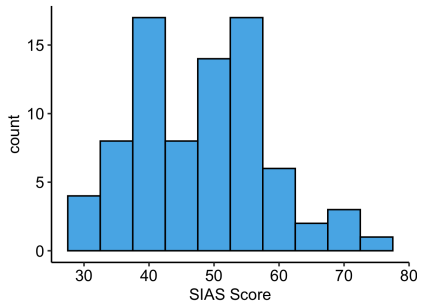


Fig. 1. SIAS score histogram for all participants. Bin width is set to 5

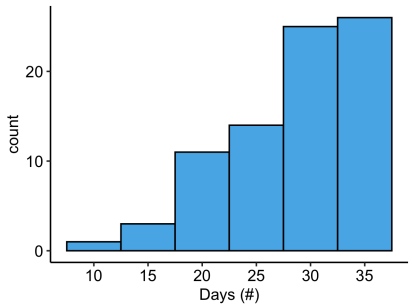


Fig. 2. Distribution of # of days of data submitted by participants. Bin width is 5.

3 STUDY DESIGN

We conducted a five-week study at the University of Virginia with 114 participants. The study was approved by the university institutional review board. Participant eligibility was based on scoring above 29 on the Social Interaction Anxiety Scale (SIAS) [43], indicating moderate to severe social anxiety symptoms (higher SIAS scores indicate greater symptom severity).

34 out of the 114 total participants were dropped from the current analysis due to low compliance with the subjective measure surveys (i.e. less than 50% of expected responses; 17 days minimum threshold). Table 1 shows the demographic information and Figure 1 shows the distribution of SIAS scores for the remaining N = 80 participants (M_{age} = 20.51, SD_{age} = 3.18). Figure 2 shows the distribution of the number of days of data submitted by these 80 participants.

Participants received either $25 or 1.5 credit hours for participating in a baseline (in-lab) session (as part of the larger study). The payment for the five-week EMA monitoring period was calculated based on the percentage of available surveys that were completed, for a maximum of $70. In the baseline session, participants completed various trait inventories, including measures of interpretation bias, cognitive reappraisal tendencies, and social anxiety symptoms. Additionally, a third-party mobile application, MetricWire[1], was installed on the personal smartphones of participants.

4 DATA

4.1 Passive Data

The MetricWire smartphone application passively recorded the following four data streams: (1) GPS; (2) Pedometer; (3) Accelerometer; (4) Activity; (5) Call; (6) Text. Activity states (e.g., Stationary, Walking, Running, Automotive, Cycling) were computed using a built-in classifier in MetricWire on collected raw sensor data. Accelerometer data were collected at a sampling frequency of 1 Hz, whereas GPS and Pedometer data collection was event-driven (i.e., new data samples were collected only if a change in the location was detected). Phone call and Text data were downloaded at the end of the 5-week study from the participants' smartphones using the iMazing[2] application.

To effectively characterize participants' daily behaviors in terms of these passive data streams, we calculated a large number of features at day level as shown in Table 2. For pedometer, MetricWire collected start/end timestamps, number of steps taken, distance covered, pace (in seconds per meter), cadence (in steps per second). Using this information we computed other statistical (average, min, max, std) features. For accelerometer,

[1] https://metricwire.com/
[2] https://imazing.com/

Table 2. Passive data streams and computed features at day level. Features labelled with superscript * were dropped during modeling as they were collinear with the remaining features of data streams.

Data stream	Features
Pedometer	Steps (total, avg, min*, max*, std*), Pace (avg, std, min*, max*), Cadence (avg, std, min*, max*), Activeminutes (total, avg*, min*, max*, std*), distance (total*, avg*, min*, max*, std*)
Accelerometer	Magnitude (average, min, kurtosis, max*, std*, median*, skew*, energy*), Signal Magnitude Area* (SMA), Signal Vector Magnitude* (SVM)
GPS	Number of places visited, entropy, proportion of day time spent at home, proportion of day time spent at other places
Activity	Stationary proportion, walking proportion, running proportion, automotive proportion, cycling proportion, number of records
Call	Count of incoming/outgoing calls, average duration of incoming/outgoing calls, number of unique contacts contacted
Text	Count of incoming/outgoing calls, number of unique contacts texted

MetricWire collected timestamps, X, Y, and Z coordinates. With X, Y, and Z, we computed Magnitude, Signal Magnitude Area (SMA), Signal Vector Magnitude (SVM) and other statistical (average, min, kurtosis, max, std, median, skew, energy) features. Signal Magnitude, SMA, and SVM were computed as

$$Magnitude = \sqrt{X^2 + Y^2 + Z^2}$$
$$SMA = |X| + |Y| + |Z|$$
$$SVM = X^2 + Y^2 + Z^2$$

For GPS, MetricWire provided timestamps and location coordinates (latitude, longitude). We then computed location based features such as location entropy and cummulative staying time in locations using the methodology outlined in [12]. For activity types, MetricWire outputted whether a participant was stationary, walking, cycling, running, or driving at a particular time instant. Using this information, we computed the proportion of day time a participant was doing either of these mentioned activities. All computed features are of continuous type.

Given the breadth of our feature space, multicollinearity posed a significant problem. Thus, we removed highly correlated features within each data stream a priori, using Pearson Correlation coefficients to identify highly correlated features. For modeling, we selected only those features that had a correlation coefficient in the range of -0.75 to 0.75. While existing works have supported our choice of correlation threshold [45, 53], we ultimately set the threshold of ± 0.75 empirically in order to balance the retention of redundant features with our goal of condensing the final feature space used in the predictive modeling step. Within each data stream, collinear features which were dropped during the modeling are shown in blue color in Table 2.

4.2 Subjective Measures

Social anxiety disorder is characterized by increased state social anxiety in daily life, but state social anxiety is not necessarily uniformly elevated throughout the day – it is particularly heightened in certain situations more than others (i.e., socially evaluative situations, such as speaking up in a group or asking a person on a date; [5]). Times of elevated state social anxiety are likely good times for intervention. Thus, we aimed to identify times of

Table 3. Daily subjective measures and corresponding questions collected through EMAs at 10 PM everyday.

Subjective Measure	Question	Response (on a scale of 0 to 10)
Social Effectiveness	I felt that I interacted ___ with others today	Not effectively (0) ... Very effectively (10)
Social Enjoyment	I ___ my social interactions today	Really disliked (0) ... Really enjoyed (10)
Social Acceptance	During my social interactions today, I felt ___	Very rejected (0) ... Very accepted (10)
Social Avoidance	I avoided my social interactions today ___	Not at all (0) ... Very much (10)
Evaluation Concern	I was ___ with what people might think of me.	Very worried (0) ... Very comfortable (10)
Negative Affect	Throughout my whole day, I felt ___	Very negative (0) ... Very positive (10)
Anxiety	Throughout my whole day, I felt ___	Very calm (0) ... Very anxious (10)

elevated state, rather than trait social anxiety towards capturing a better temporal resolution for understanding social anxiety, and holds promise as a step towards just-in-time adaptive interventions (JITAIs).

The Metricwire app was programmed to deliver EMAs daily and collect subjective measures of participants' experiences. Subjective measures were collected using three-minute surveys, which were delivered at the end of every day at 10 PM. These surveys remained active for two hours and then closed automatically at midnight if left unanswered. We gathered seven subjective measures related to perceived social effectiveness, social enjoyment, perceived social acceptance, social avoidance, concern about evaluation, negative affect, and anxiety. The question phrasing and response scales corresponding to each of these subjective measures are shown in Table 3. We will refer to the subjective measures by the identifiers listed in the 'Subjective Measure' column of Table 3. The response for each question was presented on a sliding scale of 0 to 10.

While the current focus on end of day reports (vs. reports throughout the day) means we do not have fine-grained resolution for predicting the right time for an intervention in the specific moment of distress, this evaluation shows the feasibility of the approach and would allow for prediction at the daily level. Notably, some interventions might occur at a daily level, such as planned completion of thought records to reappraise thoughts about fears of negative evaluation.

The Pearson correlation plot for all subjective measures is shown in Figure 3. The plot shows that most of the subjective measures are not significantly correlated with each other. This implies that it is important to analyze these subjective measures separately.

