Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing

RUI WANG, Dartmouth College, USA
WEICHEN WANG, Dartmouth College, USA
ALEX DASILVA, Dartmouth College, USA
JEREMY F. HUCKINS, Dartmouth College, USA
WILLIAM M. KELLEY, Dartmouth College, USA
TODD F. HEATHERTON, Dartmouth College, USA
ANDREW T. CAMPBELL, Dartmouth College, USA

43

There are rising rates of depression on college campuses. Mental health services on our campuses are working at full stretch. In response researchers have proposed using mobile sensing for continuous mental health assessment. Existing work on understanding the relationship between mobile sensing and depression, however, focuses on generic behavioral features that do not map to major depressive disorder symptoms defined in the standard mental disorders diagnostic manual (DSM-5). We propose a new approach to predicting depression using passive sensing data from students' smartphones and wearables. We propose a set of *symptom features* that proxy the DSM-5 defined depression symptoms specifically designed for college students. We present results from a study of 83 undergraduate students at Dartmouth College across two 9-week terms during the winter and spring terms in 2016. We identify a number of important new associations between symptom features and student self reported PHQ-8 and PHQ-4 depression scores. The study captures depression dynamics of the students at the beginning and end of term using a pre-post PHQ-8 and week by week changes using a weekly administered PHQ-4. Importantly, we show that symptom features derived from phone and wearable sensors can predict whether or not a student is depressed on a week by week basis with 81.5% recall and 69.1% precision.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Applied computing → Life and medical sciences**;

Additional Key Words and Phrases: Mobile Sensing, Mental Health, Depression

ACM Reference Format:
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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 43 (March 2018), 26 pages. https://doi.org/10.1145/3191775

Authors' addresses: Rui Wang, Dartmouth College, Computer Science, Hanover, NH, 03755, USA, ruiwang@cs.dartmouth.edu; Weichen Wang, Dartmouth College, Computer Science, Hanover, NH, 03755, USA, we-wang@cs.dartmouth.edu; Alex daSilva, Dartmouth College, Department of Psychological and Brain Sciences, Hanover, NH, 03755, USA, awdasilva21@gmail.com; Jeremy F. Huckins, Dartmouth College, Department of Psychological and Brain Sciences, Hanover, NH, 03755, USA, jhuckins@gmail.com; William M. Kelley, Dartmouth College, Department of Psychological and Brain Sciences, Hanover, NH, 03755, USA, bill.kelley@dartmouth.edu; Todd F. Heatherton, Dartmouth College, Department of Psychological and Brain Sciences, Hanover, NH, 03755, USA, todd.f.heatherton@dartmouth.edu; Andrew T. Campbell, Dartmouth College, Computer Science, Hanover, NH, 03755, USA, campbell@cs.dartmouth.edu.

2474-9567/2018/3-ART43 $15.00
https://doi.org/10.1145/3191775

1 INTRODUCTION

Clinical depression or major depressive disorder (MDD) is one of the most common and debilitating health challenges of our time. In 2015, an estimated 6.7% of all U.S. adults had at least one major depressive episode in the past year [63]. Major depressive disorder accounts for a staggeringly high proportion of illness-related burden worldwide [55, 72], and is the second leading cause of years lost to disability in the U.S. [54]. College age young adults 18 to 25 are more likely to have major depressive episodes than any other age groups. In 2015, an estimated 10.3% of young adults had a major depressive episode over the past year with 6.5% reporting the episode resulted in severe impairment [63]. The college years introduce major stressors for young adults that may exacerbate students' propensity for psychopathology, including increased academic pressures, social challenges, unfamiliar living and physical environments, financial pressures, and cultural differences that affect self-worth [33, 37]. Furthermore, students must negotiate loss of familiar support systems and social networks (e.g., high school friends). Consequently, many young adults can feel overwhelmed, struggle to find their place, and become more susceptible to depression or other mood disorders. Surveys at colleges across the U.S. found that 53% of respondents experienced depression at some point after entering college with 9% reporting suicidal ideation [26]. A recent study of Facebook profiles showed that 25% of students displayed depressive symptoms [53]. In addition, a 2016 study by the American College Health Association found that 38.2% of students at 2- and 4-year institutions reported feeling "so depressed that it was difficult to function" in the past year [3]. At Dartmouth College, a 2016 survey shows that depression and anxiety are the most common health problems, exceeding national averages in the adult population; 19% of Dartmouth students report being diagnosed with depression, and 24% say that depressive symptoms had harmed their academic performance [19]. Importantly, up to 84% of college students who screen positively for depression never seek mental health services [22]. Many students become aware (i.e., insight to illness) of their depression only after experiencing significant functional deterioration [37]. Many colleges offer mental-health services and counseling, but the stigma associated with mental illness is a major barrier to care-seeking [12, 18]. Evidence also suggests that higher education institutions' current approaches are not addressing depression adequately [37]. Depression rates continue to increase [26]. There is a need to understand what is happening on our campuses with these increasing depression rates. One thing is clear: the demand for mental services on US campuses is increasing with many institutions not capable of dealing with the rising needs of students. Clinicians, mental health counsellors, teachers, and administrators on our campuses do not understand why this inflection toward higher rising risk has occurred.

Identifying early warning signs of depression (i.e., "red flags") could mitigate or prevent major depression disorder's negative consequences [13, 37]. However, if students do not pay attention to their clinical condition or do not seek care when needed, depression can lead to devastating outcomes, including self-injurious behavior and suicide [26, 27]. There is a growing realization in academia and industry that everyday mobile phones and wearables (e.g., fitbits, smartwatches), which passively collect and analyze behavioral sensor data 24/7, will complement traditional periodic depression screening methods (e.g., PHQ-9 survey [40, 41, 67]) and visits to mental health specialist – ultimately, if validated at scale mobile sensing has the potential to replace periodic screening questionnaires such as PHQ-9. Recently, researchers have made progress in understanding the relationship between behavioral sensor data from phones and mental health [6, 14, 45, 61, 62, 75]. In addition, there is considerable activity in the startup space in the area of mental health. A number of companies are starting to use mobile technologies to assess and help people living with depression [28, 71]. However, while progress is being made (e.g., significant correlations between mobile sensing data and depression have been found across different studies [61, 75]) to the best of our knowledge there is no mobile passive sensing technology capable of predicting rising risk, impending depressive episodes, or occurrence from a combination of smartphone and wearable passive sensing to date.

The standard mental disorders diagnostic manual (DSM-5 [5]) defines 9 common symptoms associated with major depression disorders: depressed mood, loss of interest or pleasure in activities, sleep changes, weight change, fatigue or loss of energy, restlessness or feeling slow, diminished interest or pleasure in activities, diminished ability to concentrate, feelings of worthlessness, and thoughts of death and suicide. Existing work [6, 14, 45, 61, 62, 75] have found relationships between depression and generic behavioral features from passive sensing. However, they do not discuss how these behavioral features are associated with the well-defined depression symptoms. In this paper, we take a different approach and propose for the first time a set of depression sensing symptom features (called *symptom features* for short) derived from phone and wearable passive sensor data that represent proxies for the DSM-5 depression symptoms in college students; that is, we design a set of behavioral features to capture the characteristics of the depression symptoms that take into account lifestyles of students (e.g., going to class, working in study areas, socializing on campus). Specifically, we hypothesize: (1) the *sleep change symptom* can be measured by sleep duration, start time, and end time inferred from passive sensor data from phones [15, 74, 75]; (2) the *diminished ability to concentrate symptom* can be associated with excessive smartphone use [20, 44], specifically when measured in study spaces across campus (e.g., libraries, study rooms, quiet working areas associated with cafes, dorms, etc.) where students typically focus on their course work; (3) the *loss of interest or pleasure in activities symptom* may cause changes in activity and social engagement patterns [5], thus can be associated with changes in activity, conversation, and mobility patterns inferred from mobile sensing data; and finally (4) the *depressed mood symptom* and *fatigue or loss of energy symptom* relate to changes in physiology, thus, may be associated with changes in heart rate data passively measured by wearables (e.g., prior work found heart rate data is associated with depressed mood [35, 38] and fatigue [65]). To test our hypothesis, we conduct a study of 83 undergraduate students at Dartmouth College across two 9-week terms during the winter and spring terms in 2016. Each student installed an updated StudentLife app [75] on their own Android or Apple phones and were given a Microsoft Band 2 [52] for the duration of a 9-week term. The StudentLife app continuously collects behavioral passive sensing data from smartphones and physiological sensing data from Microsoft Band 2 [52]. We compute the symptom features from the passive sensing data and conduct correlation analysis between the symptom features and PHQ-8/PHQ-4 depression outcomes. We look at the correlations between the symptom features and the PHQ-8 [43] and PHQ-4 [42] depression groundtruth. We further look to predict PHQ-4 depression subscale states (i.e., *non depressed* and *depressed*) [42] using the proposed symptom features.

