Prediction of Mood Instability with Passive Sensing

MEHRAB BIN MORSHED, Georgia Institute of Technology, USA
KOUSTUV SAHA, Georgia Institute of Technology, USA
RICHARD LI, University of Washington, USA
SIDNEY K. D'MELLO, University of Colorado Boulder, USA
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, USA
GREGORY D. ABOWD, Georgia Institute of Technology, USA
THOMAS PLÖTZ, Georgia Institute of Technology, USA

Mental health issues, which can be difficult to diagnose, are a growing concern worldwide. For effective care and support, early detection of mood-related health concerns is of paramount importance. Typically, survey based instruments including Ecologically Momentary Assessments (EMA) and Day Reconstruction Method (DRM) are the method of choice for assessing mood related health. While effective, these methods require some effort and thus both compliance rates as well as quality of responses can be limited. As an alternative, We present a study that used passively sensed data from smartphones and wearables and machine learning techniques to predict mood instabilities, an important aspect of mental health. We explored the effectiveness of the proposed method on two large-scale datasets, finding that as little as three weeks of continuous, passive recordings were sufficient to reliably predict mood instabilities.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; **Empirical studies in ubiquitous and mobile computing**.

Additional Key Words and Phrases: Smartphone Sensing; Wearable Sensing; Mood Instability; Mental Health; Early Intervention

ACM Reference Format:
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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 75 (September 2019), 21 pages. https://doi.org/10.1145/3351233

1 INTRODUCTION

Situated communities consist of geographically co-located, diverse, and close-knit communities of individuals, who share distinctive social ties [57]. Understanding the well-being of such communities can foster the adoption of preemptive steps to facilitate the psychological needs of individuals and the communities that they are part of. Understanding psychosocial well-being can help design informed and tailored care and intervention strategies to

Authors' addresses: Mehrab Bin Morshed, Georgia Institute of Technology, Atlanta, GA, 30332, USA, mehrab.morshed@gatech.edu; Koustuv Saha, Georgia Institute of Technology, Atlanta, GA, 30332, USA, koustuv.saha@gatech.edu; Richard Li, University of Washington, Seattle, WA, 98195, USA, lichard@uw.edu; Sidney K. D'Mello, University of Colorado Boulder, Boulder, USA, sidney.dmello@colorado.edu; Munmun De Choudhury, Georgia Institute of Technology, School of Interactive Computing, Atlanta, GA, 30332, USA, mchoudhu@cc.gatech.edu; Gregory D. Abowd, Georgia Institute of Technology, School of Interactive Computing, Atlanta, GA, 30332, USA, abowd@gatech.edu; Thomas Plötz, Georgia Institute of Technology, School of Interactive Computing, Atlanta, GA, 30332, USA, thomas.pleotz@gatech.edu.

2474-9567/2019/9-ART75 $15.00
https://doi.org/10.1145/3351233

proactively prevent the onset of mental health challenges among individuals and, as a result, communities at large.

Mental health challenges account for almost one-half of the disease burden in the United States alone [25], and the country spends over 200B USD per year to treat mental disorders [48]. Most lifetime mental disorders appear by the age of 24 [27], and when developed in crucial periods of transition to adulthood impedes an individual's psychosocial functioning, vocational development, and access to social capital [25]. Furthermore, if left untreated, mental health challenges can negatively impact academic success, productivity, and social relationships [28, 67].

Mood is a vital construct of an individual's mental health. Long or short term changes in the psychological state of an individual, such as anxiety or depression, are usually reflected by a corresponding mood change [3, 8, 35, 42] For example, consistent negative affective state is ine of the a diagnostic criterion for depression [8], and frequent mood swings are symptoms of bipolar disorder (BPD) [2, 3]. The lack of temporal stability in a person's mood is defined as mood instability and it can be of clinical significance [36]. For example, depression, anxiety, life satisfaction are all associated with instability of both positive and negative moodes [20, 31]. Hence, the measurement of mood instability can be a crucial component towards understanding and treating various mental health outcomes.

One of the most common methods for measuring mood instability is to ask individuals to respond to self-report questionnaires (e.g., Affective Lability Scale [21], Affect Intensity Scale [33], or General Emotional Dysregulation Measure [39]). Though generally reliable, surveys have a number of biases that reduce their validity, including memory recall bias, social-desirability bias, frame-of-reference bias, amongst others. An alternative to self-report questionnaires is situated active measurement approaches through Ecological Momentary Assessments (EMA), i.e., micro-questionnaires that actively probe an individual via electronic prompts [50, 66]. EMA questions are designed to *repeatedly* capture *real-world data* in *real-time* in naturally occurring contexts [60]. However, it is not without its limitations, most severe being the fact that it can be disruptive so cannot be used in perpetuity.

What is needed is a passive method to model mood states at scale. Smartphones provide a viable option with approximately 96% and 92% of US young adults between the age of 18-29 and 30-49, respectively, owning a smartphone[44] . This widespread adoption of ubiquitous computing technologies, specifically smartphones, creates an opportunity for using passive sensing modalities to assess aspects of mental health (e.g., mood instability). Smart devices also allow for combining passive sensing with active user querying through EMAs [51, 68, 69].

In this article, we assess mood instability of individuals in situated communities (where the individuals are geographically colocated [57] and share distinctive social ties, such as students on college campuses, and employees at workplaces by using a combination of passive sensing modalities (e.g., smartphones, wearables) and an automated classification approach. Specifically, our contributions are as follows:

(1) We developed a model to predict mood instability only using passively sensed data from both smartphone sensors and wearable sensors of individuals in situated communities.
(2) Based on the evidence that mood instability might be a relatively stable trait [36], we present computational methods to infer mood stability from three weeks of actively queried EMA responses or three weeks of passively collected sensor data. We validate our findings with two different kinds of situated communities: college students on campus and peers in workplaces.
(3) We discuss the implications of our results for situated communities, and also illustrate the privacy and ethical implications of collecting and using passive sensing data at scale.

2 RELATED WORK

We first define mood instability and its use as an indicator of mental health. We then examine related work that highlights the difficulty in accurately capturing mood instability. We discuss the use of short and frequent self-report surveys, so-called Ecological Momentary Assessments (EMAs) for capturing other mental health conditions, and how it could be applied to mood instability. Finally, we review related work on using passive sensing to address limitations of EMAs.

2.1 Mood Instability and Mental Health

Mood instability is often referred to by clinicians and psychologists as affective lability, emotional instability, affective and emotional dysregulation, or mood swings [36]. Even though it has been widely described in the psychology literature, there is a lack of agreement in its precise definition [34, 65, 71], though this is true of other psychological constructs as well. Mood instability encompasses a variety of distinct features such as "frequent affective category shifts, disturbances in affect intensity, overdramatic expressions, excessive reactivity to psychological cues, delayed return to emotional baseline", etc [29]. According to Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [9], in the Borderline Personality Disorder population (BPD), mood instability reflects "marked reactivity of mood" including intense irritability and intense anxiety, which might last a few hours or rarely more than a few days. Mood instability is identified as a common feature of many mental health conditions. For example, it is one of the widely agreed upon symptoms for BPD [9, 16]. In addition, studies indicate that of all the BPD diagnostic criteria, mood instability is the strongest predictor of suicidal behavior [72]. Mood instability is strongly associated with attention deficit hyperactivity disorder (ADHD), and it has argued that mood instability should be considered as a diagnostic criteria for ADHD [62]. Mood instability is also frequent in depression[7] and anxiety disorder [5].

Despite these well known mental health associations, measuring and characterizing mood instability has proven to be challenging [19]. Traditional methods of measuring mood instability rely on respondents' recall, which is captured through surveys or interviews [65]. One of the most commonly used clinical survey instruments for quantifying mood instability in BLPD and other psychiatric disorders [22, 30] is the Affective Lability Scale (ALS)[21]. ALS asks respondents to rate a statement on how closely it characterizes them (e.g., "One minute I can be feeling OK, and then the next minute I'm tense, jittery, and nervous."). First, there is no mention of how much historical context the participants should consider while answering this question so it can vary from participant to participant ; Second, individuals are most likely to recall experiences that are consistent with the current mood state. Hence, they might suffer from recall bias, in which they are biased when recalling past experiences, especially when they are asked to aggregate moods or experiences over a long period of time [24]. Previous research suggests that the responses to instruments, which attempt to capture retrospective mood, are primarily influenced by the peak and end of an affectively intense period [18]. Hence, it is very unlikely that the participants can report the variation in their own mood if they are asked to reflect on it after a long period of time. In this paper, we address some of these challenges by investigating whether we can assess mood instability using passively sensed data from smartphones and social media.