We also aimed to understand the variance of responses to each of the subjective measures across different study days and across the participants. Figure 4 shows the

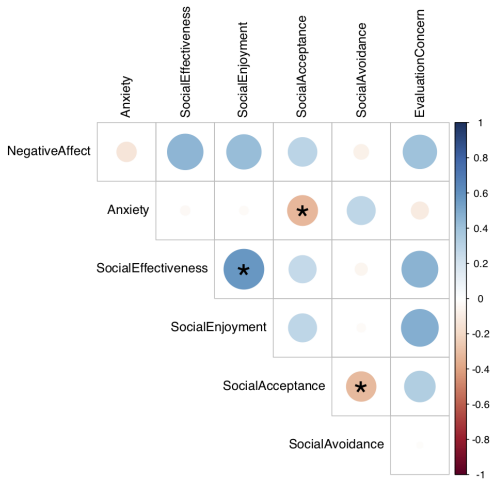


Fig. 3. Correlation plot of subjective measures. The color and size of the circles reflect the magnitude of the correlation coefficient and starred cells denote statistically significant values (i.e. p-value < 0.05).

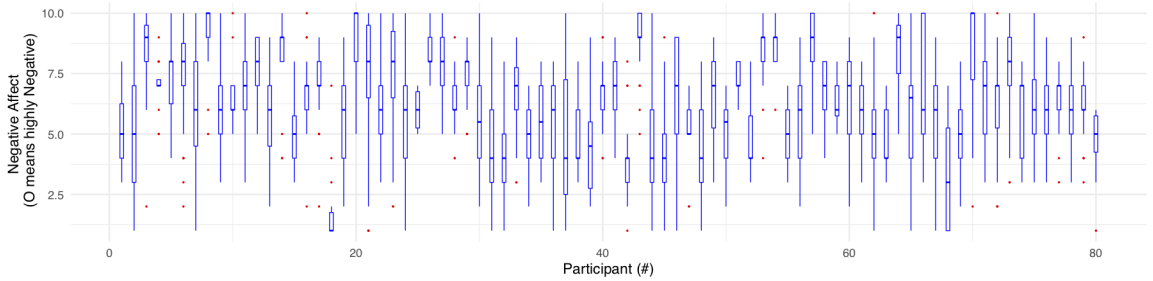


Fig. 4. Boxplots showing the variation in responses to Negative affect subjective measure over the course of the study for different participants.

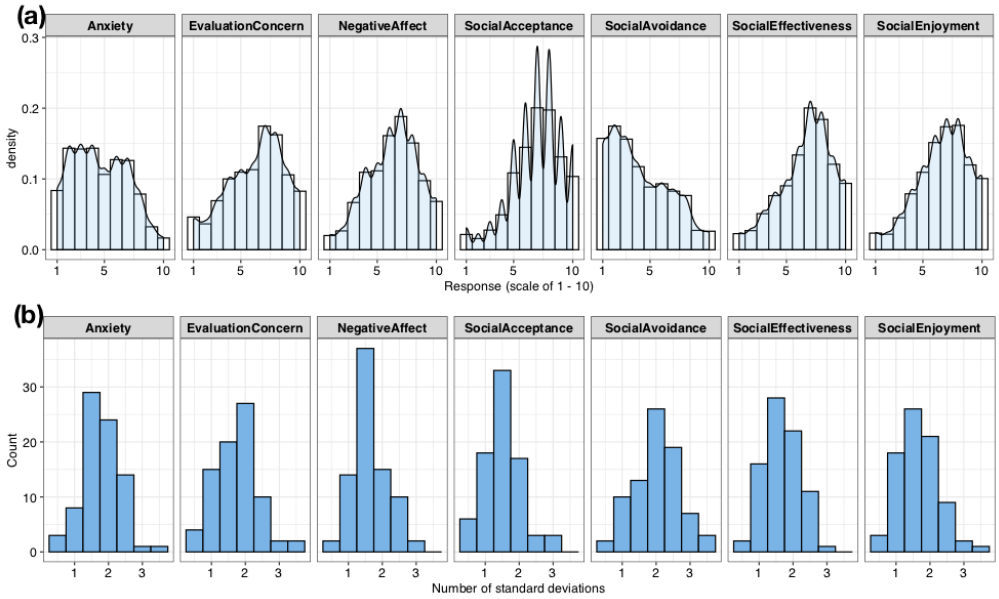


Fig. 5. **(a)** Distribution of participants' responses to different subjective measures. Anxiety (M = 4.69, SD = 2.35), Evaluation-Concern (M = 6.19, SD = 2.45), NegativeAffect (M = 6.25, SD = 2.20), SocialAcceptance (M = 6.85, SD = 2.10), SocialAvoidance (M = 4.18, SD = 2.52), SocialEffectiveness (M = 6.60, SD = 2.25), SocialEnjoyment (M = 6.60, SD = 2.24). **(b)** Histogram showing the frequency of standard deviation in daily subjective responses for all 80 participants. Bin width is set to 0.5.

variability in daily responses to the "Negative Affect" measure for each of the 80 participants. Notably, there is considerable deviation in responses within most participants, i.e., most participants respond to the same subjective measure differently on different days, pointing to the measure's sensitivity to dynamic changes. Furthermore, the boxplots vary in their median values, signifying that participants responded differently to the same subjective measure. Participants' responses to the remaining six subjective measures exhibited similar variability. The demonstrated variability in responses to subjective measures makes the task of predicting these measures using passive data particularly interesting and complex.

Figure 5(a) shows the distribution of participants' responses to different subjective measures. And Figure 5(b) shows the distribution of standard deviations, computed on each of the seven measures for each participant separately. These histograms show that responses provided to each of the subjective measures differ in terms of standard deviations. For example, the histogram for the "Anxiety" measure shows that 29 participants had a standard deviation of 1.5 in their daily responses, meaning they differed by 1.5 from their daily mean response.

4.3 Data Missingness

Our dataset contains data from several mobile sensor data streams (see Table 2 for a complete list). Well-known prediction based machine learning algorithms (e.g., Random Forest, AdaBoost, Decision Trees, kNN) drop rows of data that have missing values in one or more of the features, thus reducing the size of the usable dataset. This potentially drops useful information and reduces the usefulness of the collected datasets to the research community.

Due to the variation in the amount of missing data across the different passive streams and participants, we calculated the percentage of missing data for each data stream on the participant level as follows:

$$\text{Missing data (\%)} = (1 - \frac{d_i}{N}) * 100$$

where $i \in$ {Accelerometer, GPS, Pedometer, Text, Call, Activity}, d_i represents the number of days a participant has submitted the data for the i^{th} data stream, and N represents the total number of days a participant has submitted daily subjective measures via EMAs. We have chosen to represent data missingness on a daily level due to the significant variations in sensing frequency (i.e., continuous streams like Accelerometer vs. polling streams like GPS) as well as inconsistencies in when sensors were enabled or disabled across participants' devices.

Figure 6 shows the percentage of missing data for all participants in the form of a heat map. Each x-axis tick mark in the heat map represents a different participant and y-axis tick-mark labels represent a different data stream. The legend at the top of the figure shows the mapping from percentage of missing data to the chosen color scale. One approach to improve the usefulness of the dataset is to impute the missing values based on the observed values in other data streams. In the following section, we describe several imputation methods and our adaptation of these methods to our mobile sensor data.

5 METHODS

In the previous sections, we have presented our target dataset and rationale for examining various data imputation methods in this problem space. In this section, we describe different data imputation methods and the predictive

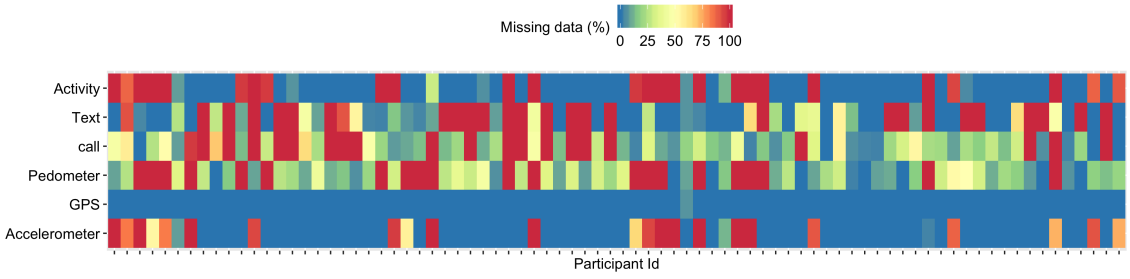


Fig. 6. Heat map showing missing data percentage in each of the six data streams for all 80 participants. Each dot in the x axis represents a participant. [Best viewed in colour.]

modeling approaches we will use to predict the subjective measures. Further, we describe our process for evaluating the performance of these models.