The contributions of this paper as follows; specifically:

- We propose a set of passive sensor based symptom features derived from phones and wearables that we hypothesize proxy 5 out of the 9 major depressive disorder symptoms defined in DSM-5. To test this hypothesize, we conduct a study with 83 undergraduate students from Dartmouth College across two 9-week term during the winter and spring in 2016. PHQ-8 and PHQ-4 are self-reported at the beginning and the end of each term, and weekly across the term, respectively, and serve as groundtruth for the study.
- We find a number of correlations between the proposed symptom features and PHQ-8 item scores. The findings evaluate the efficacy of the symptom features to capture depression symptoms.
- We identify a number of correlations between the symptom features and PHQ-8; specifically, we find students who report higher PHQ-8 scores (i.e., are more depressed) are more likely to use their smartphone more particularly at study places ($r = 0.391, p < 0.001$) in comparison with all day phone usage ($r = 0.282, p = 0.010$); have irregular sleep schedules (i.e., more variations in bed time ($r = 0.301, p = 0.024$) and wake time ($r = 0.271, p = 0.043$); spend more time being stationary ($r = 0.374, p = 0.009$) and visit fewer places during the day ($r = -0.269, p = 0.023$).

- We use ANOVA to compare the means of the symptom features between the *non depressed group* and the *depressed group*, as defined in PHQ-8 [43]. We show that these two groups are clearly identified in our data set.
- We show that a lasso regularized linear regression model predicts pre PHQ-8 scores with MAE = 2.4 and predict post PHQ-8 scores with MAE = 3.60. The predicted PHQ-8 scores strongly correlate with the groundtruth.
- We identify a number of symptom features capturing depression dynamics during the term that are associated with the PHQ-4 depression subscale using regression analysis for longitudinal data; specifically, students who report higher PHQ-4 depression subscale score (i.e., are more depressed) are around fewer conversations ($p = 0.002$), sleep for shorter periods ($p = 0.024$), go to sleep later ($p = 0.001$), wake up later ($p = 0.027$), and visit fewer places ($p = 0.003$) during the last two week period.
- We use lasso regularized logistic regression [69] to predict whether or not a student is depressed on a week by week basis. The area under the ROC curve (AUC) [31] of the prediction model is 0.809. We show that the logistic model predicts students' depressed state with 81.5% recall and 69.1% precision.
- We present a case study of a student that captures the depression dynamics across the term as an anecdotal example of the insights that behavioral sensing data can offer into the life of a student. In this case the student shows elevated depression as the term progresses with behavioral curves indicating less sleep, less conversational interaction, and fewer places visited. There is an inflection point after which the student recovers and shows resilience without seemingly any known clinical intervention (e.g., they do not visit the student health center on campus).

To the best of our knowledge we are the first to propose a set of symptom features derived from mobile phones and wearables based on specific depression symptoms defined in DSM-5 and specifically designed to capture the dynamics of depression for college students. As in the case of the original StudentLife study [75] we plan to release the depression data set in due course for other researchers to study student health issues. The structure of the paper is as follows. We present related work on mental health sensing in Section 2, followed by a detailed description of the "symptom features" in Section 3. We describe the sensing system, study design, and the dataset in Section 4. We describe our methods in Section 5. Following that, we discuss the correlations between the symptom features and PHQ-8 in Section 6. In Section 7, we present the results of using the symptom features to regress and predict PHQ-4 depression subscale. We discuss implications of our methods and results in Section 8, and limitations in Section 9. Finally, we present some concluding remarks in Section 10.

2 RELATED WORK

There is an increasing body of influential work on mobile sensing for mental health [6, 10, 11, 14, 45, 61, 62, 75]. In what follows, we focus on work related to depression sensing in addition to work on other mood disorders that has inspired us.

In [59] the authors use an early mobile sensing platform device [16] equipped with multiple embedded sensors to track a 8 older adults living in a continuing care retirement community. The authors demonstrate that speech and conversation occurrences extracted from audio data and physical activity infer mental and social well-being. The StudentLife study [75] investigates the relationship between passive sensing behaviors from mobile phones including conversation, sleep, activity and co-location with mental health outcomes, such as, depression, stress, loneliness, and flourishing for 48 college students over a 10-week term. The authors found for the first time significant correlations between passive sensing data from phones and the PHQ-9 depression scale [40, 41, 67]. Researchers at Northwestern University [62] find that mobility and phone usage features extracted from mobile phone data correlates with depressive symptom severity measured by PHQ-9[40, 41, 67] for 40 participants recruited from the general community over a two week period. Results show features from GPS data, including circadian movement, normalized entropy, location variance, and phone usage features, including usage duration and usage frequency, are associated with depressive symptom severity. It is important to note that the research team were able to reproduce the findings from their initial study [61] using the StudentLife dataset [75]. The replication of the Northwestern University researchers' study results [62] using a different dataset indicates that the mobility features could be broadly applicable in depression sensing across different communities; that is, students on a small college campus and people recruited in a metropolitan area.

There seems to be growing evidence that mobility features have significant signal when it comes to depression sensing. Canzian et al [14] develop an expanded set of mobility features over [61, 62] and found that location data correlates with PHQ-9 [40, 41, 67]. The authors show that the maximum distance traveled between two places strongly correlates with the PHQ score. Mobility features are used to track the trajectory of depression over time. Recently, the same team [51] report their preliminary results from a 30 day 25 participant study on the association between human-phone interaction features (e.g., interactions with notifications, number of applications launched) and PHQ-8 scores. Demirci et al.[20] conduct an study with 319 university students to investigate the relationship between smartphone usage and sleep quality, depression, and anxiety. They divide the participants into a smartphone non-user group, a low smartphone use group, and a high smartphone use group based on the Smartphone Addiction Scale (SAS)[44]. The authors find the Beck Depression Inventory (BDI) [8, 9] score is higher in the high use group than the low use group. This study is solely based on self-report and does not have a sensing component but illustrates that interaction usage could be important in depression sensing.

In [24], the authors present a study of 79 college-age participant from October 2015 to May 2016 in which they find a number of mobility features (e.g., home stay duration, normalized entropy) correlate with PHQ-9. The authors report using SVM with RBF kernel [66], they can predict clinical depression diagnoses with the precision of 84%. Wahle et al. [73] recruit 126 adults to detect depression levels from daily behaviors inferred by phone sensing, in addition to exploring intervention. In the study 36 subjects with an adherence of at least 2 weeks are included in the analysis. The authors compute features based on activity, phone usage, and mobility. They report 61.5% accuracy in predicting a binary depression state. In [2], the authors investigate difference in speech styles, eye activity, and head poses between 30 depressed subjects and 30 non-depressed subjects. The authors report 84% average accuracy in predicting the depression state using a combination of the speech style, eye activity, and head pose features. In [58] Place et al., report on a 12 week study with 73 participants who report at least one symptom of post-traumatic stress disorder (PTSD) [77] or depression. The study assess symptoms of depression and PTSD using features extracted from passive sensing, including the sum of outgoing calls, count of unique numbers texted, absolute distance traveled, dynamic variation of the voice, speaking rate, and voice quality. They report area under the ROC curve (AUC) for depressed mood is 0.74. Chow et al.[17] hypothesizes that time spent at home is associated with depression, social anxiety, state affect, and social isolation. The authors use passive sensing from phones to compute a participant's time spent at home during a day. The study recruits 72 undergraduates and finds participants with higher depression tended to spend more time at home between the hours of 10 am and 6 pm. The DeepMood project [68] develops a recurrent neural network algorithm to predict depression. The authors conduct a study with 2382 self-declared depressed participants using self-reports to collect self-reported mood, behavioral log, and sleeping log. They report their long short-term memory recurrent neural networks (LSTM-RNNs) [34] model predict depression state with AUC-ROC 0.886.

Detecting bipolar disorder is to some degree related to work on depression sensing. The MONARCA project [32, 56, 57, 60] first reported on findings from mobile sensing and bipolar disorder. The authors [57] discuss correlations between the activity levels over different periods of the day and psychiatric evaluation scores associated with the mania-depression spectrum. The findings reported in [1] show the automatic inference of circadian stability as a measure to support effective bipolar management. Maxuni et al. [49] extend these insights by using speech and activity levels to successfully classify stratified levels of bipolar disorder.

In summary, the existing work use either self-report or behavioral features extracted from mobile sensing to predict depression states. The proposed features, however, are not based on specific depression symptoms. We are the first to propose passive sensing features that are proxies to depression symptoms defined in DSM-5 [5] and specifically designed to capture the characteristics of students' college life.

3 DEPRESSION SENSING USING SYMPTOM FEATURES

How we assess depression has not changed in 30 years. Mental health specialists rely on depression screening tools (e.g., Patient Health Questionnaire (PHQ) [40, 41, 67], Major Depression Inventory (MDI) [7], Beck Depression Inventory (BDI) [8], Hamilton Depression Rating Scale (HDRS) [30]) and clinical interviews to assess the severity of the symptoms of major depressive disorder as defined in the DSM-5 [5]. These questionnaires rate the frequency of the symptoms that factor into scoring a severity index called a "depression score". These tools, however, rely on periodic subjective self-reports. A person is said to suffer from a depressive episode if they experiencing at least 5 of the 9 depression symptoms during the same 2-week period, including either depressed mood or loss of interest or pleasure in activities, the symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning [5], and the symptoms are not attributable to other medical conditions [5]. Existing work on predicting depression using mobile phone sensing focuses on using generic behavioral features to predict or correlate the depression scale. *We take a different approach and hypothesize that mobile sensor data from phones and wearables represent proxies for the DSM-5 depression symptoms in college students; that is, we design a set of behavioral features to capture the characteristics of the depression symptoms that take into account lifestyles of students (e.g., going to class, working in study areas, socializing on campus).* In what follows, we describe the depression symptom features that represent 5 of the 9 major depressive disorder symptoms. The symptom feature implementation details are described in Section 5.1.