2.2 Sensing Mental Health with EMA

Ecological Momentary Assessment (EMA) have emerged as a popular technique [60]. EMA items are short questions designed to capture in-situ real time information about a person's experience. This approach is also often referred to as an Experience Sampling Method (ESM) [58]. EMAs address some of the limitations of survey-based approaches. One of the advantages of EMAs is that the timing can be varied in that EMAs can be gathered at a certain time, after a certain event, or a combination of both [65]. In addition, EMAs can vary based on the platform, including text messages, voice calls [59, 61], paper-pencil diaries, smart devices [50, 66] etc.. Hence, EMA provides several advantages over traditional methods (e.g., surveys, interviews, etc.) when a dynamic psychological process (e.g., mood) is of interest to researchers [45]. In fact, it has been used widely as a tool for in situ measurement of affective phenomenon such as mood [60]. Furthermore, EMAs can also be configured to record context during measurement. For example, characteristics of the environment (e.g., location, time, etc.)

can be recorded while an EMA is being answered [65]. Finally, EMAs are easier to answer compared to surveys simply due to their shorter length and are also less prone to recall bias [60].

2.3 Sensing Mental Health with Passive Sensors and EMAs

Researchers from various disciplines have used EMAs alongside passive sensors for assessing the mental health of their target populations [53, 68, 69]. For example, Wang et al. [68, 69] gathered mental health data on mood, stress, among other measures, of (Dartmouth) college students using EMAs. They also recorded data about students' daily activities (e.g., walking, conversation) using passive sensors and found a correlation between the two data sources [68]. Saha et al. inferred mood instability of a Georgia Tech student sample based on EMA and social media data [53]. Several similar studies on various samples have been conducted [1, 14, 51], using EMAs to collect ground truth data of mental health markers (e.g., stress, mood, depression, etc).

Even though EMAs have several advantages for collecting data "in-the-wild", there are several drawbacks for EMA-based studies [58]. Since these studies usually prompt their participants frequently, they pose a challenge of response burden for the participants [63]. This creates a trade-off between increasing the granularity of data collection and reducing data burden on the participants. In addition, recruiting participants over long periods of time incurs substantial logistical and financial constraints [19]. Moreover, low response rates in EMA-based studies can be a major limitation. Heron et al. conducted a survey of EMA-based studies and found that the survey completion rate was only 76% [23]. Even when response rates are good, this does not imply that the responses are of good quality [11].

Passing sensing technologies might provide an attractive alternative to EMAs. These technologies are becoming more prevalent with the advent of smart devices (e.g., smartphones, smartwatches), especially in college campuses in the US. According to a recent survey, 91% of US-based college students own smartphones [44]. In addition, social media is also popular medium of expression for college students [43], which presents a unique opportunity to investigate how, in such a technology-centric community, we can leverage their choice of technology to assess their mental health. In this paper, we leverage such passively sensed data collected via ubiquituous technologies to predict mood instability levels.

Despite the advantages of passive sensing data in terms of their density and high fidelity, this data can only be collected in a prospective fashion. That is, they are typically only collected through the duration of a study. In addition, passive sensing streams are unable to capture the external or surrounding factors that can potentially influence the fluctuations on an individual's mood.

On the other hand, social media enables us to collect historical and longitudinal data that is self-recorded in the present. Social media doubles up as a verbal sensor, and in the psychology and psycholinguistics literature, psychological health states can be inferred via language [12]. Drawing on that, in recent years, social media have been employed as a passive sensor in mental health studies [15, 55]. When considered via the lens of the Social Ecological Model [10, 52], the attributes of individuals can be considered to be deeply embedded in the complex interplay between an individual, their relationships, the communities they belong to, and the societal factors. Social media provides a passive mechanism to gather quantifiable signals about the social ecological dimensions relating to an individual's behavior. Drawing on this theoretical construct and a rich body of prior work, this paper incorporates similar situated community-specific temporal mood to control for the contextual effects on individuals' mood instability. In this paper, we add to this body of work by investigating how passive sensing can be effective for inferring mood instability in two situated communities — college campuses and workplaces.

3 STUDY AND DATASETS

In order to address our research questions as formulated in the previous section, we conducted a study based on the StudentLife dataset [1] and the Tesserae project.

[1] http://studentlife.cs.dartmouth.edu/dataset.html

Table 1. Student Distribution according to their Academic Year in the StudentLife Dataset

Student Category	**# Students**
Undergraduate Students	
Freshmen	2
Sophmores	6
Juniors	14
Seniors	8
Graduate Students	
1st Year Master's	14
2nd Year Master's	1
Ph.D.	3
Total	48

Table 2. Demographic Information of Participants in the StudentLife Dataset

Gender	**# Students**
Male	38
Female	10

(a) Gender Distribution of StudentLife Dataset

Ethnic Identity	**# Students**
Caucasian	23
Asian	23
African American	2

(b) Ethnic Identity of the Participants in the StudentLife Dataset

3.1 StudentLife Dataset

Wang et al. collected and released the StudentLife dataset as a part of their research effort on inferring mental, physical, and academic well-being of students on a US college campus (Dartmouth University) through smartphone-based data sensing and analysis [68]. The research team recruited 60 students to participate for a period of 10 weeks during the Spring 2013 semester. Among these, five students dropped out of the class, and seven students did not continue with the study, resulting in 48 students completing the study. Table 1 and 2 summarizes the dataset through its descriptive statistics.

3.1.1 Types of Data. The StudentLife dataset consists of three types of data: *i)* passive sensor data; *ii)* Ecological Momentary Assessment (EMA) data; and *iii)* survey data. In what follows we will provide an overview of the dataset as it is of relevance for the work presented in this paper.

Passively Sensed Data An Android application ("StudentLife", developed by the Dartmouth research team) collected and stored sensor data from students' smartphones. The research team used a range of computational methods for inferring higher level activities (e.g., conversation, activity, etc.) from raw sensor data. Table 3 lists the types of passively sensed information that were calculated from various sensors and shared as a part of the public dataset.

Ecological Momentary Assessment (EMA) Data Participants were prompted to answer EMA items multiple times a day. These items asked about the in-situ experience of the students with respect to psychological (e.g., mood, stress, etc.) and behavioral measures, such as the number of people the study participants encountered, sleep duration, amount of time spent on different activities, etc. Through these EMAs, the research team also deployed the photographic affect meter (PAM) [46] to record in-situ mood reports. PAM is a validated instrument for capturing self-reported moods of people, which is based on Russel et al.'s circumplex model of affect [49].

Table 3. Passively Sensed Data in the StudentLife Dataset

Sensed Phenomenon	Sensor Modalities
Inferred Activity	Accelerometer
Audio	Microphone
Inferred Conversation	Audio Data
Sociability	Bluetooth
Darkness	Light Sensor
Indoor Mobility	WiFi Data
Outdoor Mobility	GPS
Phone Charging Event	
Phone Locking Event	

Table 4. Overview of the passively sensed data as recorded though StudentLife.

Measure	Mean (Days)	Median (Days)	Stdev. (Days)
Indoor Mobility	56.31	58	10.82
Activity	59.49	63	9.21
Conversation	56.10	58	10.10
Outdoor Mobility	58.47	63	9.60
Photographic Affect Meter	46.76	51	19.11

Survey Data The third type of actively queried data comprises responses to the following validated mental health questionnaires, which were administered before and after the study. However, none of these survey data was used for this study.

3.2 The Tesserae Project

Our second dataset comes from the Tesserae study that recruited 757 participants[2] who are information workers in cognitively demanding fields (e.g. engineers, consultants, managers) in the U.S. [37, 38, 52, 56]. The participants were enrolled from January 2018 through July 2018. The study was approved by Institutional Review Board at the University of Notre Dame who was the lead institution.

Participants responded to a set of self-reported survey questionnaires at the onset of the study and provided passively sensed data through four major sensing streams : bluetooth beacons; wearable; smartphone agent; and social media. They were provided with an informed-consent document with descriptions of each sensing stream and the data collected from each, and they were required to consent to each sensing stream individually. The data was de-identified and stored on secured data servers with limited access privileges.

The enrollment process consisted of responding to a set of initial survey questionnaires related to demographics (age, gender, education, and income). The participant pool consists of 350 males and 253 females, where the average age is 34 years (stdev. = 9.34). The majority of the participants had a Bachelor's (52%) or Master's degree (35%).

Participants were additionally required to answer periodic EMAs once per day for approximately two months. Of present interest is daily affect as measured by the PANAS-Short scale [64].

[2] Note that this was an ongoing study at the time of writing and this paper uses sensed data collected until August 21st, 2018. A randomly selected 154 participants were "blinded at source", and their data is only available for external validation. This paper only analyzes data from the remaining 603 "non-blinded" participants.