5.1 Data Imputation Algorithms

An important consideration in applying imputation methods to our dataset is to check the assumption that the missing data is missing completely at random (MCAR). To confirm that our dataset satisfied this assumption, we ran Little's MCAR test [8, 41]. Because we obtained a p-value of 0.44, we concluded that the missing data is indeed MCAR and that the imputation methods are viable for our dataset. We used following four imputation methods, each of these is from a different category of imputation methods as discussed in the Related Work section.

Multiple Imputations by Chained Equations (MICE) [6]: This method imputes the missing values multiple times instead of using a single value such as mean, mode, or last observation carried forward. This results in less biased missing values. MICE consists of the following four steps:

(i) *Input data*: Take the dataset with missing values as input.
(ii) *Impute data*: Create several copies of the input dataset and then replace missing values in each copy with regression. Running regression on each copy of the dataset with different observations results in different regression coefficients and hence results in different imputed values [6].
(iii) *Analyze results*: Run a single statistical analysis method on all the copies of imputed datasets and assess the results using several statistical measures (e.g., standard deviation). This step determines which of the imputed copies will be used in the next step.
(iv) *Pool results*: Aggregate selected copies of imputed datasets by computing the mean, and then output a final imputed copy of the dataset.

Matrix Completion: Suppose M represents a matrix of all features with some values missing. The Matrix Completion approach finds a solution matrix, X, by solving the following equation

$$minimize\ ||M - X||_F^2 + \lambda||X||_*$$

where $||.||_F$ is the Frobenius norm, calculated on the non-missing entries of M, and $||X||_*$ is the nuclear norm (sum of singular values) of X. Please refer to [44] for the details of the method.

K-nearest Neighbor (KNN): In this method, the missing values are imputed by taking the weighted average of k-nearest neighbors. The nearest neighbors are chosen based on some distance measure. So, this method requires the selection of a distance metric and value of k.

Last Observation Carried Forward (LOCF): This method is commonly used for imputing longitudinal repeated measures data. It replaces missing values by the recent last known values.

5.2 Predictive Modeling

Existing approaches to predicting subjective responses from passive data features have framed the task as either a classification or regression problem [11, 34, 63]. While classification based approaches result in coarse-grained output values and do not show the variability at a detailed level, regression based approaches result in more fine-grained output values, along the same scale as collected subjective measures. We evaluated seven (linear regression, DecisionTree, XGBoost, LightGBM, Random Forest, MERF, and CatBoost) different prediction methods.

But, in this paper we present only those methods (Random Forest, MERF, and CatBoost) which resulted in lower prediction error. These methods were able to capture the non-linearity in the dataset.

Random Forest (RF): RF is a "bagging" ensemble based machine learning method. It creates various subsets of the input data chosen randomly with replacement. On each subset, decision trees are trained separately first and then the final prediction value is taken as an average of the outputs of all the decision trees.

CatBoost: Catboost [56] is a "boosting" ensemble-based machine learning method. It also uses decision trees as in the case of Random Forests, but the decision trees are arranged sequentially; i.e., results of previous decision tree are fed to the next one so as to produce the final predicted value. It is specially used for the categorical data.

Mixed Effects Random Forests (MERF): MERF [28] falls in the category of linear mixed effect models, i.e., models having two components, a fixed or population-averaged and a random or cluster specific component. Such models are suitable for modeling data that are divided into various clusters. For example, our dataset contains data from 80 participants, which represents 80 clusters. The underlying assumption in these models is that observations belonging to the same cluster are more similar to each other than the observations found in other clusters. So, the features which differentiate clusters should be modeled separately from the rest of the features. The former ones are called random effects and the latter ones are called fixed effect features. MERF extends the RF to operate on clustered data. It is defined as follows:

$$
\begin{aligned}
&y_i = f(X_i) + Z_i b_i + \epsilon_i, \\
&b_i \sim N(0, P),\ \epsilon_i \sim N(O, Q_i),\ i = 1, \ldots n
\end{aligned}
$$

where y_i represents a vector of responses for n_i observations in cluster i, X_i and Z_i represent matrices of fixed-effects and random-effects covariates, b_i represents unknown vector of random effects for cluster i, ϵ_i is the vector of errors, $f(X_i)$ represents an unknown non-linear function and is estimated using Random Forests. P and Q represent covariance matrices of b_i and Q_i respectively.

Baseline: We used a baseline mean method as a benchmark for comparing the performance of the above three methods. In this method, the output is predicted as the mean of previous observed response values. For example, if the training data has subjective measures values for any response as 5, 6, 5, 5, 7, then during testing the predicted value will be 5.6 ($= \frac{5+6+5+5+7}{5}$)

5.3 Evaluation

To evaluate the performance of our chosen predictive models, we used 5-fold cross-validation to ensure the consistency and robustness of our results. We used 80% of data for training and the remaining 20% for testing in each fold while ensuring that 80% of each of the participant's observations went into training and the remaining 20% of their data was alloted for testing. Parameters shown in Table 4 for each of the prediction methods (except Baseline) were set using the GridSearchCV method from the scikit-learn Python library [51]. All data imputation methods were run with default parameter settings, except KNN, in which k was set to 3 with the

Table 4. Parameters set in different prediction methods after using Grid search technique.

Method	Parameters
CatBoost	depth = 7, iterations = 400, l2_leaf_reg = 4, learning_rate = 0.1
MERF	n_estimators = 200, max_iterations = 50
RF	n_estimators = 10, random_state = 0, min_samples_split = 2

ALGORITHM 1: Steps in imputing missing data

Input: A dataset $D[P, F]$ with missing values as shown in Figure 6, where $P \in \{P_1, \ldots, P_{80}\}$ represent participants ids, $F \in \{F_1, \ldots, F_N\}$ represents different passive features

Output: Imputed dataset, $\widehat{D}[P, F]$ with no missing values

$\widehat{D}[P, F] \leftarrow [\,]$ /* Empty matrix for storing imputed data */
for $i \leftarrow 1$ **to** *sizeof(P)* **do** /* for each participant */
  $S_i \leftarrow$ subset D with the data of P_i
  $N \leftarrow [nrows(S_i), 1]$ /* Empty matrix for storing Null features of S_i */
  for $j \leftarrow 1$ **to** *sizeof(F)* **do** /* for each feature */
    if *($S_i[j] == Null$)* **then** /* all values of the feature are NULL */
      Remove $S_i[j]$ from S_i and add to N
  $\widehat{S_i} \leftarrow$ Apply imputation method to S_i
  $\widehat{S_i} \leftarrow \widehat{S_i} \cup N$ /* concatenate features column wise */
  $\widehat{D} \leftarrow \widehat{D} \cup \widehat{S_i}$ /* concatenate row wise */
if *($\widehat{D}$ has any Null value)* **then**
  $\widehat{D} \leftarrow$ Apply imputation method to $\widehat{D}$
return $\widehat{D}$

GridSearch technique. We also used the scikit-learn implementation for Random Forest and publicly available Python implementations for CatBoost[3], MERF[4], LOCF[5], Matrix Completion[6], KNN[7], and MICE[7].

MICE and KNN were run in two steps: (i) The imputation method was applied on each participants' data separately. Only those data streams were imputed that had some of the values missing. Data streams with all values missing were left out; (ii) The imputation method was applied on the whole dataset obtained from the previous step. This step imputes only left out missing values using the data from all participants. The rationale behind applying these techniques in two steps was first to leverage the local patterns found within each participant's data and, then leverage the global patterns found in the entire dataset. Note that the second step did not impute the missing values which were imputed in the first step. Algorithm 1 shows the steps followed in implementing MICE and KNN in detail. The two step methodology was not appropriate for LOCF and Matrix Completion. So, for each of these methods, each participants' data was inputted one at a time and later all participants' imputed data

[3] https://catboost.ai/
[4] https://github.com/manifoldai/merf
[5] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.fillna.html
[6] https://web.stanford.edu/~hastie/swData/softImpute/vignette.html
[7] https://github.com/eltonlaw/impyute

were collated to create a final imputed dataset. Only Steps 1 - 10 of Algorithm 1 were used in the implementation of LOCF.

5.3.1 Evaluation metric. We quantified the prediction errors of the predictive models using Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{N}}$$

where N represents number of days, y_n represents actual value of a subjective measure on day n, and $\hat{y}_n$ represents the predicted value of the subjective measure. The lower the value of RMSE, the better is the prediction accuracy.