Sleep changes may result in students experiencing difficulty sleeping (insomnia) or sleeping too much (hypersomnia) and changes in sleep schedules. Many times because of the demands of the term (e.g., assignment due dates, exams, social life, sports, etc.) students experience changes in the regular sleep patterns as demands on the term increase. We infer students' sleep time, wake up time, and sleep duration using passive sensing from phones [15]. We use phone sensing to determine if depressed students sleep more or less than non-depressed students and if depressed students have more irregular sleep schedules; that is, more variation in the time they go to bed and wake up. As a result we can accurately infer the sleep changes symptom.

Diminished ability to concentrate may cause students to appear distracted, unfocused, and unable to perform well. We use phone usage to measure if a student is more likely to be distracted. Previous work [20] shows that smartphone overuse may lead to increased risk of depression and/or anxiety. Specifically, we measure the number of unlocks of the phone and associated usage duration across the day and at specific locations; for example, dorm, in study areas and in the classroom, etc. When a student is at the classroom or study areas, they are supposed to focus on work at hand or studying. In such locations we assume the more phone usage (i.e., spend more time on their phones) may indicate they are having difficulty focusing on their work. We hypothesize that phone usage in the classroom and study places is a potential indicator of a student's diminished ability to concentrate in comparison to regular phone use in social spaces, dorm, gym, or walking around campus. By using phone location data and contextual labeling of the campus [76] we can differentiate these different use cases accurately.

Diminished interest or pleasure in activities may cause changes in students' activity and social engagement patterns that are not easily explained by external forces like deadlines or exams. We can observe changes in a student's physical activity (i.e., more/less active) and mobility (e.g., visit more/fewer places on campus). Specifically, we look at the time spent at different types of places. On campus buildings and spaces in buildings are usually associated with a primary function (e.g., study area, classroom, library, gym, social, cafes, etc). Most Dartmouth undergraduate students study in a large number of shared study spaces across campus including libraries and typically dormitory buildings are used to rest, sleep and socialize and rarely used to study. We compute time spent in all areas but specifically in study places and dorms; specifically, we compute a number of behavioral features including dwell time at locations, phone usage (unlock frequency and duration), and the number and duration of conversation students are around in these spaces. All the location based behavioral features are normalized by the dwelling duration. We use the fusion of these features to proxy a student's diminished interest or pleasure in activities.

Depressed mood and **Fatigue or loss of energy** are difficult to detect from passive behavioral sensing from phones. Instead we consider heart rate from wearables. Previous work has found that heart rate data is associated with depressed mood [35, 38] and fatigue [65]. We determine if depressed students' heart rate is different from non-depressed students. We consider this might be a potential signal and proxy for depressed mood or fatigue or loss of energy. In addition, we also determine if more depressed students visit on-campus health facilities more.

4 DATA COLLECTION

We collect a smartphone and wearable sensing dataset from 83 Dartmouth College undergraduate students across two 9-week terms during winter (56 students) and spring (27 students) terms in 2016. The average age of the participants is 20.13 (std=2.31) and 40 are male and 43 are female (26 are Asian, 5 are African American, 24 are Caucasian, 1 is multiracial, and 26 not specified). This study is approved by the Institutional Review Board at Dartmouth College. Figure 1 shows the sensing system, symptom feature mappings, and the depression ground truth. In what follows, we discuss the sensing system, the study design, and the dataset.

4.1 Mobile Sensing System for Phones and Wearables

The StudentLife sensing app is built based on our prior sensing work [1, 46, 74, 75] for Android. For this study we implemented the StudentLife app on iOS phones as well. The StudentLife app continuously infers and record participants' physical activities (e.g., stationary, in a vehicle, walking, running, cycling), sleep (duration, bed time, and rise time) based on our prior work on sleep inference using phones [15], and sociability (i.e., the number of independent conversations a participant is around and their duration). The app also collects audio amplitude, location coordinates, and phone lock/unlock events. A built-in MobileEMA component [75] is used to administer self-reported PHQ-4 [42] EMAs. The app uploads the data to the secured server when the user is charging their phones and under WiFi. StudentLife is extended to collect data from wearables; specifically, we collect physiological data from Microsoft Band 2 [52] given to each of the students in our study. The StudentLife app collects the heart rate, galvanic skin response (GSR), skin temperature, and activity data from the band in real time. Band data is uploaded to the StudentLife app over Bluetooth and then uploaded to our servers as described above. Note, during data modeling and analysis we found poor data quality issues associated with GSR and skin temperature data from the band. First, the GSR sample rate provided by the Microsoft Band SDK [52] is too low (0.2 HZ) to be useful in analysis. Such a low sample rate limited extracting useful GSR features. We also found that the skin temperature sensor reading is affected by the ambient environment temperature. The temperature differences between indoor and outdoor during New Hampshire winter can be as large as 70 degree Fahrenheit. We observed significant drops in skin temperature when participants are outside during the winter term. For these reasons we collected but did not uses GSR and skin temperature data in our modeling. While the StudentLife app infers sleep data from the phone only within +/- 25 mins of error [15] the band has much better sleep measurements. However, because the band only lasted one day due to limited battery and the demands of continuous sensing most students wore the band during the day and recharged it at night. The result is that we have limited sleep data from the band. We therefore only use sleep measurements from our phone data. Even though we collect GSR, sleep, skin temperature we end up only using heart rate and activity data from the band.

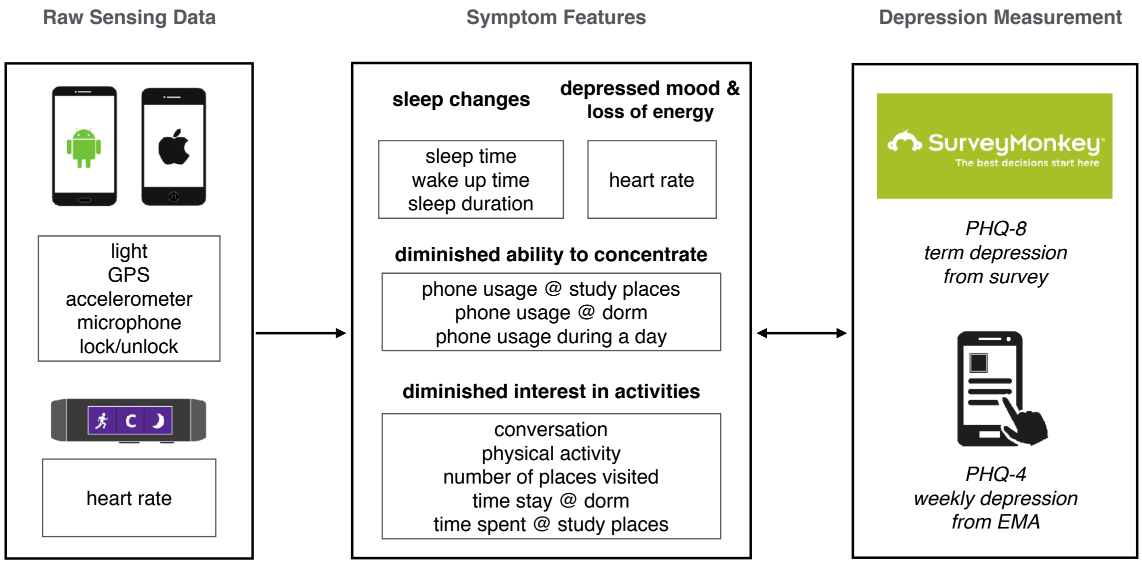
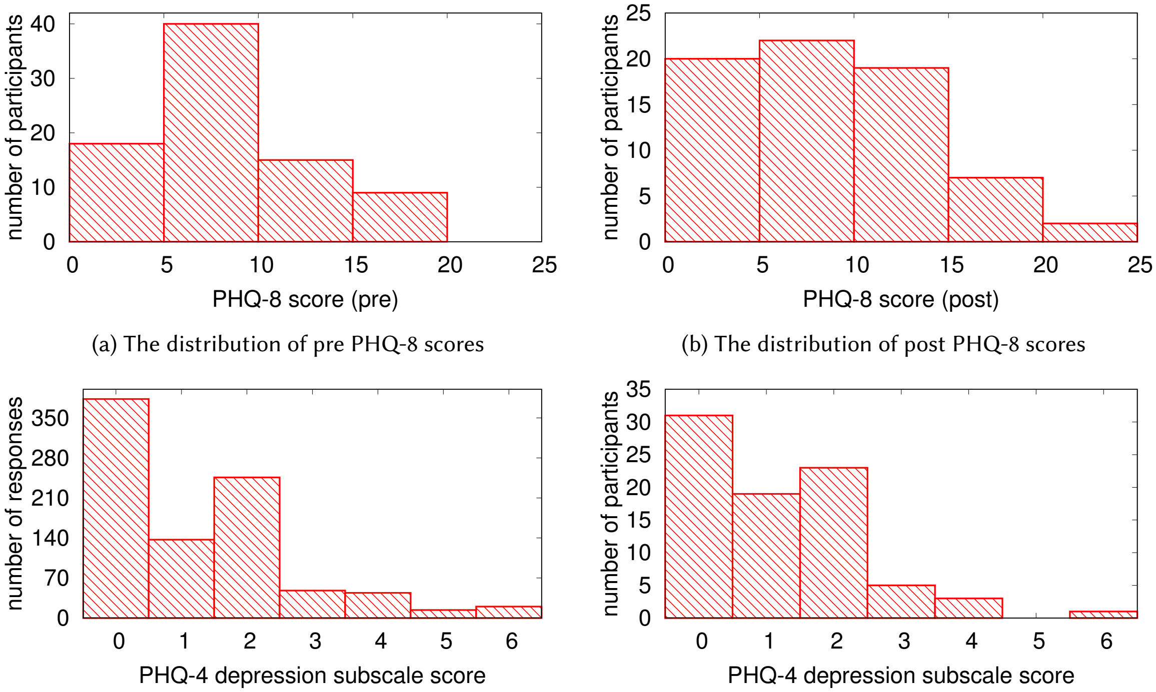


Fig. 1. We continuously collect behavioral passive sensing data from Android and Apple iOS smartphones and physiological sensing data from Microsoft Band 2. We compute the symptom features from the passive sensing data. The symptom features map smartphone and wearable passive sensing to 5 depression symptoms defined in DSM-5: sleep changes, diminished ability to concentrate, diminished interest in activities, depressed mood, and fatigue or loss of energy. We look for associations between the symptom features and the PHQ8/PHQ4 depression outcomes.