Table 5. Descriptive statistics of # days data collected in the Tesserae Dataset

Type of Data	Range (Days)	Median (Days)	Stdev. (Days)
Study Duration	16:205	99	46.7
Study Duration (while PANAS-short was administered)	3:56	48	13
Bluetooth (Entire Study Period)	1:159	37	32.6
Wearable (while PANAS-short was administered)	27:56	46	7.2
Smartphone (while PANAS-short was administered)	35:56	48	5.9

To passively collect data about participants' behavior, this study deployed three continual sensing streams: bluetooth beacons, smartphone application, and wearable. However, we used two of these streams, which are explained below:

Wearable Participants were provided with a fitness band (Garmin Vivosmart [4]), which they would wear throughout the day. The wearable continually tracks health measures, such as heart rate, stress, and physical activity in the form of sleep, footsteps, and calories burnt.

Smartphone Application A modified verions of the Student Life application [68] was installed on participants' spersonal martphones (android and iPhones). This application tracks their phone use such as lock behavior, call durations, and uses mobile sensors to track their mobility and physical activity.

The participants were enrolled over 6 months (January to July 2018) in a staggered fashion, averaging 111 days of data per participant. Table 5 reports the descriptive statistics of the number of days of passively sensed data that we collected per participant through each of the sensor streams. We obtained an average of 108 days data per participant from the wearable and 101 days/participants of data from the phone application. When limited to days when the PANAS-short EMA was administered, the average available days of data was 49 and 51 for wearable and smartphone, respectively

4 METHODOLOGY

In this section, we discuss how we estimate mood instability from self-reports of individual's affective states, how we excluded participants based on the amount of availabe data, how we calculated features, and finally our approach to the predicting mood instability score.

4.1 Calculating Mood Instability

The participants, in the StudentLife study, responsed to the PAM EMAs by selecting the picture which best represents their mood at that particular time. Since, valence and arousal can have four distinct integer values: $\{-2, -1, +1, +2\}$ [46], PAM can represent 16 distinct mood states. Table 6 illustrates those 16 mood states and their corresponding valence and arousal values.

The participants, in the Tesserae Project, reported affect using the PANAS-Short scale. PANAS-Short measures five positive (alert, excited, enthusiastic, inspired, determined) and five negative (distressed, upset, scared, afraid, nervous) emotions on a scale of 1 (low) to 5 (high).

We quantify each of these affect responses on the two dimensions of Russell's Circumplex [47], valence and arousal, using the Affective Norms for English Words (ANEW) lexicon [40]. ANEW is an affect dictionary, curated after rigorous psychometric studies that contains a list of over a thousand affect categories, and their associated valence and arousal scores, and has been used in prior work to understand expressions of mood and affect [57].

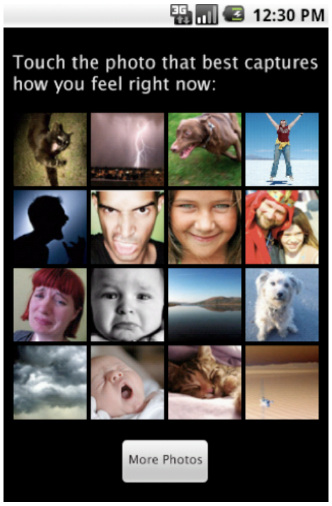
Recall that EMAs were scheduled at random time throughout the day for the StudentLife data set. To calculate mood instability, it is necessary to compute successive differences of valence and arousal responses of the participants over the entire study period. Hence, we adopted a method proposed by Jahng et al. [26], and computed the Adjusted Successive Difference (ASD) by adjusting both valence and arousal responses with respect to time (See Saha et al. for an example of using this method to infer mood instability of college students based on their social media usage).

The method works as follows. Let x_i be the valence or arousal of a participant's logged mood state at time t_i such that we can compute the ASDs based on Equations 1 and 2:

$$ASD_{i+1} = \frac{x_{i+1} - x_i}{[(t_{i+1} - t_i)/\text{median}(t_{i+1} - t_i)]^{\lambda}} \tag{1}$$

$$\begin{aligned} SSEE(\lambda) &= \sum_i [EAASD_{(t_{i+1}-t_i)}(\lambda) - C(\lambda)]^2 \qquad (2) \\ &= \sum_{i=1}^{N-1} \left\{ E\left\{ \frac{|x_{i+1} - x_i|}{[(t_{i+1} - t_i/\text{median}(t_{i+1} - t_i))]^{\lambda}} \right\} - C(\lambda) \right\}^2 \end{aligned}$$

In Equation 1, λ is chosen by minimizing the cost function, $SSEE(\lambda)$, as defined in Equation 2.



Touch the photo that best captures how you feel right now:

Fig. 1. Screenshot of the Photographic Affect Meter (PAM) deployed on an Android smartphone, which was used for gathering mood reports in the StudentLife dataset

Table 6. Mapping of moods on the PAM scale to valence and arousal values as proposed in [46] while working with the StudentLife dataset

Mood	Valence	Arousal
Afraid	-2	2
Angry	-1	1
Calm	1	-1
Delighted	2	2
Excited	1	2
Frustrated	-2	1
Glad	2	1
Gloomy	-2	-2
Happy	1	1
Miserable	-2	-1
Sad	-1	-1
Satisfied	2	-1
Serene	2	-2
Sleepy	1	-2
Tense	-1	2
Tired	-1	-2

We performed a non-parametric smoothing regression method called lowess [13] to calculate the Expected Adjust Successive Difference (EASD). Afterwards, we calculate the Expected Adjusted Absolute Successive Difference (EAASD). This eliminates the dependency of EASD on the time intervals. $C(\lambda)$ in Equation 2 is the $EAASD(\lambda)$ at median time interval.

$$MIS = MAD(ASD_{valence}) + MAD(ASD_{arousal}) \tag{3}$$

Then, we calculated the mean absolute deviation for both valence and arousal for each participant. We refer to the Mean Absolute Deviation for valence as MAD($\text{ASD}_{\text{valence}}$) and Mean Absolute Deviation for arousal as MAD($\text{ASD}_{\text{arousal}}$). After calculating MAD($\text{ASD}_{\text{valence}}$) and MAD($\text{ASD}_{\text{arousal}}$), we add them to obtain the Mood Instability Score (MIS) for each participant (refer to Equation 3). A high MIS (such as for participant 15 (StudentLife Dataset) and participant b (Tesserae project) in Figure 2) indicates that the participant has high variation of

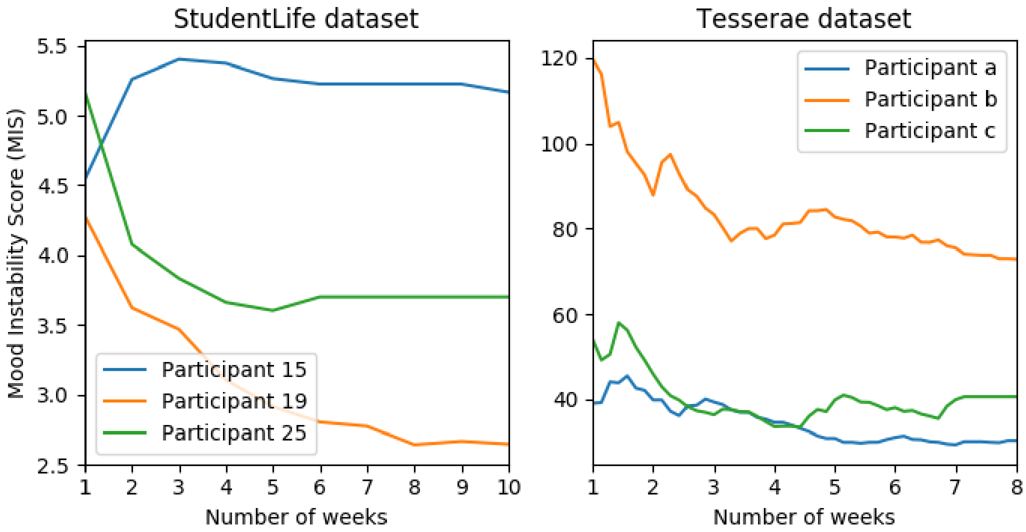


Fig. 2. Example Mood Instability Scores (MIS) of three participants from the StudentLife and Tesserae Project datasets.

mood over the study period, whereas low MIS scores (e.g., participant 19 ((StudentLife Dataset)) and participant a (Tesserae project) in Figure 2) indicates that the participant has low variation of mood.

4.2 Exclusion Criteria

For both datasets, the amounts of collected data varied for all study participants. Note that the focus of this paper is on predicting Mood Instability (MI) using passive sensors, and we calculate MIS (Mood Instability Score) based on Equation 3. In Equation 3, MIS is an estimate of the MI of an individual. Given that MI is a personality trait [36], we investigated whether the numerical estimate of MI (MIS) stabilize (within 5% error margin) after a certain time.