6 RESULTS

In this section, we first address **RQ1** regarding the task of predicting subjective measures of SAD using passive data collected from smartphones. Specifically, we present the prediction results obtained from the non-imputed dataset and compare the performance across the different regression models. We then move to **RQ2** and present the corresponding results derived from the imputed dataset and compare both the between-model and overall performance to our non-imputed results. Finally, we explain the variation of prediction performance of each of the seven subjective measures in relation to both fluctuations in participants' reports and SIAS score (i.e., trait social anxiety).

6.1 Predicting Daily Subjective Measures from Passive Data Features

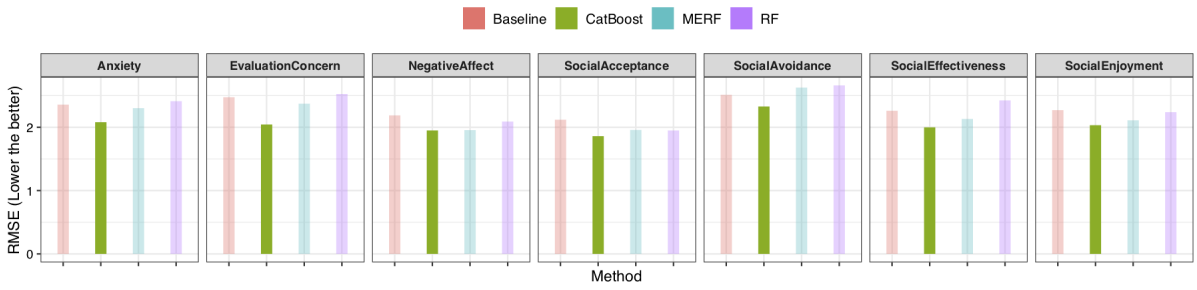


Fig. 7. RMSE for different prediction methods using non-imputed dataset (i.e., none of the missing values were imputed.) [Best viewed in color.]

Figure 7 shows the RMSE values for RF, MERF, CatBoost, and a baseline prediction method using the non-imputed dataset. The lower the RMSE value, the better the prediction accuracy. Each subplot of the figure corresponds to the prediction results of a different subjective measure. Among the three prediction methods used, CatBoost performed consistently better than RF, MERF, and Baseline for all seven subjective measures. We hypothesize that the reasons for this performance improvement are two-fold: i) CatBoost uses all the observations in the dataset, where as RF and MERF drops observations that have one or more missing values; ii) CatBoost handles missing values internally; thus, the performance improvement may result from the additional affordances of this built-in handling approach.

Among the seven subjective measures, "Social Acceptance" and "Social Avoidance" had the lowest and highest overall RMSE values, respectively. The extreme deviation in RMSE values likely reflects the underlying deviation in the subjective measure responses as shown in Figure 5(b). Histograms in the figure show that "Social Acceptance"

had the least number (~ 16) of participants having ≥ two standard deviations whereas "Social Avoidance" had the highest number (~ 45) of participants having ≥ two standard deviations.

6.2 Effect of Imputation on Predictive Model Performance

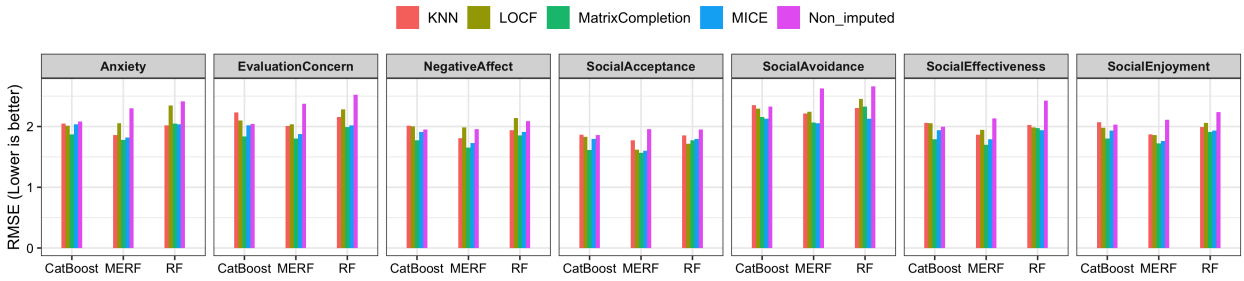


Fig. 8. RMSE for different prediction methods (CatBoost, MERF, and RF) with Non-imputed and imputed data (obtained with KNN, LOCF, Matrix Completion, and MICE). [Best viewed in color]

6.2.1 Model Performance on Imputed Datasets. Figure 8 shows the RMSE values for all model predictions made using four imputed datasets obtained with KNN, LOCF, Matrix Completion, and MICE imputation methods. The Figure shows: (i) Overall, all three prediction methods (CatBoost, MERF, and RF) for all seven subjective measures resulted in lower RMSE when using the imputed dataset obtained with the Matrix Completion imputation method. (ii) For all subjective measures, MERF resulted in lower RMSE compared to RF and CatBoost. This performance improvement in MERF makes sense in this application due to the completeness of the imputed dataset and MERF's suitability for clustering data (given each participant's data is considered as a separate cluster). By using all the observations, this approach can detect the clustering nature of the data. Predicting subjective measures while accounting for the fact that each participant is different resulted in lower prediction error. The performance of CatBoost improved marginally as compared to its performance on the Non-imputed dataset. Although CatBoost handled missing values internally in the non-imputed dataset, the performance was still not on par with the performance obtained using the imputed dataset. This indicates the internal missing data handling approach of CatBoost was not as robust as applying the Matrix Completion method.

We also aimed to understand whether the results of different machine learning based prediction methods differ statistically from the baseline method. Thus, we ran Student's t-test between a pairwise set of the baseline method results and the different prediction method results. We ensured the applicability of t-test by checking the normality and variance assumptions by using Shapiro-Wilk[8] and Levene[9] tests respectively. With t-test, almost all pairs of results were found to be significantly different as shown in Table 5; thus, we can conclude that all of the machine learning based prediction methods' results differed significantly from the baseline results.

6.2.2 Performance Improvement. To understand the differences in model performance between the imputed and non-imputed dataset, we computed the percentage decrease in prediction error (RMSE) when using the Matrix Completion imputed dataset for each of the subjective measures. Figure 9 shows that average percentage decrease in the RMSE values of RF, MERF, and CatBoost across all seven subjective measures are 18.16%, 22.25%, and 3.7% respectively. The higher RMSE percentage decrease found in RF and MERF when using the imputed dataset can

[8] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html
[9] https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.levene.html

Table 5. *t*-value (*p*-value) computed between the results of the Baseline method and each of the prediction methods (CatBoost, MERF, and RF) while using imputed datasets produced by KNN, LOCF, Matrix Completion, and MICE. The greater the magnitude of *t*-value, the greater is the evidence against the null-hypothesis (i.e., prediction method and the Baseline method results are same.)

	KNN	**LOCF**	**Matrix Completion**	**MICE**
CatBoost	2.7(0.018)	3.5(0.003*)	5.7(0.000*)	5.0(0.000*)
MERF	5.0(0.000*)	3.8(0.002*)	6.8(0.000*)	6.7(0.000*)
RF	3.4(0.004*)	1.5(0.141)	3.8(0.002*)	5.0(0.000*)