4.2 Depression Groundtruth

We use the self-reported PHQ-8 [43] and PHQ-4 [42] as groundtruth for depression outcomes in our study. This is a widely used measure with excellent validity. PHQ-8 is administered at the beginning and the end of the study period as a pre-post depression measures. PHQ-4 is administered once a week and used to capture depression dynamics across the term. The PHQ-8 scores 8 of the 9 major depressive disorder symptoms over the past two weeks where each item (i.e., question) is scored by the user from 0 (not at all) to 3 (nearly every day). The PHQ-8 does not score the *thoughts of death and suicide symptom* nor do we consider this in our study. The resulting PHQ-8 depression score ranges from 0 to 24 indicating five levels of depression: (1) none to minimal depression (range 0-4); (2) mild depression (range 5-9); (3) moderate depression (range 10-14); (4) moderately severe depression (range 15-19); and finally (5) severe depression (20 to 24). The score can also be interpreted as *no depression (range 0-9)* and *current depression (range 10-24)* according to [43]. Figure 2(a-b) shows the distribution of the pre-post PHQ-8 responses. The mean score for the pre PHQ-8 is 6.09 (std = 4.33), where 16 out of 82 students are in the *depressed group* (PHQ-8 $\geq$ 10). The mean score for the post PHQ-8 is 6.69 (std = 5.46), where 17 out of 71 students are in the depressed group. We receive fewer post PHQ-8 surveys because some participants do not complete the survey.

The PHQ-4 is an ultra-brief tool for screening both anxiety and depression disorders over the past two weeks. It uses a depression subscale to score depression and an anxiety subscale to score anxiety. We only consider



(a) The distribution of pre PHQ-8 scores

(b) The distribution of post PHQ-8 scores

(c) The distribution of PHQ-4 depression subscale scores

(d) The distribution of mean PHQ-4 depression subscale scores

Fig. 2. The distribution of the PHQ-8 and PHQ-4 responses. (a) The mean score for the pre PHQ-8 is 6.09 ($N = 82$, std = 4.33), where 16 students are in the depressed group (PHQ-8 $\geq$ 10). (b) The mean score for the post PHQ-8 is 6.69 ($N = 71$, std = 5.46), where 17 students are in the depressed group. (c) The mean score of the PHQ-4 depression subscale is 1.34 ($N = 707$, std = 1.50), where 108 responses are above the depressed cutoff (≥ 3). (d) The mean per-participant PHQ-4 depression subscale score is 1.31 (std = 1.17), where 4 participants' mean PHQ-4 depression subscale score is above the depressed cutoff (≥ 3).

the depression subscale in our study on depression. The depression subscale comprises 2 questions from the PHQ-4 that score the *diminished interest or pleasure in activities symptom* and the *depressed mood symptom*. The score of the depression subscale ranges from 0-6, where a score of 3 or greater is considered depressed according to [42]. We collected in total 707 PHQ-4 responses across the terms. Figure 2(c) shows the distribution of the PHQ-4 depression subscale scores from all responses. The mean score of the PHQ-4 depression subscale is 1.34 (std = 1.50), where 108 responses are above the depressed cutoff (≥ 3). Figure 2(d) shows the distribution of each student's PHQ-4 depression subscale. The mean per-student PHQ-4 depression subscale score is 1.27 (std = 1.15), where 5 students' mean PHQ-4 depression subscale score is above the depressed cutoff (≥ 3).

5 METHODS

In what follows, we present our symptom features and methods to evaluate the efficacy of using the symptom features to predict depression severity in college students.

5.1 Symptom Features

We present our symptom features listed in Table 1 that capture the 5 out of 9 major depressive symptoms based on data collected using phone and wearable passive mobile sensing data.

Table 1. Depression symptom features.

DSM symptom	Symptom features
sleep changes	sleep duration sleep start sleep end
diminished ability to concentrate	unlock duration unlock duration at dorm unlock duration at study places
diminished interest or pleasure in activities	stationary time conversation duration number of places visited time at dorm time at study places
depressed mood and fatigue or loss of energy	heart rate

Sleep. We compute three sleep features to measure the *sleep change symptom*: sleep duration, sleep onset time, and wake time. We assume students experiencing this symptom may sleep significantly more or less than normal, or experience irregular sleep schedules (i.e., more variations in sleep onset time or wake time). The sleep inferences are based on four phone sensors: ambient light, audio amplitude, activity, and screen on/off [15]. The sleep classifier does not infer naps. It simply computes the longest period of inferred sleep. The sleep classifier approximates sleep duration within +/- 30 minutes and has been used in a number of other studies [1, 74, 75].

Physical activity. Students who experience the *diminished interest or pleasure in activities symptom* may change their activity pattern (e.g., being less mobile and more stationary) [5]. We compute the stationary duration during a day to measure students' sedentary levels. The app continuously infers physical activities using the Android activity recognition API [29, 74] and iOS Core Motion [4]. Activity recognition infers whether a user is on foot, stationary, in vehicle, on bicycle, tilting, or doing an unknown activity. We compute the non-physically active duration (i.e., the stationary duration) using the still label from the classifier. Both Android and iOS activity recognition API detects the stationary state with high accuracy.

Speech and conversational interaction. Students who experience the *diminished interest or pleasure in activities symptom* may experience social withdrawal and change their social engagement patterns [5]. We compute the number of independent conversations and their duration everyday as a proxy for social interaction. The StudentLife app infers the amount of speech and conversation a participant is around [74, 75]. In [74, 75] we discuss the detailed in design of the conversation classifier that continuously runs on the phone in an energy efficient manner (i.e., duty cycled). In brief, it represents a two level classifier. At the lowest level we detect speech segments and at the higher level we determine if the set of segments represent a unique conversation. The conversation classifier does not identify speakers. Therefore, we do not know that if the participant is actively involved in the conservation or not (e.g., they could be sitting at a table in a cafe where others around them are speaking). However, we have validated this in a number of studies and it is capable of capturing levels of social interaction.

Location and mobility. There is evidence that people's mobility patterns are related to depression. We use mobility and location features as a proxy for the *diminished interest or pleasure in activities symptom*. Prior work indicates [14, 24] that people with this symptom avoid leaving their homes. We compute students' distance traveled, the number of places visited and time spend at dorms and study areas across campus based on location data [14, 62, 75] and a semantic understanding of locations across Dartmouth campus. We use DBSCAN [23] to cluster GPS coordinates collected during the day to find significant places that students dwell at. The DBSCAN algorithm groups location/GPS coordinates that are close together as a significant place where students visit. We compute the number of places visited as a feature. We label every on-campus building and spaces in buildings (e.g., classrooms, study area, dorms, libraries, cafes, social spaces, gyms, etc.) according to their primary function; for example, we label each student's dorm as the place where they dwell between 2-6 am. In addition, we also determine the number of times a student visits the on-campus health center as a contextual mobility feature.

Phone usage. We use phone usage to measure the *diminished ability to concentrate symptom*. Phone overuse has been linked to depression in college students [20, 44]. We compute the number of phone lock/unlock events and the duration that the phone is unlocked during a day, when a student is at their dorms and study areas. To avoid the impact of stay duration on the location based phone usage features (e.g., a student tends to record higher phone usage when they stay at a place longer), we normalize the phone usage features for location based usage data by duration of their stay. Excessive smartphone usage at study places or in the classroom may indicate students are experiencing difficulty in concentrating on the work at hand.

Heart rate. We use students' physiological signals to detect if they are experiencing *depressed mood* or *fatigue or loss of energy symptoms*. Previous work has found that heart rate variability is associated with depressed mood [35, 38] and fatigue [65]. The average of beat-to-beat or NN intervals (AVNN) is one of the heart rate variability measures [48]. The heart rate (HR) is the inverse of the AVNN in milliseconds: HR = 60000/AVNN. The StudentLife app collects heart rate data from Microsoft Band 2 in real time. The accuracy of wrist heart rate monitors depends on many factors. The heart rate measured by Microsoft Band 2 is accurate when the user wears the band correctly (i.e., not too loose nor too tight) and is relatively stationary during measurement periods. Based on our testing we found that the heart rate error is within 2 beats for the band. However, the accuracy suffers if moving their arms because of motion artifacts in the signal. In order to get an accurate measure of daily heart rate, we compute the median heart rate during each day. The median heart rate is a more robust measure of daily heart rate than the mean, maximum, and minimum heart rate because median heart rate is less likely to be influenced by outliers.