Hence, we varied (incrementally from 1 to 10 in Student Life, and from 1 to 8 in Tesserae project) the number of weeks for mood responses (e.g., PAM and PANAS-short) for each participant and calculated their MIS based on the responses for only those weeks. Then, we calculated the difference of MIS with respect to the original MIS, which is the MIS value comprising all self-reports. Figure 3 illustrates the results of this investigation.

Both average and median errors of MIS with respect to their final values (MIS after 10 weeks and 8 weeks) reduce with the increase in number of weeks. The blue line in Figure 3 highlights 5% error margin. Based on our analysis, at least 3 weeks of consecutive PAM responses and 4 weeks of consecutive PANAS-short responses were needed from the beginning of the study to reliably infer (within 5% error margin) final MISs of each participant. Hence, any participant who did not have at least 3 weeks of consecutive PAM reports or 3 weeks of PANAS-short responses were excluded from further analysis.

4.3 Feature Extraction

For predicting the mood instability of participants, we adopted a regression-based approach since our goal is to estimate the mood instability scores based on the features. In this section we explain which features we extracted from the input data that were then fed into the regression backend.

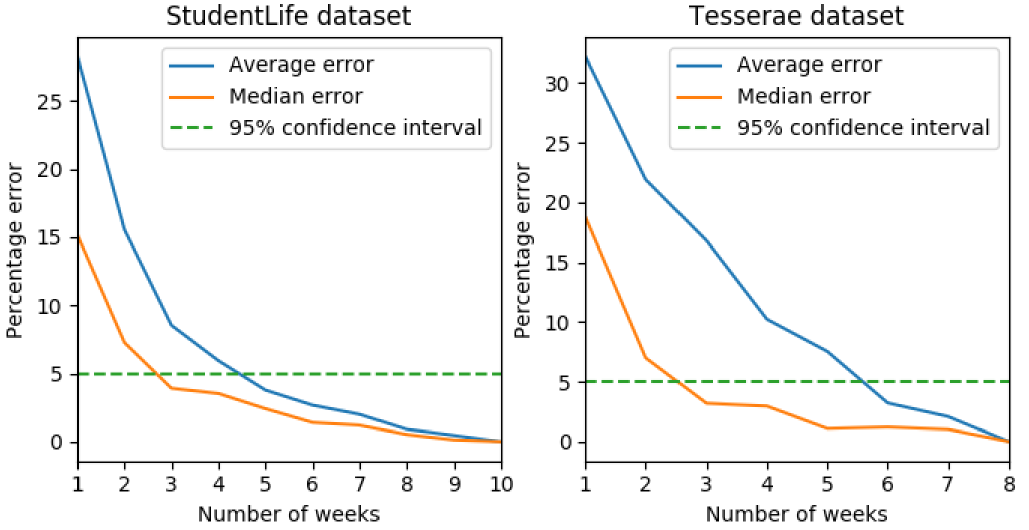


Fig. 3. Mean and median error is calculated by increasing the number of self-reports in terms of weeks and comparing the MIS with the MIS obtained by taking all ten weeks (StudentLife) and eight weeks (Tesserae), respectively, of self-reported data for each participant.

4.3.1 Activity. For the Student Life Dataset, Wang et al. [68] inferred four types of activities based on accelerometer data: *i)* stationary; *ii)* walking; *iii)* running; and *iv)* unknown (but active). The StudentLife application inferred and logged these activities automatically. We calculated the daily active minutes based on these logged data resulting in the number of (physically) "active" (i – iii above) minutes for each participant per day. We calculated the mean and median of this number of active minutes per day over 10 weeks for each participant, and used those two values as features for representing their activity.

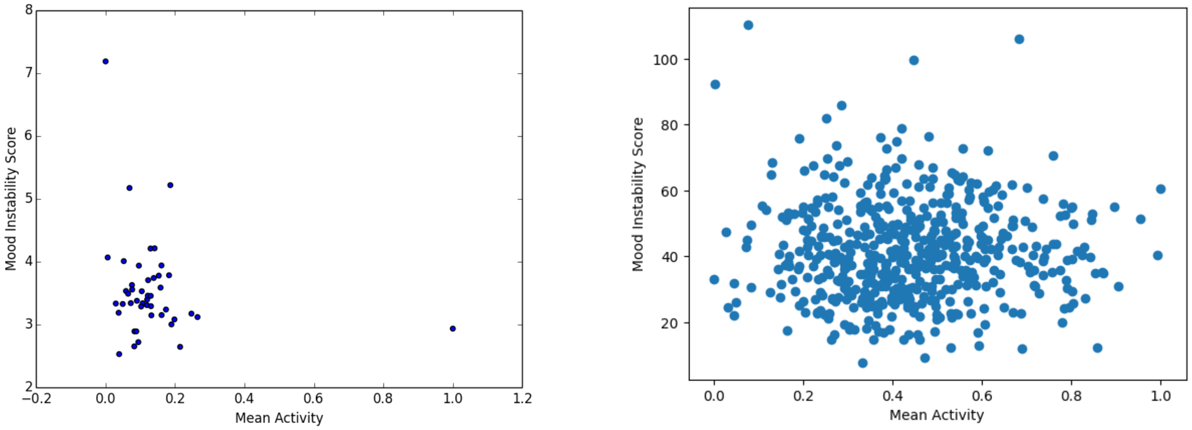
For the data in the Tesserae project, activity was inferred from the Garmin wearable and the smartphone sensor data. However, the mean activity calculated from wearables and smartphones were highly co-linear. Hence, we only used the mean activity inference from the wearable. We used the Garmin data for estimating the duration an individual was active in a day. We calculated the mean and median of the active minutes per day over the 8 weeks for each participant and used these values as features for representing their activity.

4.3.2 Conversation. The StudentLife dataset contained information about inferred conversation of participants. The conversation classifier logged start and end times whenever it inferred that the participant was engaged in a conversation or near a conversation. We calculated the amount of conversation (in minutes) per day for each participant. Subsequently, we calculated the mean and median number of conversation minutes per day over 10 weeks for each participant, and used those two values as features for representing conversations.

There was no conversation data in the Tesserae project.

4.3.3 Indoor Mobility. Wang et al. collected WiFi data from participants' phones, and later mapped these to building locations. This helped us in estimating the mobility of each participant, since a change in location would mean that the participant is mobile. We calculated the number of changes in locations in a day and used that number as indoor mobility for each participant. We take the mean and median of indoor mobility over 10 weeks for each participant, and used those two values as features for representing their indoor mobility.

There was no indoor mobility data in the Tesserae project.



(a) Scatter plot with respect to mean activity in the Student Life dataset

(b) Scatter plot with respect to mean activity in the Tesserae Project

Fig. 4. Scatter plots of two datasets with respect to one of the features, mean activity of participants, used for modeling the regressors

4.3.4 Outdoor Mobility. Mobility was also inferred from GPS. It is a reliable estimator of outdoor mobility. The GPS data in the dataset had three movement inferences: *i)* moving; *ii)* stationary; and *iii)* no inference. We used the "moving" inferences from GPS to calculate the outdoor mobility per day for each participant. Afterwards, we calculated the number of mean and median outdoor mobility per day over 10 weeks for each participant, and used those two values as features for representing their outdoor mobility.

There was no outdoor mobility data in the Tesserae project.

4.3.5 Sleep. For the Tesserae project, Garmin objectively inferred the sleep duration of individuals. It helped us estimating the duration an individual was sleeping in 24 hours. We calculated the mean and median of the sleep duration, in minutes, per day over the 8 weeks for each participant and used those values as features for representing their sleep.

There was no objective sleep data in the StudentLife dataset.

4.3.6 Scaling of the Features. Once we calculated features from all modalities, we used min-max scaling to scale each feature between 0 (min) and 1 (max). For each feature, if the original value of any data point was x, the maximum value x_{max}, minimum value is x_{min}, then the scaled value (x_{scaled}) can be calculated as follows:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4}$$

4.4 Automated Assessment, Training, and Testing Protocols

In order to infer the mood instability scores (MIS) of individuals, we used a regression-based (ridge-regression with regularization) approach since scores are continuous. For both datasets, we found that features and mood instability scores can be represented with a linear relationship. Our choice of linear regression compared to non-linear regression was to combat overfitting, which is often the case with non-linear regression. Figure 4 represents the scatter plots of mood instabilty scores of both datasets with respect to mean active number of minutes. For evaluting the prediction of our regressor, we used an evaluation metric called Symmetric Mean Absolute Percentage Error (Equation 5).