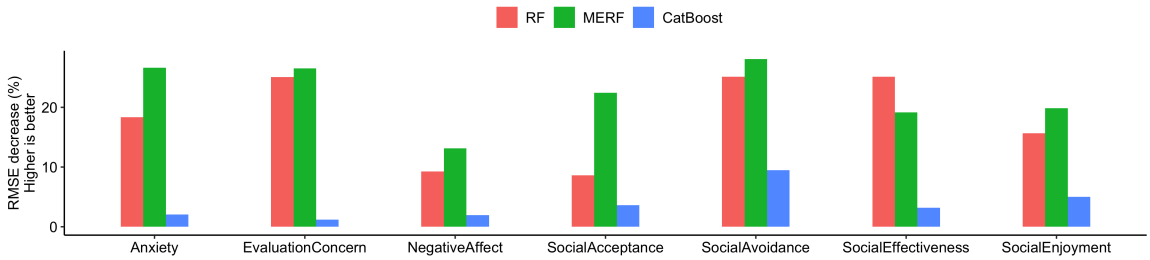


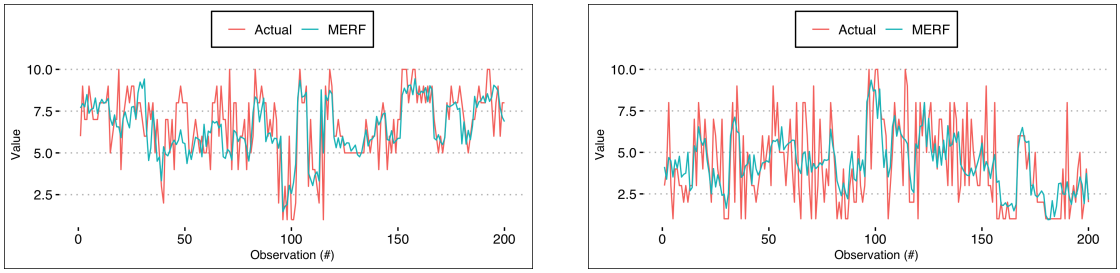
Fig. 9. RMSE percentage decrease while using the imputed data with different methods. [Best viewed in color.]

be explained by the substantial increase in useful observations afforded by the Matrix Completion imputation method. Specifically, the missingness in the non-imputed dataset significantly decreased the amount of usable data (only 13% of the data; i.e., 305 out of 2,268 observations were usable). This confirms our expectation that applying the Matrix Completion imputation algorithm to our dataset (and thus increasing the number of useful observations) enables our regression models to better predict the daily subjective measures from the patterns within the passive data features.

6.3 Subjective Measure Variability & Model Performance

6.3.1 Deviations in Subjective Measures. Thus far, we have presented aggregated metrics of model performance in predicting the seven subjective measures. However, another important consideration is the variation of predicted values in comparison to the actual reported values. To investigate this pattern within our dataset, we plotted MERF predicted and actual values in Figure 10. We selectively plotted results for "Social Acceptance" and "Social Avoidance" measures in order to study the variation across measures for which the model had produced both low and high RMSE values. Our analysis shows that the average deviation of predicted values in "Anxiety", "Evaluation Concern", "Negative Affect", "Social Acceptance", "Social Avoidance", "Social Effectiveness", and "Social Enjoyment" subjective measures from the actual value is 1.4, 1.7, 1.3, 1.1, 1.7, 1.3, 1.2 respectively.

To further understand the correlation between actual and predicted values, we computed Pearson correlation coefficient between the actual and MERF predicted values for all subjective measures separately. The values for "Anxiety", "Evaluation Concern", "Negative Affect", "Social Acceptance", "Social Avoidance", "Social Effectiveness", and "Social Enjoyment" subjective measures were found as 0.63, 0.65, 0.61, 0.65, 0.57, 0.61, and 0.64 respectively. All are significantly correlated with p-value < 0.01.



(a) Social Acceptance

(b) Social Avoidance

Fig. 10. Actual and MERF predicted values of subjective measures: (a) 'Social Acceptance', (b) 'Social Avoidance'.

6.3.2 Individual Differences in Trait Anxiety. To understand the relationship between our model predictions and trait social anxiety, we plotted each participant's SIAS score, standard deviations in subjective measure, and RMSE values (for MERF results) in Figure 11. We arranged participants in decreasing order of their SIAS score (with lower scores indicating relatively less severe social anxiety symptoms) to interpret linear trends with respect to SIAS. Through visualizing the relationship between trait social anxiety and model performance in this format, we identified a number of notable patterns. First, we found that as SIAS decreases, standard deviations decrease. That is, participants who were more socially anxious (i.e., had a higher SIAS score) produced higher standard deviations in their subjective measure reports, whereas the subjective measures of less socially anxious participants exhibited lower standard deviations. This suggests that highly social anxious participants report a greater degree of variability in self-report measures across days as compared to participants with less severe social anxiety symptoms.

Secondly, we found that the RMSE decreases as the SIAS score decreases. This suggests that predicted subjective measures for low socially anxious participants are more accurate and reliable than those for high socially anxious participants. This may be attributable to dynamic changes as captured by the standard deviation of daily responses (see middle panel of the Figure).

Furthermore, Figure 11 shows that the relationship between trait anxiety and reports of social anxiety symptoms varies across different subjective measures. For example, for the same participant, the SD values are different for the "Social Acceptance" vs. the "Social Avoidance" subjective measures.

7 DISCUSSION

7.1 Towards Predicting Subjective Measures from Passively Collected Mobile Sensing Data

Predicting participant self-report surveys from passively collected data streams allows us to understand whether objective physical behaviors that can be passively measured with ubiquitous mobile technology are informative about a person's subjective experience. Prior work has demonstrated the potential of utilizing passive data to predict subjective measures [12, 50, 82]. In the current work, we collected seven different measures related to social anxiety and then evaluated various machine learning algorithms for predicting each of these measures separately to provide a more comprehensive evaluation. Our results demonstrate that, on average, the predicted subjective reports varied by 1.3 units (on a scale from 0 to 10) from participants' actual subjective reports (see example in Figure 10). Further, we found that our predicted values for each measure were highly correlated with their respective observed values (mean Pearson's r = 6.2, all p-values < 0.01). This indicates that, despite the minor deviations from participants' reported experiences of social anxiety, our predictions align relatively well with participants' lived experience.

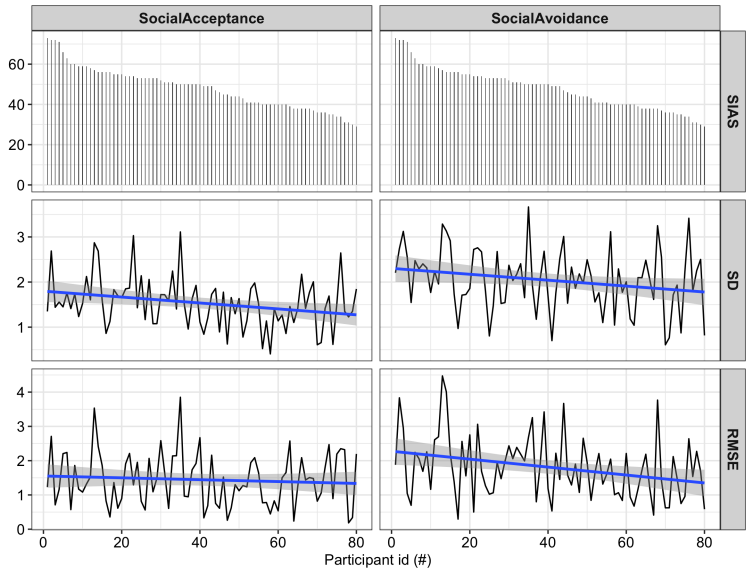


Fig. 11. Grid plot showing SIAS score of participants in top row, standard deviations (SD) of daily responses in middle row, and RMSE with MERF in bottom row. Columns correspond to 'Social Acceptance' and 'Social Avoidance' subjective measures.

We were also interested in comparing more complex, machine learning based predictive models with a baseline averaging approach. As shown in Figure 7, the machine learning based prediction methods (i.e. RF, MERF and CatBoost) performed better than the baseline method in predicting all of the subjective measures, with the exception of "Social Avoidance" and "Social Effectiveness." Since the RMSE values produced by all four predictive modeling approaches were within 0.5 of each other across all subjective measures, it is difficult to draw definitive conclusions from these trends. However, further investigation of the instruments used in this study and the respective self-reporting behavior of participants could yield valuable insights into how they might be combined or predicted in tandem to improve model performance.

In the context of training predictive models on mobile sensor data, the configuration of cross-validation (CV) is an important factor. Prior research has presented differing insights on the affordances of random cross-validation (e.g. 5-fold CV) as compared with leave-one-out cross validation (LOOCV) [29]. In our analysis, we pose the major differences between the two approaches as follows: 5-fold CV ensures that the predictive model is trained on a portion of each participant's data whereas LOOCV renders the model completely blind to participants' observed behavior when making predictions. We thus evaluated our predictive modeling task using both the 5-fold CV and LOOCV approach. We found that the average RMSE produced via the LOOCV approach for all the subjective measures was 2.5; on the other hand, the average RMSE produced with the 5-fold CV approach was around 2 (as shown in Figure 8). The decreased prediction error with 5-fold cross-validation approach suggests our prediction models need to be trained with each participant's data explicitly. In the context of our dataset, this suggests that the relations between passive and self-report data vary from participant to participant.