5.2 Feature Set Construction

We construct a PHQ-8 dataset and a PHQ-4 dataset to look at association between the symptom features and the depression groundtruth.

The PHQ-8 dataset uses students' pre-post PHQ-8 responses as the ground truth. We look to find correlations between the PHQ-8 scores and the symptom features. In addition to the PHQ-8 scores, we include the PHQ-8 depression group assignment (i.e., *non depressed group (range 0-9)* and *depressed group (range 10-24)* as defined in [43]) as students' binary pre-post depression state. We compute the term mean, standard deviation, and the slope of each symptom features described in Section 5.1. The term mean features describes the average of the daily symptom features over the 9-week term. For example, the mean time spend at study places is the average time a student spends at study places every day during the term. The term standard deviation describes the variations in the daily symptom features. For example, a higher standard deviation in sleep start time and end time indicates that the student has an irregular sleep schedule. The term slope features describe how the daily symptom features change over the term. We fit the daily feature time series with a linear regression model and use the regression coefficient as the slope. The slope describes the direction and the steepness of change. For example, a positive slope in conversation duration indicates the student is around more and more conversations as the term progresses, whereas a negative slope indicates the surrounding conversations decrease over time. The absolute value of the slope shows how fast the conversation duration changes. In addition to the symptom features, we also include the mean PHQ-4 score from each student in the dataset. The correlations between the mean PHQ-4 and PHQ-8 show the validity of the EMA administered PHQ-4.

The PHQ-4 dataset uses students' weekly self-administered PHQ-4 depression subscale scores as the groundtruth. We include the PHQ-4 depression group assignment (i.e., non depressed group (range 0-2) and depressed group (range 3-6) as defined in [42] as students' binary weekly depression state. For each PHQ-4 response, we compute the mean symptom features using data from the past 2 weeks. This is because PHQ-4 asks for participants' symptoms during the last two weeks.

5.3 PHQ-8 Association and Prediction Analysis

We test our hypothesis that the mobile sensing derived depression features represent proxies for the depression symptoms for college students by first running Pearson correlation analysis to assess the relations between the term symptoms features and the pre-post PHQ-8 item scores. Each of the PHQ-8 items maps to one of the major depression disorder symptoms defined in DSM 5 except the suicidal ideation symptom. The correlations between the PHQ-8 item scores and symptom features may give us preliminary insight about whether or not the symptom features are likely to be associated with individual symptoms.

We then run Pearson correlation analysis to assess the relations between the term symptoms features and the pre-post PHQ-8 scores. PHQ-8 scores are validated depression severity measures. The correlations suggest how the symptom features are related to the overall depression severity. We report the correlation coefficients and the p values. We also apply the False Discovery Rate (FDR) [78] to address the multiple testing problem [21, 25].

In addition to the correlation analysis, we use ANOVA [64] to test whether or not the mean of the symptom features are significantly different between the non depressed group and the depressed group. The PHQ-8 defines a non depressed group (< 10) and a depressed group (≥ 10) [43]. We use ANOVA [64] to test whether or not the mean of the symptom features are significantly different between the non depressed group and the depressed group. Analysis of variance (ANOVA) is a statistical model that is widely used to analyze the differences among group means [64]. We report the F statistics and the p value from ANOVA. The F statistics indicate the ratio of between-group variability and the within-group variability. The p value indicates whether or not the group means are significantly different.

Finally, we use lasso regularized linear regression [69] to predict pre-post PHQ-8 scores. The lasso regularization selects features that are predictive of the outcome by penalizing irrelevant features' weights to zeros [69]. Before training the model, we normalize each feature to have zeros mean and one standard deviation. Feature normalization avoids the different feature scales adversely affecting regularization. We use 10-fold cross-validation to select the regularization hyperparameter, which controls the penalizing strength of non-zero-weight features. We choose the hyperparameter that minimizes the mean squared error (MSE). We report the mean absolute error and correlations between the predicted PHQ-8 scores and the groundtruth. We use the population PHQ-8 mean as the prediction baseline to compare the prediction performance.

5.4 PHQ-4 Regression and Prediction Analysis

We further test our hypothesis by identifying a number of associations between 2-week symptom features and the PHQ-4 depression subscale scores using regression analysis. We use the 2-week symptom features to predict if a student is considered depressed according the PHQ-4 depression subscale. The periodic PHQ-4 responses are longitudinal across the term where every student provides multiple responses. Ordinary linear regression and correlation cannot be applied to analyze longitudinal data because the responses from the same individual are likely correlated. We run bivariate regression analysis using the generalized linear mixed model (GLMM) [50] to understand associations between the 2-week symptom features and the PHQ-4 depression subscale scores. GLMM is a widely used model to analyze longitudinal data. It describes the relationship between two variables using coefficients that can vary with respect to one or more grouping variables. In our case, the group variable is student. GLMM better explains the intra-individual differences. We normalize each symptom feature to have zero mean and one standard deviation. The regression coefficient of a normalized feature b can be interpreted as a unit increase in the feature value is associated with b increases in the associated PHQ-4 depression subscale value. A positive coefficient indicates that a greater feature value is associated with a greater depression score, whereas a negative coefficient indicates that a greater feature value is associated with a smaller depression score. PHQ-4 are collected at a weekly rate. The 2-week symptom features may artificially create dependency among consecutive PHQ-4 scores. To address the dependency problem, we first run GLMM on all PHQ-4 data, then we remove consecutive PHQ-4 responses by skipping a week's PHQ-4 response. The skip-a-week dataset is half of the PHQ-4 dataset in size. We compare the regression results from the complete PHQ-4 dataset and the skip-a-week dataset. Similar results from the two datasets would suggest dependency does not have an impact on the analysis.

We then use lasso regularized logistic regression [69] to predict whether or not a student is depressed week by week (i.e., the reported PHQ-4 depression subscale is ≥ 3). We use the PHQ-4 data to train a logistic regression model to predict each student's PHQ-4 for a given week. We first use 10-fold cross-validation to select the regularization hyperparameter, which controls the penalizing strength of non-zero-weight features. We then choose the hyperparameter that maximizes the regression deviance to train a generic PHQ-4 prediction model. We report the prediction performance from the 10-fold cross-validation.

6 PHQ-8 RESULTS: ASSESSING DEPRESSION ACROSS THE TERM

In what follows, we first report the correlations between the symptom features and PHQ-8 item scores to show whether or not the symptom features are likely to be associated with individual symptoms. We then report the correlations between the symptom features and PHQ-8 scores. Finally, we report ANOVA results to show whether or not the mean of the symptom features are significantly different between the non depressed group and the depressed group.

6.1 Correlations Between Symptom Features and PHQ-8 Item Scores

In what follows, we discuss the relationship between the specific PHQ-8 item scores and the proposed symptom features. Figure 3 shows the correlation matrices of the term mean symptom features and eight pre-post PHQ-8 item scores. We omit correlations with $p > 0.05$. In what follows, we discuss our findings.

Higher sleep changes (more irregular sleep patterns) item score is associated with shorter sleep duration (in line with our hypothesis), longer unlock duration during the day at dorm and study places, more time being stationary, and spending more time at on-campus health facilities. The sleep start time and end time, however, are not correlated with the sleep changes item score.

Higher mood item score is associated with longer unlock duration at dorm and study places, more time being stationary, and spending more time at on-campus health facilities. The heart rate is not correlated with the mood item score.

Higher loss of energy (fatigue) item score is associated with spending more time at dorm. The heart rate is not a predictor of this PHQ-8 item.

Higher diminished ability to concentrate item score is associated with longer unlock duration at study places (in line with our hypothesis) and more time being stationary.