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t|+|F_t|)/2} \tag{5}$$

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \tag{6}$$

SMAPE is a variation of mean absolute percentage error (MAPE), which is an error estimation between the predicted values and actual values. However, MAPE is not feasible to calculate when the actual value is 0 (Refer to the denominator in Equation 6). Hence, that cannot be used for meaningful interpretation. That is where SMAPE helps us estimating the percentage of error with respect to the predicted and actual values. In addition, SMAPE is particularly useful for comparing non-negative predicted values in comparison to actual values [32], which is our case. If, for a regressor, a SMAPE score is $n\%$, that should be interpreted as the regressor is predicted a value that within $n\%$ of the final value. Hence, the lower the SMAPE score, the better the regressor.

Our data analysis framework was implemented in Python. Regression models were trained using the scikit-learn machine learning library for Python [3].

5 RESULTS

In this section, we present our results for the two research questions: whether we can predict mood instability with passive sensing, and how early we can predict mood instability.

5.1 Mood Instability Inference

In the previous section, we defined how we calculated the mood instability score for our participants (equation 3). In this section, we present the results of the automated MI estimator for all (remaining) StudentLife and Tesserae Project participants. Table 7 shows the results based on all possible sensing modalities. In this table, A refers to Activity, C refers to Conversation, O refers to Outdoor Mobility, I refers to Indoor Mobility, S refers to sleep data during the study period, Baseline refers to the Baseline regressor. Baseline regressor refers to a classifier that always predicts the mean of the entire sample. The score from baseline predictor serves as an indicator for inferring whether our trained model is learning from the datasets. Note that from this point onward, we will refer to modalities using their one letter code (i.e., I, A, etc.). In addition, combinations of letters—referring to the modalities—without comma (e.g., ACI, AS, etc.) denote the features that were used to train and test the regressor. For example, ACI means that features generated from activity, conversation, and indoor mobility were used for training and testing the classifier. Performance in the following section refers to mean Symmetric Mean Absolute Percentage Error (SMAPE) of leave-one-out validation.

One Modality. The first four rows of Table 7 illustrate the regressor's performance using only one of the sensing modalities on the Student Life dataset. Our results highlight that if we are using only one of the sensing modalities, in StudentLife dataset, for assessment, *I* was the strongest predictor for classifying MI comnpared to other modalities. However, I did not perform significantly better than other single modalities.

For the Tesserae project, both *A* and *S* performed similarly (third last and second to last row in Table 7). There was no difference in their performance. This can be explained with the coefficients we got from the previous section. We found that both activity and sleep show negative correlations with MIS. (Section 5.3) The coefficient values are also very similar. Hence, we do not see any changes in the SMAPE score.

[3] https://www.python.org/

Table 7. Performance with all possible combinations of passive sensing modalities: Activity (A), Conversation (C), Indoor Mobility (I), Outdoor Mobility (O), and Sleep (S). The Symmetric Mean Absolute Percentage Error (SMAPE) scores are reported based on leave-one-out protocol. Lower SMAPE scores correspond to better regressors.

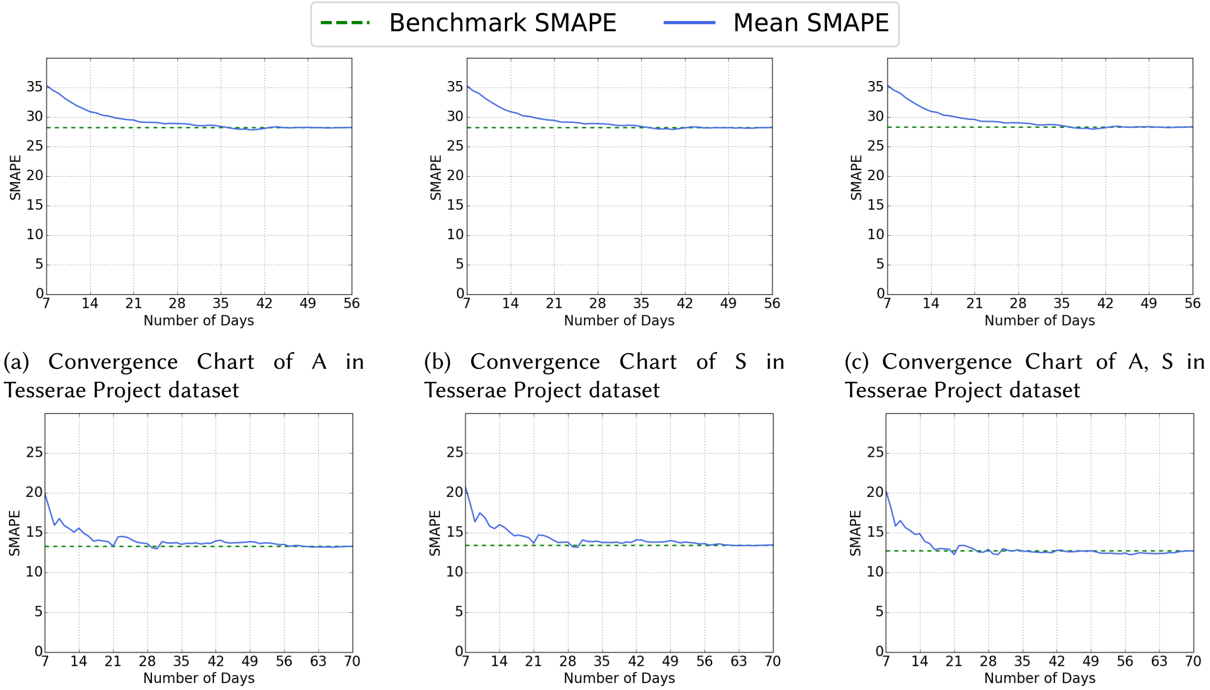
Dataset	Modality	SMAPE Score (%)
StudentLife	Baseline	16.63
	Activity (A)	13.26
	Conversation (C)	13.43
	Indoor Mobility (I)	12.74
	Outdoor Mobility (O)	12.99
	A, I	11.70
	A, C	13.37
	A, O	12.91
	C, I	12.83
	C,O	13.12
	I, O	12.42
	A, C, I	12.8
	A, C, O	13.08
	C, I, O	12.63
	A, I, O	12.39
	A, C, I, O	12.58
Tesserae	Baseline	33.74
	A	28.21
	Sleep (S)	28.22
	A, S	28.30

Two Modalities. For the Student Life dataset, if we take all possible combinations of any two modalities train regressor, we find that some of the combinations work better than individual modalities. For example, *AI* performed best as combined modalities to predict mood instability. In addition, *AI* performed better than *A* or *I* individually. Only in one circumstance, combining two modalities actually hurts the automated assessment procedure compared to their individual performances. Note that, *I* was also the best performer, even though not statistically significant, among the single modalities. We observed that whenever *I* was added to any other modality, it slighty improved the performance. This effect is observed more substantially in the convergence behavior (next section).

For the Tesserae project, we did not see any change in the regression performance when combining the two modalities. We found that for Tesserae project, neither sleep nor activity outperforms each other in terms of SMAPE score. This is due to the fact that both of them have similar, if not the same, effect on the regressor. We have discussed the correlation of sleep and activity from Garmin with respect to the mood instability score. And, correlation coefficients are very close. Hence, we do not see any improvement in the performance of the regressor when we combine these modalities.

Three Modalities. For the Student Life dataset, we evaluated all possible combinations of any three modalities. Some combinations outperform the individual performance of single modalities but none of them were able to outperform AI combination. For example, *ACI* outperformed *A* and *C*, but was not able to outperform *AI*. AIO performed the closest to *AI*

Four Modalities. Afterwards, for the Student Life dataset, we combined every sensing modalities and found that it does not outperform *AI*.



(a) Convergence Chart of A in Tesserae Project dataset

(b) Convergence Chart of S in Tesserae Project dataset

(c) Convergence Chart of A, S in Tesserae Project dataset

(d) Convergence chart of A in Student Life dataset

(e) Convergence chart of C in Student Life dataset

(f) Convergence chart of A, I in Student Life dataset

Fig. 5. Convergence characteristic of our classifier using various combinations of sensing modalities: conversation (C), activity (A), indoor mobility (I), outdoor mobility (O), and sleep (S).

5.2 Early Prediction of Mood Instability

Recall that three weeks of PAM responses and four weeks of PANAS-short responses were sufficient predicting mood instability; this validates the theory that mood instability is a personality trait [36]. However, setting up studies to collect self-reported information has several challenges such as response burden for participants [63], validity of the responses since they are collected very frequently [11], scalability issues [58], etc. Hence, the main objective of our study is to investigate whether early predictions of mood instability can be made based on the analysis of passively sensed data only.

For deriving a measure of an early prediction, we varied the number of weeks for all combinations of sensing modalities for each participant, trained a regression model using a portion of the data and compared the accuracy of our model with the benchmark accuracy for the respective combination of modalities (SMAPE after ten weeks for student life and SMAPE after eight weeks for Tesserae Project). For any combination of sensing modalities, we investigated whether the SMAPE score converges in a stable manner to the final SMAPE score of the respective combination. Stability of the convergence means that the SMAPE plot does deviate from the baseline (dashed green line on each subfigure in Figure 5).