7.2 Missingness in Mobile Sensing Data

Real-world mobile sensing studies will inevitably result in missing data. This unavoidable missingness should not always limit the usability of a dataset and the ability to answer research questions posed by the study. Missing

values are generally imputed using a statistical approach or by relying on domain knowledge. In our dataset, domain knowledge cannot be used to impute the missing values as it is difficult to predict the participant's behaviors a priori. However, several data streams are collected simultaneously from the smartphone. Thus, there is a possibility of inferring the missing values in one data stream from the remaining streams [62]. For example, accelerometer data varies by activity type (walking, running), implying that accelerometer data could be used to infer values in the Activity type data stream. Exploring this idea further, we investigated the efficacy of four imputation algorithms towards alleviating the issue of missing data to predict subjective measures from passive data streams. Our results reveal a number of important insights on the potential for data imputation to overcome existing limitations in mobile sensing work, particularly for predicting self-reported outcomes.

First, we highlight the impact of imputing missing values in passive mobile sensor data streams on the performance of the applied predictive models. We demonstrate that the imputation methods we evaluated in this analysis resulted in consistent improvements in model performance across seven distinct target measures of social anxiety. As compared to results from the non-imputed dataset, on average (across the prediction methods and the seven subjective measure), LOCF, KNN, MICE, and Matrix Completion resulted in the reduction of RMSE by 11%, 13%, 17%, and 20% respectively. Our results suggest that the Matrix Completion method was able to account for the underlying relations between different features of each of the participants' passive data. However, similarly to the non-imputed predictive modeling results, the relatively minute differences in RMSE across the various imputation methods indicates that further investigation is required to determine the superiority of any given method for similar applications.

We additionally compared the performance of single-iteration imputation methods (i.e. LOCF, KNN) with multiple-iteration methods (i.e. MICE, Matrix Completion). Our results showed that single-iteration methods consistently underperformed multiple-iterations methods (average RMSE of 2.0 vs. 1.8). Furthermore, we found that predictive models trained on the dataset imputed using the LOCF method consistently underperformed models that leveraged the other imputation methods we evaluated. This is a particularly important finding given the common use of LOCF in mobile sensing literature, particularly in the context of mental health prediction tasks [20, 31]. For a sequence of missing values, LOCF imputes the entire sequence using the same value. Further, it does not leverage other features within the dataset to impute the values of a feature with missing values.

Alternatively to LOCF, mobile sensing researchers have often leveraged machine learning approaches that internally account for missing data samples in order to mitigate negative effects on predictive performance. In our analysis, we aimed to formally compare these internal imputation methods to independent imputation methods used prior to the prediction step. We found that although the CatBoost model internally handled of missing data, the resulting increase in model performance was less than that achieved by other predictive modeling approaches coupled with the Matrix Completion imputation method.

Our findings suggest that simple averaging or carrying forward of previous values is not an effective approach in imputing values in datasets with multiple features. The application of a broader range of imputation algorithms to future mobile sensing applications may allow us to more effectively make sense of the underlying behavioral patterns of mental health populations in spite of the data sparsity. In particular, shifting from naive single-iteration methods, such as LOCF, towards more robust and complexity-preserving techniques, including multiple-iteration methods like MICE and Matrix Completion, is critical to advancing mobile sensing applications.

7.3 Implications for Understanding Social Anxiety

SAD is characterized by elevated feelings of anxiety and negativity throughout daily life, especially in the context of social situations. For some people, the intensity of their anxious and negative feelings might be so intolerable that they avoid social situations altogether. For others, they may enter into social situations but find those social experiences to be unenjoyable or to lead them to worry that those they talk to are judging them negatively. For

others still, their anxiety might be driven by thinking that other people tend to reject them socially. Or, perhaps they feel a part of their social group, but instead have a tendency to think back on their social interactions and criticize themselves for things they said that they think came out incorrectly (i.e., they believe themselves to be socially ineffective). These daily aspects of social anxiety are expected to interrelate differently for each individual, given differences in symptom presentation, making it important to sample these features separately in order to characterize each individual's experience of SAD. Our results show that not all of the subjective measures can be predicted with a similar accuracy (see Figures 7 and 8). This suggests other passive features may need to be collected to better explain the variability in the subjective measures. Such features may include sleep quality and duration, social interaction features (pitch, energy, etc).

By leveraging EMA to collect information on multiple aspects of a person's subjective experience of anxiety, we investigated seven subjective measures related to perceived social effectiveness, social enjoyment, perceived social acceptance, social avoidance, concern about evaluation, negative affect, and anxiety. Our analysis shows that these measures of social anxiety exhibit different levels of variability, across both measures and participants (as shown in Figure 5(b)). While these variations are in part expected due to the diversity of peoples' lived experiences, the question remains as to whether this variability translates to varying degrees of predictability. In working towards clinical applications, it will be important to assess which subjective measures are most relevant to any given person's SAD symptom profile to target interventions appropriately. Further, some of the subjective measures used in this analysis were found to be reliably correlated with one another (see Figure 3). By investigating the relative predictability of these interrelated instrument variables, we hope to identify redundancies in our EMA survey design and improve future iterations of this study deployment, thus reducing user burden.

For example, in the current data, social effectiveness and social enjoyment were highly positively correlated with each other. This strong, positive correlation may indicate that, in a sample of individuals all high in trait social anxiety severity, these survey items do not capture unique information. We included both items in the current study because previous research has shown that for people with more (vs. less) severe social anxiety symptoms, high negative affect in social interactions relates differently to social effectiveness and social enjoyment [25]. However, unlike the undergraduate sample studied by Geyer and colleagues, the current study only included participants who were above a certain threshold of social anxiety severity. The high correlation between social effectiveness and social enjoyment in the current study may suggest that researchers studying features of social anxiety in daily life within an exclusively anxious sample may opt to use just one of these survey items. That these constructs are so highly correlated in a socially anxious sample is not surprising, as an anxious person who believes that they did not perform well in their social interactions that day is likely to have also not enjoyed those interactions. Our results show that, for the correlated subjective measures (i.e. 'Social Effectiveness' and 'Social Enjoyment'), the prediction results by different prediction methods have almost the same prediction error (RMSE). This means that we do not need to collect/predict each of them separately given each of them can be inferred from the other.

We additionally aimed to account for the impact of participants' varying levels of baseline social anxiety, as measured by a clinically-validated scale. Although we recruited only participants whose trait social anxiety score was elevated (here, at least 29 out of 80 on the SIAS) , there was still considerable variability in trait social anxiety across eligible participants. A participant scoring a 29 would, on average, endorse slight to moderate experiences of the 21 symptoms measured by the SIAS, whereas a person scoring 80 would endorse universally extreme symptoms. While both people are indeed socially anxious, the person with the higher score would likely be much more functionally impaired than the less anxious person: the person who scored 80 might rigidly avoid nearly all social interactions, view themselves as unwaveringly ineffective in the few social situations they do enter, and never enjoy interacting with others. The person endorsing greater levels of trait social anxiety might experience more consistently intense daily social anxiety, whereas the less anxious person's experience of various aspects of

anxiety might fluctuate more from day to day. Figure 11 shows that the prediction error for high socially-anxious participants is more as compared to low socially-anxious participants.

7.4 Limitations and Future Work

Notably, the current study does not evaluate the relative importance of the seven different subjective measures to individuals diagnosed with SAD. Individuals with SAD vary in their symptom presentations; it is likely that feature importance will vary between individuals. For example, one person with SAD might experience difficulty related to avoiding situations where they need to present at work, and another person with SAD might participate in lots of social situations for fear of missing out on something, but experience a lot of fear that others are judging them and not enjoy social situations. Future work should address the relative importance of each of these measures to the experience of social anxiety in daily life in order to inform future study design and clinical application in the personalized treatment of SAD.

Researchers often limit the number of survey items per EMA as much as possible to reduce participant burden, given that participants are asked to answer surveys multiple times throughout the duration of the study. In the present study, participants answered the nightly survey 35 times. As a result, we opted to keep the nightly survey brief to reduce participant burden. Further, we elected to refrain from predicting a composite score, in favor of predicting each single-item construct separately, because we were interested in identifying unique aspects of the socially anxious experience rather than some global composite. We expect this level of specificity to be more useful to clinicians who are interested in better understanding the unique ways in which the socially anxious experience does or does not manifest itself within daily life. For instance, if a clinician sees elevated behavioral avoidance of social situations, then the intervention is likely to focus on planned social threat exposures, while if the elevated item centers on fears of negative evaluation, then cognitive restructuring may be more appropriate. Thus, the items reflect different meaningful intervention targets, so while a global composite could in some ways be useful, it might obscure meaningful within-construct variance over time.