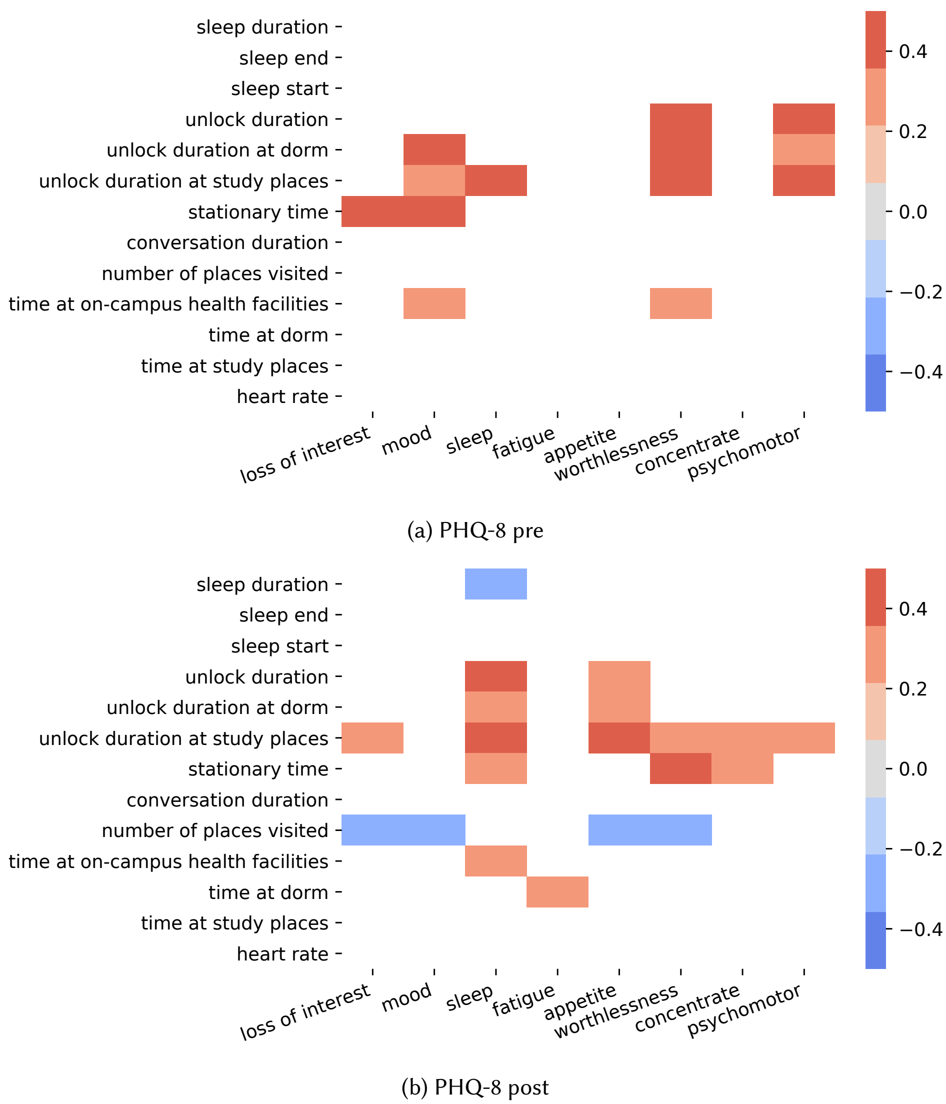


Fig. 3. The correlation matrix of proposed symptom features and PHQ-8 pre-post item scores. Correlations with $p > 0.05$ are omitted.

Higher diminished interest in activities item score is associated with longer unlock duration at study places, more time being stationary (in line with our hypothesis), and visiting fewer places a day (in line with our hypothesis).

Higher feeling worthless item score is associated with longer unlock duration during the day at dorm and study places, spending more time at on-campus health facilities, more time being stationary, and visiting fewer places a day.

Higher psychomotor retardation/agitation item score is associated with longer unlock duration during the day, at dorm, and at study places.

Higher appetite changes item score is associated with visiting fewer places a day, longer unlock duration during the day, at dorm, and at study places.

We find more statistically significant correlations between the symptom features and the post PHQ-8 item scores. We believe this is because post PHQ-8 scores better capture students depression states during the term whereas pre PHQ-8 scores capture the depression states when students started the academic term. The symptom features computed from the data collected during the term better capture students' depression symptoms during the term. The results show there are indeed association between the proposed symptom features and symptoms scores. We also find that some symptom features correlate with symptoms that are considered relevant. For example, phone use data (e.g., unlock duration) is associated with sleep changes.

6.2 Correlations Between Symptom Features and PHQ-8 Depression Scores

Table 2. Pearson correlations between the term symptom features and pre-post PHQ-8 scores

		PHQ-8 pre		PHQ-8 post	
		r	p	r	p
sleep duration	all		> 0.05		> 0.05
sleep start	std	0.236	0.059	0.301	0.024
sleep end	std	0.183	0.145	0.271	0.043
unlock duration	mean	0.282	0.010*	0.268	0.024
unlock duration at dorm	mean	0.245	0.027	0.206	0.085
unlock duration at dorm	std	0.270	0.014*	0.222	0.062
unlock duration at study places	mean	0.391	< 0.001**	0.322	0.006*
unlock duration at study places	std	0.260	0.018	0.120	0.319
stationary time	mean	0.256	0.040	0.347	0.009*
conversation duration	slope	0.467	< 0.001**	0.223	0.062
number of places visited	mean	-0.066	0.556	-0.269	0.023
time at on-campus health facilities	mean	0.210	0.059	0.029	0.812
time at dorm	all		> 0.05		> 0.05
time at study places	all		> 0.05		> 0.05
heart rate	all		> 0.05		> 0.05
PHQ-4 depression subscale	mean	0.743	< 0.001**	0.849	< 0.001**
PHQ-4 depression subscale	std	0.328	0.003*	0.521	< 0.001**
PHQ-4 depression subscale	slope	0.045	0.688	0.438	< 0.001**

*FDR < 0.1, **FDR < 0.05

The correlation results are presented in Table 2. In what follows, we discuss the correlation results in detail. Sleep duration, sleep start time, and sleep end time are proxies to measure the sleep changes symptom. We do not find correlations between sleep duration and PHQ-8 score. However, we find students who have more variations in their bed time schedule report higher pre ($r = 0.236, p = 0.059$) and post ($r = 0.301, p = 0.024$) PHQ-8 scores. Students who have more variations in their wake up time report higher post PHQ-8 scores ($r = 0.271, p = 0.043$). The results show that students with more irregular sleep patterns tend to be more depressed.

Phone unlock duration during the day, at their dorms, and in study places are proxies to measure the diminished ability to concentrate symptom. We find all 3 unlock duration features correlate with the PHQ-8 scores. In general, students who use their phones more during the day report higher pre PHQ-8 score ($r = 0.282, p = 0.010$) and post PHQ-8 scores ($r = 0.268, p = 0.024$). When students are at their dorms, those who use their phone more report higher pre PHQ-8 score ($r = 0.245, p = 0.027$). We find strong correlations when students are at study places where the typically goal is to focus on school work. Students who spend more time using their phones in study areas report higher pre PHQ-8 scores ($r = 0.391, p < 0.001$) and higher post PHQ-8 scores ($r = 0.322, p = 0.006$). The results show that context aware device usage can be used to detect distractions and measure the ability of students to concentrate.

Stationary time during the day, conversation duration, number of places visited, and time spent at dorms and study places are proxies to measure the diminished interest or pleasure in activities symptom. Students who report higher post PHQ-8 scores are likely to spend more time being stationary ($r = 0.256, p = 0.040$ for the pre survey and $r = 0.374, p = 0.009$ for the post survey) and visit fewer places during the day ($r = -0.269, p = 0.023$ for the post survey). Students who report higher pre PHQ-8 scores see an increase in conversation duration (i.e., conversation duration slope) as the term progresses ($r = 0.467, p < 0.001$). However, there is no correlation between the mean conversation duration and the post PHQ-8 scores. We speculate this is because students who are depressed at the beginning of the term are not social at the beginning of the term. However, as the term progresses, students become more socially engaged. This may because some students in the cohort seek help at on campus health facilities.

The daily median heart rate is a proxy for the depressed mood symptom and the fatigue or loss of energy symptom. However, it does not correlate with the PHQ-8 score. Students who report higher pre PHQ-8 scores at the beginning of the term spend more time at the campus health center ($r = 0.210, p = 0.059$). However, the correlation does not hold for the post PHQ-8 scores. The results show that although students who are more depressed at the beginning of the term actively seek medical help, students who become depressed during the term may not seek out help at campus health center. When talking to the clinicians and mental health counselors at the Dartmouth campus health center they indicate that the peek demand time they see students is during midterm weeks, during once per term social festivals, and at the end of term (week before, during and after the final exam period). Clearly, there seems to be a barrier for depressed students to visit the campus health center. Many issues might stop a student reaching out including not understanding they are experiencing depression or stigma associated with mental illness. The result is consistent with what other colleges experience [22].

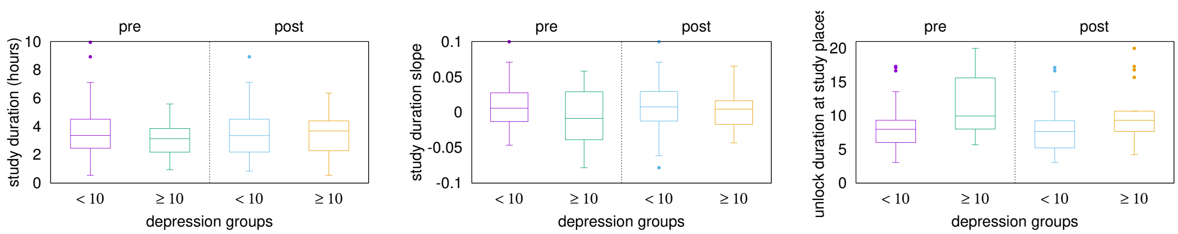
We administer weekly PHQ-4 EMAs to track how students' depression changes as the term progresses. We compute the PHQ-4 depression subscale term average for each student and compare the subscale scores with the PHQ-8 scores. The PHQ-4 depression subscale scores strongly correlate with both pre ($r = 0.743, p < 0.001$) and post ($r = 0.849, p < 0.001$) PHQ-8 scores. The result show that PHQ-4 responses are consistent with PHQ-8, which gives us confidence to use the PHQ-4 depression subscale to track depression changes during the term. The PHQ-4 depression subscale standard deviation correlate with both pre ($r = 0.328, p = 0.003$) and post ($r = 0.521, p < 0.001$) PHQ-8 scores. It suggests students who are more depressed have more variations in depression severity over the term. The PHQ-4 depression subscale slope correlate with post ($r = 0.438, p < 0.001$) PHQ-8 scores but not with the pre scores, which suggests the symptom severity may increase over the term for students who are more depressed by the end of the term.