For StudentLife dataset, we found that the convergence behaviors were similar for almost all of the regressors, and that they all converged to the baseline SMAPE score when using a maximum of only three weeks of data. Recall that our best performing modality combination for mood instability regression was Activity and Indoor mobility. We found that *AI*'s convergence plot is closer to the baseline after 3 weeks compared to *A* or *I* alone (Figure 5) The blue line on each plot in Figure 4 represents mean SMAPE scores of the regression model for the

Table 8. Pearson correlation coefficients of the sensed phenomena with respect to mood instability. All coefficient values are statistically significant.

Dataset	Phenomenon	Coefficient
Student Life Project	Active Duration	-0.74
	Indoor Mobility	0.13
	Outdoor Mobility	-0.79
	Conversation	-0.10
Tesserae Project	Active Duration	-0.12
	Sleep Duration	-0.14

respective sensing modalities and the X value for the scores denote how many days of sensor data we used from the beginning of the study.

In the case of Tesserae Project, we found that activity and sleep both had similar convergence patterns, which we discuss in the next subsection by discussing the correlation of these features with respect to mood instability. For this dataset, we found that the SMAPE score drops in a consistent manner and comes close to the baseline SMAPE when using four weeks of data.

While our findings are similar in general for both datasets, we observe a variation in performance across the two datasets. This can be attributed to multiple things. Firstly, the datasets represent mental and physical markers for two different populations: student community and employees at workplace. For the student community, everyone was in the study for the same timeline, which is for an entire semester. However, for the second dataset, participants were recruited over a period of six months. Secondly, the instruments used for collecting the moods of individuals were different (Section 4.1). The student population responded to PAM scale for reporting their mood and workplace participants responded to PANAS-Short scale. These instruments were administered at different frequency as well. Finally, the sensor streams used for collecting sensor data were not the same across participants. For example, sleep data was not present for the student population, and the conversation inference data was not present for the workplace population.

To summarize, using only four weeks sensor data we can reliably infer mood instability in situated communities as demonstrated for both datasets explored in our study.

5.3 Correlation Analysis of Features

We investigated whether the choice of features are correlated with our target variable, mood instability score (Table 8). We found that activity duration, sleep duration, outdoor mobility, conversation was negatively correlated with mood instability score. This validates theoretical grounding that lack of activity [6], sleep[6], sociability (estimation of conversation) contribute to mood instability. The complete set of convergence plots can be found in the appendix.

6 DISCUSSION

As part of our analysis, we found that three weeks of self-reported affect measures is a reliable estimator for inferring one's mood instability. Mood Instability Scores (MIS) are an estimation of the Mood Instability (MI) of individual students, and MI is a personality trait [36]. We used a data driven approach to validate this theory. Based on our results, we highlight two major observations: *1)* it is possible to predict Mood Instability using multi-modal passive sensing data from smartphones and wearables; and *2)* we are able to make an early prediction of MI using less than three weeks of data. In the following subsections, we outline the implications of our results for academics studying mood instability, and for a variety of stakeholders such as campus and workplace administrators, policy-makers, and caregivers.

6.1 Implications for Stakeholders at Situated Communities

Given that individuals have distinctive social ties in situated communities, psychological concerns have the potential to affect both individual as well as collective well-being [55]. This concerns a variety of stakeholders including administrators and policy-makers on college campuses and workplaces. University and office administrators all across the United States have been adopting concrete measures for improving the well-being on campuses [41, 70]. As part of such initiatives, various validated self-tracking tools (e.g., WellTrack [4]) are offered for free to students. Such tools use self-report methods for tracking the mental health of individuals. WellTrack asks respondents to report their in-situ mood. The self-report aspect is likely to create a response burden for individuals, which can lead to low compliance and thus limited effectiveness. We used a data-driven approach to assess the MI in situated communities using only passive sensing. Hence, if MI is of interest to the stakeholders of situated communities such as campus or office administrators, our results highlight the significance of using passive sensing modalities on smartphones and wearables. Since 94% of the young population between 29-34 years in the United States use smartphones [44], they can opt in to share their passive sensing data to the well-being tracking applications, if such applications are designed with the privacy implications discussed in the following section.

The stakeholders have three options if mood instability is of interest to them. First, they can ask the members in community (students or employees) to take surveys (e.g., Affective Lability Survey) to screen them for MI. The second option is to ask them to report their mood for three weeks with a self-reporting application such as WellTrack, and convert these mood report to mood insatbility score. Finally, they can use passively sensed data from the individuals at situated communities for predicting mood instability. The latter two options are the insights from our work. Each of these options have advantages and drawbacks.

6.2 Practical Implications

The focus of our paper has been to establish the potential of passive sensing data to infer a clinically significant mental health measure, mood instability. We have answered this question with two situated communities: college students and workplace employees. Our research have implications at scale. For example, in the events of acute crisis (e.g., gun violence, natural disaster, etc), such subtle trait of individuals can be, with permission, monitored to provide them necessary intervention. Since smartphones and wearables are becoming more and more prevalent, our study has implications for extending beyond our observed situated communities.

Most smartphones are capable of sensing data that we have used in our study. Hence, with consent, it might be feasible to collect longitudinal data passively at scale. However, labeling these passively sensed data with mental health measures is challenging and impractical. Hence, it would be interesting to investigate whether supervised model learned from a controlled study such as ours can be deployed in an unsupervised dataset to reliably label the unlabelled data.

6.3 Implications for Researchers

Researchers have to put in investment (time, money, etc.) for gathering data that can be useful for answering questions related to mental health. Our study showed that, for understanding the trait of mood instability, we need to collect at least three weeks of passive and actively sensed data. Hence, our work's implications for research is to collect at least 3 weeks of data from situated communities of interest. In addition, further investigation could be made to understand what social and psychological aspects influence these three weeks period, and why we need three weeks. Understanding these aspects may help us get even earlier estimation of mood instability, and answer broader but related questions corresponding to an individual's psychological dynamics.

[4] https://welltrack.com/

6.4 Privacy and Ethics

User privacy has always been a concern in any behavior tracking application. Privacy in the context of mental health is even more sensitive since this topic is very little understood and often stigmatized [17]. Hence, privacy concerning tracking psychological data is very sensitive, and it must be well regulated. Note that for inferring MI, we used high-level activity data from each participant (e.g., total number of minutes active in day, total number of minutes in a conversation, etc.). Such questions pose relatively little privacy challenges, especially in contrast to asking someone about specific locations they visited throughout the day. Irrespective of the granularity of such data, strong regulations need to be established first. In addition, ethical decisions about when to intervene, the granularity of intervention (e.g., individual level, community level, etc.) and how such interventions align with the individual preference for receiving intervention needs to be well thought out through ethical review boards on respective college campuses.

6.5 Limitations and Future Work

Social media data has been proven very useful for predicting mood instability in situated communities, and it additionally functions as a verbal sensor [53, 54]. However, we did not include social media data for either of the population in this paper. Future work can complement longitudinal and historical social media data with sensory data to build better models of predicting such psyhological constructs [56].

In addition, these results should not be taken out of context. We do not claim that our results will be reproducible with the same metrics in other campuses in the US. However, using such a data-driven approach and similar study design should produce a consistent result for other colleges such as Dartmouth. We caution against making clinical claims on the basis of our results. In particular, we are unaware of the fraction of our study population who suffer from clinical mood disorders, and our findings do not identify the clinical nature of mood disorders. Our work is aimed at complementing existing methodologies to assess psychological well-being both at individual and collective level. We further acknowledge that our work does not study causality between individual behavior and their mood. Therefore, even though our study finds certain forms of behaviors, activity, and context of individuals to be associated with certain kinds of mood instability, their causal relationship remains to be explored in future research.

ACKNOWLEDGMENTS

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800007. The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copy-right annotation therein.

In addition, we acknowledge the contribution of the StudentLife team from Dartmouth, and thank them for making their dataset publicly available. Finally, we thank the CampusLife team at Georgia Tech for their feedback on our research.

REFERENCES

[1] Phil Adams, Mashfiqui Rabbi, Tauhidur Rahman, Mark Matthews, Amy Voida, Geri Gay, Tanzeem Choudhury, and Stephen Voida. 2014. Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 72–79.

[2] Jules Angst. 2007. The bipolar spectrum. *The British Journal of Psychiatry* 190, 3 (2007), 189–191.

[3] Jules Angst and Giovanni Cassano. 2005. The mood spectrum: improving the diagnosis of bipolar disorder. *Bipolar disorders* 7 (2005), 4–12.