Further, with respect to our handling of missing data, our omission of participants due to lack of compliance with the subjective measures introduces a number of notable limitations in our results. Specifically, our results can only be extended to moderately sparse datasets (i.e. $> 50\%$ of data available). Future work should investigate approaches that are more suitable to extremely sparse datasets without introducing bias and extrapolating limited observations of behavioral patterns over a longitudinal period of data collection. Furthermore, researchers should also continue efforts towards improving engagement and compliance with EMA-based data collection.

8 CONCLUSION

The proliferation of mobile technologies have allowed researchers to collect passive data continuously in an unobtrusive manner towards characterizing individuals' lived experiences. However, due to the noisy nature of real-world data collection systems, missing values often limit the completeness and thus predictive usability of mobile sensing datasets. dataset collected to monitor persons high in social anxiety symptoms in their daily lives. In this work, we aimed to build upon existing methodologies for predicting the subjective measures of social anxiety from passive data streams. So, to answer the research question and to handle the data missingness. Specifically, we evaluated the efficacy if four well-known publicly available data imputation approaches towards mitigating the gaps in our dataset and thus improving the performance of our predictive models. After comparing both the imputed and non-imputed versions of the same dataset for predicting seven different subjective measures of social anxiety, our results show that leveraging sophisticated imputation methods, such as Matrix Completion, improved our model(MERF) performance (measured by RMSE) by 22%. This suggests that data imputation is a promising direction for dealing with missing values found in passively collected smartphone data.

ACKNOWLEDGMENTS

Research reported in this publication was supported by a University of Virginia Hobby Postdoctoral and Predoctoral Fellowship Grant, a University of Virginia 3 Cavaliers Grant, and the National Institute of Mental Health under award numbers R01MH113752.

REFERENCES

[1] ADAA. [n.d.]. Social Anxiety Disorder. https://adaa.org/understanding-anxiety/social-anxiety-disorder
[2] Ashley Archiopoli, Tamar Ginossar, Bryan Wilcox, Magdalena Avila, Ricky Hill, and John Oetzel. 2016. Factors of interpersonal communication and behavioral health on medication self-efficacy and medication adherence. *AIDS care* 28, 12 (2016), 1607–1614.
[3] Nanna Yr Arnardottir, Annemarie Koster, Dane R Van Domelen, Robert J Brychta, Paolo Caserotti, Gudny Eiriksdottir, Johanna Eyrun Sverrisdottir, Lenore J Launer, Vilmundur Gudnason, Erlingur Johannsson, and T.B. Harris. 2012. Objective measurements of daily physical activity patterns and sedentary behaviour in older adults: Age, Gene/Environment Susceptibility-Reykjavik Study. *Age and ageing* 42, 2 (2012), 222–229.
[4] Joost Asselbergs, Jeroen Ruwaard, Michal Ejdys, Niels Schrader, Marit Sijbrandij, and Heleen Riper. 2016. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. *Journal of medical Internet research* 18, 3 (2016), e72.
[5] American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.
[6] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. 2011. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research* 20, 1 (2011), 40–49.
[7] Jakob E Bardram, Mads Frost, Károly Szántó, Maria Faurholt-Jepsen, Maj Vinberg, and Lars Vedel Kessing. 2013. Designing mobile health technology for bipolar disorder: a field trial of the monarca system. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2627–2636.
[8] AA Beaujean. [n.d.]. R Package for Baylor University educational psychology quantitative courses [Internet]. CRAN; 2012./citado 10 feb 2015.
[9] Dror Ben-Zeev, Christopher J Brenner, Mark Begale, Jennifer Duffecy, David C Mohr, and Kim T Mueser. 2014. Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. *Schizophrenia bulletin* 40, 6 (2014), 1244–1253.
[10] Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal* 38, 3 (2015), 218.
[11] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Pentland. 2014. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*. 477–486.
[12] Mehdi Boukhechba, Philip Chow, Karl Fua, Bethany A Teachman, and Laura E Barnes. 2018. Predicting Social Anxiety From Global Positioning System Traces of College Students: Feasibility Study. *JMIR mental health* 5, 3 (2018).
[13] Mehdi Boukhechba, Alexander R Daros, Karl Fua, Philip I Chow, Bethany A Teachman, and Laura E Barnes. 2018. DemonicSalmon: Monitoring mental health and social interactions of college students using smartphones. *Smart Health* (2018).
[14] Mehdi Boukhechba, Yu Huang, Philip Chow, Karl Fua, Bethany A. Teachman, and Laura E. Barnes. 2017. Monitoring Social Anxiety from Mobility and Communication Patterns. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp '17)*. ACM, New York, NY, USA, 749–753.
[15] Nicole Michelle Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, J. Chris Karr, Emily Giangrande, and C. David Mohr. 2011. Harnessing Context Sensing to Develop a Mobile Intervention for Depression. *J Med Internet Res* 13, 3 (12 Aug 2011), e55.
[16] Emmanuel J Candès and Benjamin Recht. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9, 6 (2009), 717.
[17] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 1293–1304.
[18] Larry Chan, Vedant Das Swain, Christina Kelley, Kaya de Barbaro, Gregory D Abowd, and Lauren Wilcox. 2018. Students' Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–20.
[19] Philip Chow, Haoyi Xiong, Karl Fua, Wes Bonelli, Bethany A Teachman, and Laura E Barnes. 2016. SAD: Social anxiety and depression monitoring system for college students. (2016).
[20] Philip I Chow, Karl Fua, Yu Huang, Wesley Bonelli, Haoyi Xiong, Laura E Barnes, and Bethany A Teachman. 2017. Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. *Journal of medical Internet research* 19, 3 (2017), e62.