6.3 Depression Groups Mean Comparison

We show the ANOVA group comparison results in Table 3. In what follows, we discuss the group differences in sleep, conversation, and study behaviors.

Table 3. ANOVA significance of mean term symptom feature differences between the non depressed and depressed group

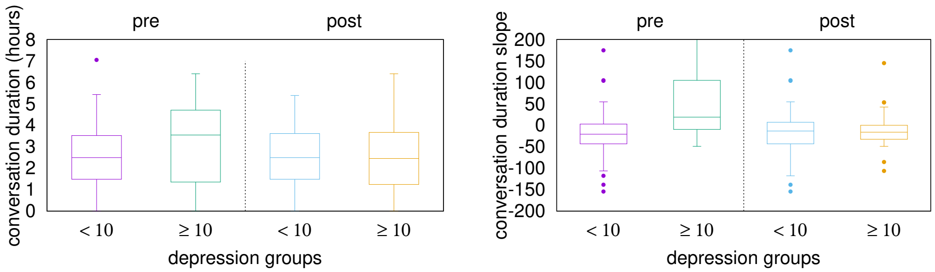
		PHQ-8 pre		PHQ-8 post	
		F	p	F	p
unlock duration	mean	5.179	0.026	5.733	0.019
unlock duration at study places	mean	11.599	0.001	6.084	0.016
unlock duration at study places	std	5.694	0.019	1.426	0.237
unlock duration at dorm	mean	4.748	0.032	6.121	0.016
unlock duration at dorm	std	5.042	0.027	5.443	0.023
conversation duration	slope	13.46	< 0.001	0.379	0.540
time at study places	slope	4.199	0.044	0.546	0.462
PHQ-4 depression subscale	mean	22.240	< 0.001	57.256	< 0.001
PHQ-4 depression subscale	std	0.312	0.578	11.207	0.001
PHQ-4 depression subscale	slope	0.319	0.574	6.716	0.012



(a) time at study places (b) time at study places slope (c) unlock duration at study places

Fig. 4. The distribution of the time at study places, the slope of the time at study places over the term, and the unlock duration at study places of the pre-post PHQ-8 non depressed group and depressed group. Students from the depressed group

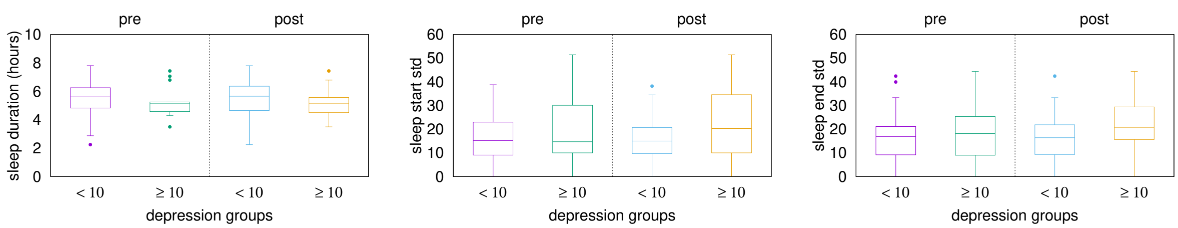
Figure 4 shows the depression groups' distribution of the time spent at study places, the slope of the time spent at study places over the term, and the unlock duration at study places. Figure 4(a) shows that there is no significant differences in time spend at study places between non depressed and depressed groups. Figure 4(b) shows that the pre PHQ-8 depressed group students spend a decreasing amount of time at study places whereas the non depressed group students spend same amount of time at study places across the term ($F = 4.199, p = 0.044$). Figure 4(c) shows that the PHQ-8 pre and post depressed group students spend more time using their phones at study places. The differences are significant with $F = 11.599, p = 0.001$ for the pre PHQ-8 groups and $F = 6.084, p = 0.016$ for the post PHQ-8 groups. Figure 5 shows the distribution of the conversation duration and the conversation duration slope of the pre-post PHQ-8 depression groups. Figure 5(a) shows that there is no significant differences in the mean conversation duration of the pre and post PHQ-8 depressed groups. However, the pre PHQ-8 depressed group shows a large in-group variation in conversation duration. Figure 5(b) shows that the pre PHQ-8 depressed group's conversation duration slope is positive whereas the non depressed group has a slight negative slope. The difference is significant with $F = 13.46, p < 0.001$. The result shows that students in the pre PHQ-8 depressed group are around an increasing amount of conversations as the term progresses whereas students in the pre PHQ-8 non depressed group are around decreasing amount of conversations. The



(a) conversation duration (b) conversation duration slope

Fig. 5. The distribution of the conversation duration and the conversation duration slope of the pre-post PHQ-8 non depressed group and depressed group.

difference in conversation duration slope does not exist in post PHQ-8 groups. The result is consistent with the correlation analysis.



(a) sleep duration (b) sleep start standard deviation (c) sleep end standard deviation

Fig. 6. The distribution of sleep duration, sleep start time standard deviation, and sleep end time standard deviation for the pre-post PHQ-8 non depressed group and depressed group. The group differences are not statistically significant according to ANOVA.

Figure 6 show the distribution of sleep duration, sleep start time standard deviation, and sleep end time standard deviation of the non depression group and the depression group for pre-post PHQ-8. Figure 6(a) shows that the sleep duration of the depressed group is shorter than the non depressed. Figure 6(b-c) show that students in the depressed group have more variations in sleep start and end times. The differences, however, are not statistically significant ($p > 0.05$) according to ANOVA.

6.4 Predicting PHQ-8 Scores

Pre PHQ-8 scores. The lasso regularization selects 10 features to predict pre PHQ-8 scores. Specifically, it selects phone usage at study places, the stationary time, time spend at on-campus health facilities, sleep start time standard deviation, unlock duration at dorm standard deviation, unlock duration at study places slope, conversation duration slope, number of places visited slope, time at on-campus health facilities slope, and heart rate slope. The MAE of the baseline model, where the mean PHQ-8 score is used as the predicted PHQ-8 score, is 3.44. Our prediction model predict the pre PHQ-8 scores with MAE of 2.40, which is 1.04 lower than the baseline. The predicted PHQ-8 score strongly correlate with the groundtruth with $r = 0.741, p < 0.001$.

Post PHQ-8 scores. The lasso regularization selects 5 features to predict post PHQ-8 scores. Specifically, it selects phone usage at study places, the stationary time, number of places visited, sleep start time standard deviation, and conversation duration slope. The MAE of the baseline model is 4.29. Our prediction model predict the pre PHQ-8 scores with MAE of 3.60, which is 0.59 lower than the baseline. The predicted PHQ-8 score strongly correlate with the groundtruth with $r = 0.578, p < 0.001$.

Most of the selected features have shown significant linear correlations with the PHQ-8 outcomes, as shown in previous sections. Interesting enough, the lasso regularization selects the heart rate term slope to predict the pre PHQ-8 scores, whereas heart rate term slope does not show correlations with the PHQ-8 scores. We suspect the heart rate data may provide extra information in predicting PHQ-8 scores when combined with other symptom features.

7 PHQ-4 RESULTS: TRACKING DEPRESSION WEEKLY DYNAMICS

In what follows, we further test our hypothesis by identifying a number of associations between 2-week symptom features and the PHQ-4 depression subscale scores using regression analysis. We use the 2-week symptom features to predict if a student is considered depressed according the PHQ-4 depression subscale.

7.1 Regression Analysis

Table 4 shows the bivariate generalized linear mixed model (GLMM) [50] regression results. We report the coefficient b and the p value associated with the variable (i.e., the 2-week symptom feature) from the bivariate regression between the 2-week symptom features and the PHQ-4 depression subscale scores. The value of the coefficient indicates the direction and strength of the association between symptom features and PHQ-4 depression subscale scores. The p-value indicates the probability that the coefficient is equal to zero. A low p-value (i.e., < 0.05) indicates that the coefficient is not equal to zero and likely to be the learned value.

The results suggest students who are around fewer conversations per day ($p = 0.002$), sleep less ($p = 0.024$), and visit fewer places ($p = 0.003$) are likely to be more depressed. Students who go to sleep late ($p = 0.027$) and wake up late ($p = 0.001$) are likely to be more depressed. The regression coefficients and p-values are similar between the full PHQ-4 dataset and the skip-a-week dataset, which suggests the dependency does not have an impact on the analysis.

Table 4. Associations between symptom features and PHQ-4 depression subscale score

	all		skip a week	
	coefficient	p	coefficient	p
number of conversations	-0.269	0.002	-0.222	0.037
sleep duration	-0.156	0.024	-0.222	0.025
sleep end	0.151	0.027	0.236	0.012
sleep start	0.223	0.001	0.317	0.001
number of visited places	-0.211	0.003	-0.224	0.016

7.2 Prediction Analysis

The regularization selects 9 features to make the prediction as shown in Table 5. Specifically, it selects the stationary time, the number of conversations, heart rate, sleep end time, time spend at a dorm, time spend at study places, phone usage at study places, and the number of places visited. Similar to predicting PHQ-8 scores, the regularization selects the heart rate feature as a predictor. It further suggests the heart rate feature helps predicting depression outcomes.