[4] Garmin Health API. 2018. http://developer.garmin.com/health-api/overview/. Accessed: 2018-11-01.

[5] Rudy Bowen, Marilyn Baetz, Judy Hawkes, and Angela Bowen. 2006. Mood variability in anxiety disorders. *Journal of Affective Disorders* 91, 2-3 (2006), 165–170.

[6] Rudy Bowen, Lloyd Balbuena, Marilyn Baetz, and Laura Schwartz. 2013. Maintaining sleep and physical activity alleviate mood instability. *Preventive medicine* 57, 5 (2013), 461–465.

[7] Rudy C Bowen, Jan Mahmood, Ali Milani, and Marilyn Baetz. 2011. Treatment for depression and change in mood instability. *Journal of affective disorders* 128, 1-2 (2011), 171–174.

[8] Timothy A Brown, Bruce F Chorpita, and David H Barlow. 1998. Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of abnormal psychology* 107, 2 (1998), 179.

[9] RJ Castillo, DJ Carlat, T Millon, CM Millon, S Meagher, S Grossman, R Rowena, J Morrison, American Psychiatric Association, et al. 2007. *Diagnostic and statistical manual of mental disorders*. Washington, DC: American Psychiatric Association Press.

[10] Ralph Catalano. 1979. *Health, behavior and the community: An ecological perspective*. Pergamon Press New York.

[11] 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), 3.

[12] Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication* (2007), 343–359.

[13] William S Cleveland. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association* 74, 368 (1979), 829–836.

[14] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. 2008. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1797–1806.

[15] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.

[16] Fifth Edition et al. 2013. Diagnostic and statistical manual of mental disorders. *Arlington: American Psychiatric Publishing* (2013).

[17] Daniel Eisenberg, Marilyn F Downs, Ezra Golberstein, and Kara Zivin. 2009. Stigma and help seeking for mental health among college students. *Medical Care Research and Review* 66, 5 (2009), 522–541.

[18] Barbara L Fredrickson. 2000. Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion* 14, 4 (2000), 577–606.

[19] Jon Froehlich, Mike Y Chen, Sunny Consolvo, Beverly Harrison, and James A Landay. 2007. MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on Mobile systems, applications and services*. ACM, 57–70.

[20] June Gruber, Aleksandr Kogan, Jordi Quoidbach, and Iris B Mauss. 2013. Happiness is best kept stable: Positive emotion variability is associated with poorer psychological health. *Emotion* 13, 1 (2013), 1.

[21] Philip D Harvey, Barbara R Greenberg, and Mark R Serper. 1989. The affective lability scales: development, reliability, and validity. *Journal of clinical psychology* 45, 5 (1989), 786–793.

[22] Chantal Henry, Vivian Mitropoulou, Antonia S New, Harold W Koenigsberg, Jeremy Silverman, and Larry J Siever. 2001. Affective instability and impulsivity in borderline personality and bipolar II disorders: similarities and differences. *Journal of psychiatric research* 35, 6 (2001), 307–312.

[23] 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.

[24] Michael R Hufford, Saul Shiffman, Jean Paty, and Arthur A Stone. 2001. Ecological Momentary Assessment: Real-world, real-time measurement of patient experience. (2001).

[25] Justin Hunt and Daniel Eisenberg. 2010. Mental health problems and help-seeking behavior among college students. *Journal of adolescent health* 46, 1 (2010), 3–10.

[26] Seungmin Jahng, Phillip K Wood, and Timothy J Trull. 2008. Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological methods* 13, 4 (2008), 354.

[27] Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, and Ellen E Walters. 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* 62, 6 (2005), 593–602.

[28] Ronald C Kessler, Cindy L Foster, William B Saunders, and Paul E Stang. 1995. Social consequences of psychiatric disorders, I: Educational attainment. *American journal of psychiatry* 152, 7 (1995), 1026–1032.

[29] Harold W Koenigsberg. 2010. Affective instability: toward an integration of neuroscience and psychological perspectives. *Journal of Personality Disorders* 24, 1 (2010), 60–82.

[30] Harold W Koenigsberg, Philip D Harvey, Vivian Mitropoulou, James Schmeidler, Antonia S New, Marianne Goodman, Jeremy M Silverman, Michael Serby, Frances Schopick, and Larry J Siever. 2002. Characterizing affective instability in borderline personality disorder. *American Journal of Psychiatry* 159, 5 (2002), 784–788.

[31] Nicole CM Korten, Hannie C Comijs, Femke Lamers, and Brenda WJH Penninx. 2012. Early and late onset depression in young and middle aged adults: differential symptomatology, characteristics and risk factors? *Journal of affective disorders* 138, 3 (2012), 259–267.

[32] Vladik Kreinovich, Hung T Nguyen, and Rujira Ouncharoen. 2014. How to estimate forecasting quality: a system-motivated derivation of symmetric mean absolute percentage error (SMAPE) and other similar characteristics. (2014).

[33] Randy J Larsen, Ed Diener, and Robert A Emmons. 1986. Affect intensity and reactions to daily life events. *Journal of personality and social psychology* 51, 4 (1986), 803.

[34] Paul S Links, Rahel Eynan, Marnin J Heisel, and Rosane Nisenbaum. 2008. Elements of affective instability associated with suicidal behaviour in patients with borderline personality disorder. *The Canadian Journal of Psychiatry* 53, 2 (2008), 112–116.

[35] Steven Marwaha, Matthew R Broome, Paul E Bebbington, Elizabeth Kuipers, and Daniel Freeman. 2013. Mood instability and psychosis: analyses of British national survey data. *Schizophrenia bulletin* 40, 2 (2013), 269–277.

[36] S Marwaha, Z He, M Broome, SP Singh, J Scott, J Eyden, and D Wolke. 2014. How is affective instability defined and measured? A systematic review. *Psychological medicine* 44, 9 (2014), 1793–1808.

[37] Stephen M Mattingly, Julie M. Gregg, Pino Audia, Ayse Elvan Bayraktaraglu, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Sidney K D'Mello, Anind K Dey, Ge Gao, Krithika Jagannath, Kaifeng Jiang, Suwen Lin, Liu Qiang, Gloria Mark, Gonzalo J Martinez, Kizito Masaba, Shayan Mirjafari, Edward Moskal, Raghu Mulukutla, Kari Nies, Manikanta D Reddy, Pablo Robles-Granda, Koustuv Saha, Anusha Sirigiri, and Aaron Striegel. 2019. The Tesserae Project: Large-Scale, Longitudinal, In Situ, Multimodal Sensing of Information Workers. In *CHI Ext. Abstracts*.

[38] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, et al. 2019. Differentiating Higher and Lower Job Performers in the Workplace Using Mobile Sensing. *Proc. IMWUT* (2019).

[39] Christina E Newhill, Edward P Mulvey, and Paul A Pilkonis. 2004. Initial development of a measure of emotional dysregulation for individuals with cluster B personality disorders. *Research on Social Work Practice* 14, 6 (2004), 443–449.

[40] Finn Årup Nielsen. 2011. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, Heraklion, Crete, Greece, May 30, 2011*. 93–98. http://ceur-ws.org/Vol-718/paper_16.pdf

[41] University of Washington Bothell. [n.d.]. Self Help Resources. https://www.uwb.edu/studentaffairs/counseling/self-help-resources

[42] Rashmi Patel, Theodore Lloyd, Richard Jackson, Michael Ball, Hitesh Shetty, Matthew Broadbent, John R Geddes, Robert Stewart, Philip McGuire, and Matthew Taylor. 2015. Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. *BMJ open* 5, 5 (2015), e007504.

[43] Tiffany A Pempek, Yevdokiya A Yermolayeva, and Sandra L Calvert. 2009. College students' social networking experiences on Facebook. *Journal of applied developmental psychology* 30, 3 (2009), 227–238.

[44] Pew Research Center: Internet, Science & Tech. 2018. Demographics of Mobile Device Ownership and Adoption in the United States. http://www.pewinternet.org/fact-sheet/mobile/. Accessed: 14-11-2018.

[45] Thomas M Piasecki, Michael R Hufford, Marika Solhan, and Timothy J Trull. 2007. Assessing clients in their natural environments with electronic diaries: Rationale, benefits, limitations, and barriers. *Psychological assessment* 19, 1 (2007), 25.

[46] John P Pollak, Phil Adams, and Geri Gay. 2011. PAM: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 725–734.

[47] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734.

[48] Charles Roehrig. 2016. Mental disorders top the list of the most costly conditions in the United States: $201 billion. *Health Affairs* 35, 6 (2016), 1130–1135.

[49] Daniel W Russell. 1996. UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure. *Journal of personality assessment* 66, 1 (1996), 20–40.