[21] M. Faurholt-Jepsen, M. Frost, C. Ritz, E. M. Christensen, A. S. Jacoby, R. L. Mikkelsen, U. Knorr, J. E. Bardram, M. Vinberg, and L. V. Kessing. 2015. Daily electronic self-monitoring in bipolar disorder using smartphones – the MONARCA I trial: a randomized, placebo-controlled, single-blind, parallel group trial. *Psychological Medicine* 45, 13 (2015), 2691–2704.
[22] Aaron J Fisher, Jonathan W Reeves, Glenn Lawyer, John D Medaglia, and Julian A Rubel. 2017. Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of abnormal psychology* 126, 8 (2017), 1044.
[23] Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personality and individual differences* 7, 3 (1986), 385–400.
[24] Yusong Gao, Ang Li, Tingshao Zhu, Xiaoqian Liu, and Xingyun Liu. 2016. How smartphone usage correlates with social anxiety and loneliness. *PeerJ* 4 (2016), e2197.
[25] Emily C Geyer, Karl C Fua, Katharine E Daniel, Philip I Chow, Wes Bonelli, Yu Huang, Laura E Barnes, and Bethany A Teachman. 2018. I did OK, but did I like it? Using ecological momentary assessment to examine perceptions of social interactions associated with severity of social anxiety and depression. *Behavior therapy* 49, 6 (2018), 866–880.
[26] Jiaqi Gong, Yu Huang, Philip I Chow, Karl Fua, Matthew S Gerber, Bethany A Teachman, and Laura E Barnes. 2019. Understanding behavioral dynamics of social anxiety among college students through smartphone sensors. *Information Fusion* 49 (2019), 57–68.
[27] John W Graham. 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology* 60 (2009), 549–576.
[28] Ahlem Hajjem, François Bellavance, and Denis Larocque. 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* 84, 6 (2014), 1313–1328.
[29] Nils Y Hammerla and Thomas Plötz. 2015. Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 1041–1051.
[30] Gabriella M Harari, Nicholas D Lane, Rui Wang, Benjamin S Crosier, Andrew T Campbell, and Samuel D Gosling. 2016. Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science* 11, 6 (2016), 838–854.
[31] Virginia Harrison, Judith Proudfoot, Pang Ping Wee, Gordon Parker, Dusan Hadzi Pavlovic, and Vijaya Manicavasagar. 2011. Mobile mental health: review of the emerging field and proof of concept study. *Journal of mental health* 20, 6 (2011), 509–524.
[32] Kristin E Heron, Robin S Everhart, Susan M McHale, and Joshua M Smyth. 2017. Using mobile-technology-based ecological momentary assessment (EMA) methods with youth: A systematic review and recommendations. *Journal of pediatric psychology* 42, 10 (2017), 1087–1107.
[33] Yu Huang, Jiaqi Gong, Mark Rucker, Philip Chow, Karl Fua, Matthew S Gerber, Bethany Teachman, and Laura E Barnes. 2017. Discovery of behavioral markers of social anxiety from smartphone sensor data. In *Proceedings of the 1st Workshop on Digital Biomarkers*. 9–14.
[34] Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. 2015. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 222–228.
[35] Amanda Jensen-Doss, Ashley M Smith, Emily M Becker-Haimes, Vanesa Mora Ringle, Lucia M Walsh, Monica Nanda, Samantha L Walsh, Colleen A Maxwell, and Aaron R Lyon. 2018. Individualized progress measures are more acceptable to clinicians than standardized measures: Results of a national survey. *Administration and Policy in Mental Health and Mental Health Services Research* 45, 3 (2018), 392–403.
[36] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. 2010. Matrix completion from noisy entries. *Journal of Machine Learning Research* 11, Jul (2010), 2057–2078.
[37] Ronald C Kessler, Wai Tat Chiu, Olga Demler, and Ellen E Walters. 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* 62, 6 (2005), 617–627.
[38] Hisashi Kurasawa, Hiroshi Sato, Atsushi Yamamoto, Hitoshi Kawasaki, Motonori Nakamura, Yohei Fujii, and Hajime Matsumura. 2014. Missing sensor value estimation method for participatory sensing environment. In *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 103–111.
[39] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications magazine* 48, 9 (2010), 140–150.
[40] Joshua D Lipsitz and Franklin R Schneier. 2000. Social phobia. *Pharmacoeconomics* 18, 1 (2000), 23–32.
[41] Roderick JA Little. 1988. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association* 83, 404 (1988), 1198–1202.
[42] Zitao Liu, Yan Yan, Jian Yang, and Milos Hauskrecht. 2015. Missing value estimation for hierarchical time series: A study of hierarchical Web traffic. In *2015 IEEE International Conference on Data Mining*. IEEE, 895–900.
[43] Richard P Mattick and J Christopher Clarke. 1998. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour research and therapy* 36, 4 (1998), 455–470.
[44] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research* 11, Aug (2010), 2287–2322.
[45] Kathryn A McGurk, Arianna Dagliati, Davide Chiasserini, Dave Lee, Darren Plant, Ivona Baricevic-Jones, Janet Kelsall, Rachael Eineman, Rachel Reed, Bethany Geary, et al. 2019. The use of missing values in proteomic data-independent acquisition mass spectrometry to

enable disease activity discrimination. *Bioinformatics* (2019).

[46] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 1132–1138.

[47] Darién Miranda, Marco Calderón, and Jesus Favela. 2014. Anxiety detection using wearable monitoring. In *Proceedings of the 5th Mexican Conference on Human-Computer Interaction*. ACM, 34.

[48] David C Mohr, Mi Zhang, and Stephen M Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* 13 (2017), 23–47.

[49] Peter CM Molenaar. 2004. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement* 2, 4 (2004), 201–218.

[50] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K D'Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–21.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[52] Brett W Pelham. 1991. On confidence and consequence: The certainty and importance of self-knowledge. *Journal of personality and social psychology* 60, 4 (1991), 518.

[53] Yasset Perez-Riverol, Max Kuhn, Juan Antonio Vizcaíno, Marc-Phillip Hitz, and Enrique Audain. 2017. Accurate and fast feature selection workflow for high-dimensional omics data. *PloS one* 12, 12 (2017).

[54] Ivan Miguel Pires, Nuno M Garcia, and Francisco Flórez-Revuelta. 2015. Multi-sensor data fusion techniques for the identification of activities of daily living using mobile devices. *Proceedings of the ECMLPKDD* (2015).

[55] Skyler Place, Danielle Blanch-Hartigan, Channah Rubin, Cristina Gorrostieta, Caroline Mead, John Kane, Brian P Marx, Joshua Feast, and Thilo Deckersbach. 2017. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *Journal of medical Internet research* 19, 3 (2017).

[56] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Advances in neural information processing systems*. 6638–6648.

[57] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 385–394.

[58] Reza Rawassizadeh, Hamidreza Keshavarz, and Michael Pazzani. 2019. Ghost Imputation: Accurately Reconstructing Missing Data of the Off Period. *IEEE Transactions on Knowledge and Data Engineering* (2019), 1–1.

[59] Panteha Hayati Rezvan, Katherine J Lee, and Julie A Simpson. 2015. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC medical research methodology* 15, 1 (2015), 30.

[60] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015), e175.

[61] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd, and Munmun De Choudhury. 2017. Inferring mood instability on social media by leveraging ecological momentary assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 95.

[62] Koustuv Saha, Manikanta D Reddy, Vedant das Swain, Julie M Gregg, Ted Grover, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, Shayan Mirjafari, Raghu Mulukutla, et al. 2019. Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 178–184.

[63] Akane Sano, Z Yu Amy, Andrew W McHill, Andrew JK Phillips, Sara Taylor, Natasha Jaques, Elizabeth B Klerman, and Rosalind W Picard. 2015. Prediction of happy-sad mood from daily behaviors and previous sleep history. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 6796–6799.

[64] Akane Sano, Sara Taylor, Andrew W McHill, Andrew JK Phillips, Laura K Barger, Elizabeth Klerman, and Rosalind Picard. 2018. Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study. *Journal of Medical Internet Research* 20, 6 (June 2018), e210.

[65] Hillol Sarker, Matthew Tyburski, Md Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H Epstein, Kenzie L Preston, C Debra Furr-Holden, Adam Milam, Inbal Nahum-Shani, et al. 2016. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 4489–4501.

[66] Joseph L Schafer and John W Graham. 2002. Missing data: our view of the state of the art. *Psychological methods* 7, 2 (2002), 147.

[67] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. 2009. Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being*. Springer, 157–180.

[68] Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4 (2008), 1–32.

[69] Sandip Sinharay, Hal S Stern, and Daniel Russell. 2001. The use of multiple imputation for the analysis of missing data. *Psychological methods* 6, 4 (2001), 317.

[70] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj* 338 (2009), b2393.

[71] Nina Stuhldreher, Eric Leibing, Falk Leichsenring, Manfred E Beutel, Stephan Herpertz, Juergen Hoyer, Alexander Konnopka, Simone Salzer, Bernhard Strauss, Joerg Wiltink, and H.H. König. 2014. The costs of social anxiety disorder: the role of symptom severity and comorbidities. *Journal of affective disorders* 165 (2014), 87–94.

[72] Hyewon Suh, Nina Shahriaree, Eric B Hekler, and Julie A Kientz. 2016. Developing and validating the user burden scale: A tool for assessing user burden in computing systems. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 3988–3999.

[73] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 6 (2001), 520–525.

[74] Michael Van Ameringen and Peter Mancini, Catherine anFarvolden. 2003. The impact of anxiety disorders on educational achievement. *Journal of anxiety disorders* 17, 5 (2003), 561–571.

[75] Thea F Van de Mortel et al. 2008. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The* 25, 4 (2008), 40.

[76] Rui Wang, Min SH Aung, Saeed Abdullah, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A Scherer, et al. 2016. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 886–897.

[77] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 3–14.

[78] Rui Wang, Emily A. Scherer, Megan Walsh, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, and John Kane. 2017. Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sept. 2017), 1–24.

[79] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.

[80] Ian R White, Patrick Royston, and Angela M Wood. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine* 30, 4 (2011), 377–399.

[81] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K Villalba, Janine M Dutcher, Michael J Tumminia, Tim Althoff, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–33.

[82] Shen Yan, Homa Hosseinmardi, Hsien-Te Kao, Shrikanth Narayanan, Kristina Lerman, and Emilio Ferrara. 2019. Estimating individualized daily self-reported affect with wearable sensors. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–9.

[83] Yuchao Zhou, Suparna De, Wei Wang, Ruili Wang, and Klaus Moessner. 2018. Missing Data Estimation in Mobile Sensing Environments. *IEEE Access* 6 (2018), 69869–69882.

[84] Xiaofeng Zhu, Shichao Zhang, Zhi Jin, Zili Zhang, and Zhuoming Xu. 2010. Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering* 23, 1 (2010), 110–121.