[50] Michael A Russell, Lin Wang, and Candice L Odgers. 2016. Witnessing substance use increases same-day antisocial behavior among at-risk adolescents: Gene–environment interaction in a 30-day ecological momentary assessment study. *Development and psychopathology* 28, 4pt2 (2016), 1441–1456.

[51] 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).

[52] Koustuv Saha, Ayse Elvan Bayraktaraglu, Andrew T Campbell, Nitesh V Chawla, Munmun De Choudhury, Sidney K D'Mello, Anind K Dey, et al. 2019. Social Media as a Passive Sensor in Longitudinal Studies of Human Behavior and Wellbeing. In *CHI Ext. Abstracts*. ACM.

[53] 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.

[54] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and Psychological Effects of Hateful Speech in Online College Communities. In *WebSci*.

[55] Koustuv Saha and Munmun De Choudhury. 2017. Modeling Stress with Social Media Around Incidents of Gun Violence on College Campuses. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 92 (Dec. 2017), 27 pages.

[56] Koustuv Saha, Manikanta D Reddy, Vedant Das Swain, Julie M Gregg, Ted Grover, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, et al. 2019. Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior. In *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*.

[57] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses. In *ICWSM*.

[58] 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.

[59] Lori N Scott, Stephanie D Stepp, Michael N Hallquist, Diana J Whalen, Aidan GC Wright, and Paul A Pilkonis. 2015. Daily shame and hostile irritability in adolescent girls with borderline personality disorder symptoms. *Personality Disorders: Theory, Research, and Treatment* 6, 1 (2015), 53.

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

[61] Jennifer S Silk, Ronald E Dahl, Neal D Ryan, Erika E Forbes, David A Axelson, Boris Birmaher, and Greg J Siegle. 2007. Pupillary reactivity to emotional information in child and adolescent depression: links to clinical and ecological measures. *American Journal of Psychiatry* 164, 12 (2007), 1873–1880.

[62] Caroline Skirrow, Gráinne McLoughlin, Jonna Kuntsi, and Philip Asherson. 2009. Behavioral, neurocognitive and treatment overlap between attention-deficit/hyperactivity disorder and mood instability. *Expert review of neurotherapeutics* 9, 4 (2009), 489–503.

[63] 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*. ACM, 3988–3999.

[64] Edmund R Thompson. 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of cross-cultural psychology* 38, 2 (2007), 227–242.

[65] Timothy J Trull, Marika B Solhan, Sarah L Tragesser, Seungmin Jahng, Phillip K Wood, Thomas M Piasecki, and David Watson. 2008. Affective instability: Measuring a core feature of borderline personality disorder with ecological momentary assessment. *Journal of abnormal psychology* 117, 3 (2008), 647.

[66] Eeske van Roekel, Luc Goossens, Maaike Verhagen, Sofie Wouters, Rutger CME Engels, and Ron HJ Scholte. 2014. Loneliness, affect, and adolescents' appraisals of company: An experience sampling method study. *Journal of Research on Adolescence* 24, 2 (2014), 350–363.

[67] Philip S Wang, Gregory E Simon, Jerry Avorn, Francisca Azocar, Evette J Ludman, Joyce McCulloch, Maria Z Petukhova, and Ronald C Kessler. 2007. Telephone screening, outreach, and care management for depressed workers and impact on clinical and work productivity outcomes: a randomized controlled trial. *Jama* 298, 12 (2007), 1401–1411.

[68] 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*. ACM, 3–14.

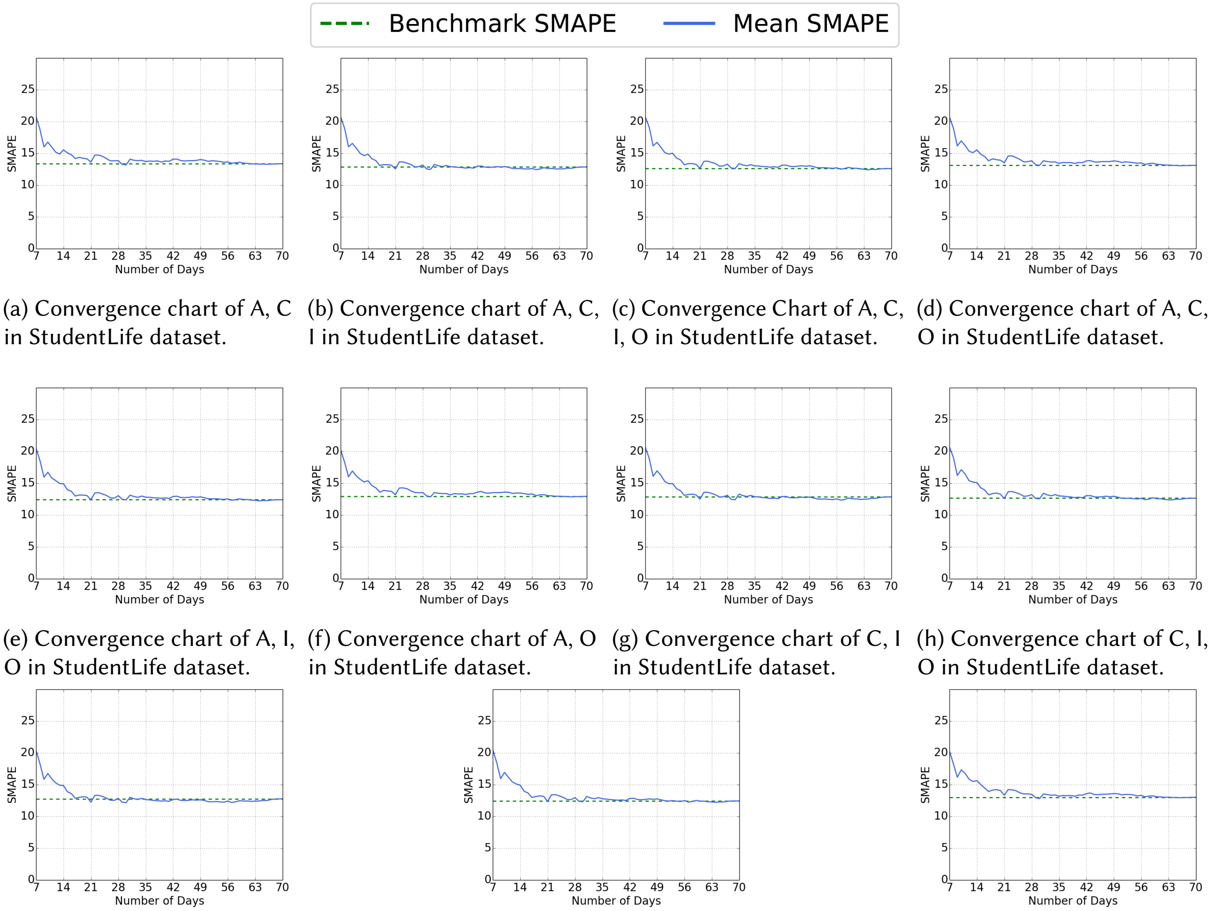
[69] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell. 2015. SmartGPA: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 295–306.

[70] WellTrack. [n.d.]. WellTrack Self-Help Tools. https://caps.ucsc.edu/resources/welltrack.html

[71] Drew Westen, Serra Muderrisoglu, Christopher Fowler, Jonathan Shedler, and Danny Koren. 1997. Affect regulation and affective experience: individual differences, group differences, and measurement using a Q-sort procedure. *Journal of Consulting and Clinical Psychology* 65, 3 (1997), 429.

[72] Shirley Yen, M Tracie Shea, Charles A Sanislow, Carlos M Grilo, Andrew E Skodol, John G Gunderson, Thomas H McGlashan, Mary C Zanarini, and Leslie C Morey. 2004. Borderline personality disorder criteria associated with prospectively observed suicidal behavior. *American Journal of Psychiatry* 161, 7 (2004), 1296–1298.

A APPENDIX



(a) Convergence chart of A, C in StudentLife dataset.

(b) Convergence chart of A, C, I in StudentLife dataset.

(c) Convergence Chart of A, C, I, O in StudentLife dataset.

(d) Convergence chart of A, C, O in StudentLife dataset.

(e) Convergence chart of A, I, O in StudentLife dataset.

(f) Convergence chart of A, O in StudentLife dataset.

(g) Convergence chart of C, I in StudentLife dataset.

(h) Convergence chart of C, I, O in StudentLife dataset.

(i) Convergence chart of I in StudentLife dataset.

(j) Convergence chart of I, O in StudentLife dataset.

(k) Convergence chart of O in StudentLife dataset.

Fig. 6. Convergence characteristic of our regressor using various combinations of sensing modalities. In the figures above, C stands for Conversation, A stands for Activity, I stands for Indoor Mobility, and O stands for Outdoor Mobility