Table 5. Selected features to predict PHQ-4 depression subscale non depressed and depressed group

lasso selected features
stationary time, number of conversations, heart rate, sleep end, time at dorm, time at study places, unlock duration at study places, unlock number at study places, number of places visited

Figure 7 shows the receiver operating characteristic (ROC) curve [31] of the logistic regression model obtained from the 10-fold crossvalidation. The ROC curve show the different true positive rate and false positive rate using a different threshold to determine if the logistic regression output is a positive case (i.e., depressed) or not. The area under the ROC curve (AUC) [31] is a widely used metric to evaluate a binary classifier. AUC ranges from 0.5 to 1. Higher score indicates better performance. Our PHQ-4 state model's AUC is 0.809, which indicate good prediction performance. The model archives 81.5% of the recall (i.e., 81.5% of the depressed cases are correctly identified) and 69.1% of the precision (i.e., 69.1% of the inferred depressed cases are correct).

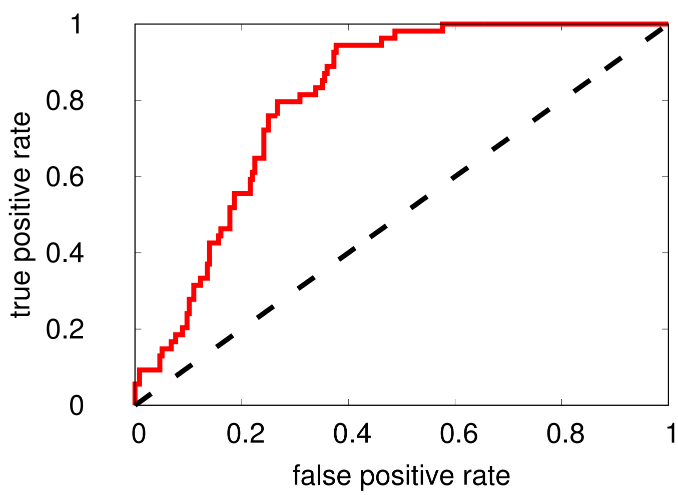


Fig. 7. The ROC curve of using lasso logistic regression to predict PHQ-4 depression states. The area under the ROC curve (AUC) is 0.809.

7.3 Case Study Showing Depression Dynamics

Many of the students in this study have interesting behavioral curves and depression dynamics. Here we highlight one anecdotal case study. Figure 8 shows the depression dynamics of a student's PHQ-4 depression subscale score, number of conversations they are around, sleep duration, bed time, wake time, and number of places visited over a 9-week term. We do not want to identify this student by detailing their academic or demographic information. The curves show the student starts the term in a non depressed state but their PHQ-4 depression subscale score deteriorates as the term progresses and peaks during week 4 (anecdotal this is the mid term week but we have no evidence that this is causal). The student's depression subscale score drops after week 4 and the student reports a non depressed state in week 6 before dropping to 0 in week 8. If we now compare the students behavioral data from the StudentLife app we can observe some interesting trends in the sensor data. Comparing the student's sleep data, the number of places visited every day and number of conversations they are around before and after their PHQ-4 depression subscale score peaks in week 4, we can observe that this student is around fewer conversations, sleep less, goes to bed later at night and wakes up earlier in the morning, and visit fewer places. This all seems indicative of a busy student who might be experiencing increased stress and anxiety because of the increasing academic demands of the term if we only looks at the behavioral curves from sensing data. However, seeing the depression dynamics from the PHQ-4 confirms that this student is struggling with the elevated levels of depression. The student has coping skills that are unknown to us and recovers from this increased risk without (from our data) any visits to the campus health center. As the term ends the student recovers showing resilience their behavioral sensing curves sleeping earlier, getting up later and therefore sleeping longer, visiting more locations on campus during the day, and being around more conversation indicating more engagement with fellow students and less isolation. All around a healthier person at the end of term. We do not present this example as some common set of curves we analyzed but an interesting one that gives insights into this student's academic term.

8 DISCUSSION

Existing research on using passive sensing to predict depression do not design features that explicitly map to depression symptoms defined in DSM 5. For example, prior work propose location features [14, 61, 62] based on the assumption that depressed persons would travel less and have more irregular mobility patterns. In this paper, we propose to use well-known symptoms to guide us design passive sensing features that are more likely to be associated with depression. Incorporating the symptom domain knowledge and the behavioral characteristics of the population (e.g., college students), we can come up with novel behavioral features that leverage multiple sensor streams. For example, with the knowledge of college students' daily routine, we can leverage multiple sensors on the smartphone to assess students are likely to be distracted when they should be studying.

The proposed symptoms features might not be able to generalize to other populations. For example, people with other occupations (i.e., non students) do not usually go to classes. Generic features, such as distance traveled and conversation duration, might not generalize well. We would expect salespersons would be around more conversations than people working in quite offices. Our method shows that we need to tailor behavioral features to different populations.

There are different approaches that can be taken to address the mental health classification and detection problem. The National Institute of Mental Health (NIMH) has launched the Research Domain Criteria (RDoC) [36, 39] project to create a framework for studying mental disorders. The RDoC framework centers around dimensional psychological constructs that are relevant to human behavior and mental disorders [36, 39]. The psychological constructs include negative valence systems, positive valence systems, cognitive systems, systems for social processes, arousal/regulatory systems. RDoC proposes to measure the systems using molecular, genetic, neurocircuit and behavioral assessments [36, 39]. We imagine that future mobile sensing approaches for mental health assessment could focus on developing new sensing modalities and physiological and behavioral features to predict the RDoC constructs. We are pursuing this idea in a related project on auditory verbal hallucinations.

9 LIMITATIONS

While the current study provides evidence that passive sensing may help up track and predict real-world changes in mental health, specifically depression, there are a number of limitations to our study. The sample size of our

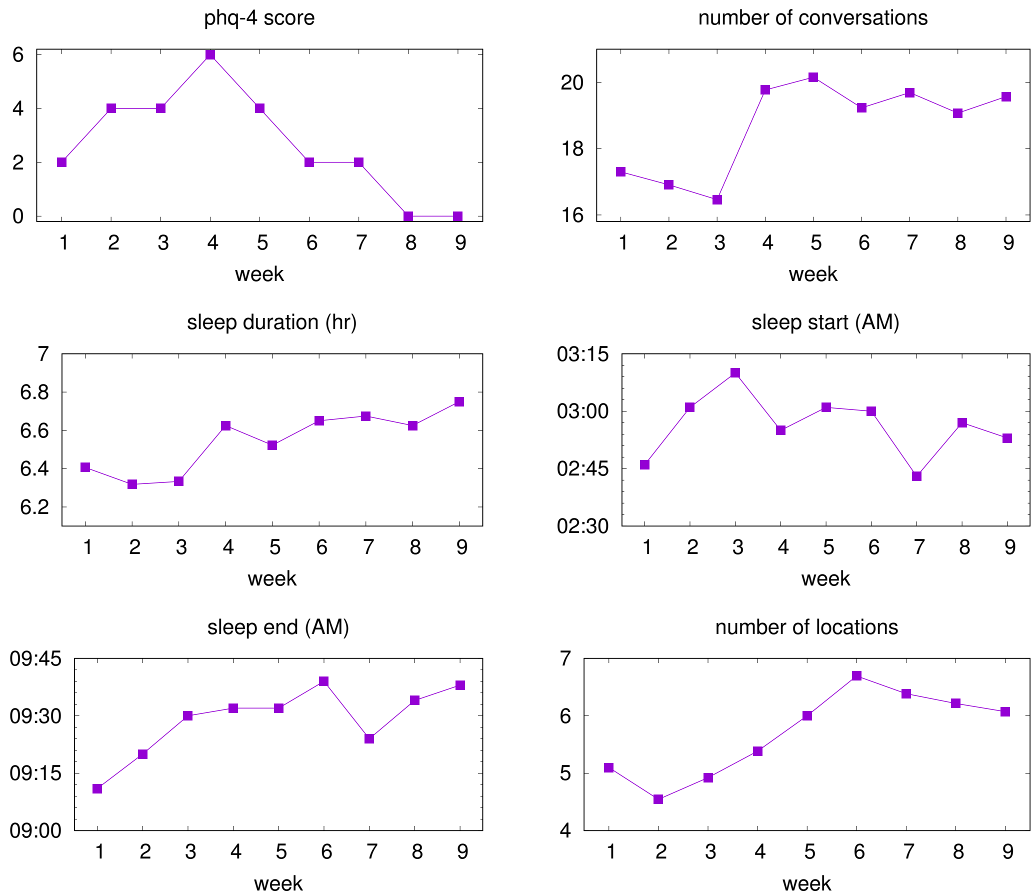


Fig. 8. The dynamics of a student's PHQ-4 depression subscale score, number of conversations around, sleep duration, bed time, wake time, and number of places visited over a 9-week term. The student starts the term in a non depressed state but their PHQ-4 depression subscale score deteriorates as the term progresses and peaks during week 4 and drops to 0 in week 8. The student is around fewer conversations, sleep less, goes to bed later at night and wakes up earlier in the morning, and visit fewer places before week 4. As the term ends the student recovers showing resilience and their behavioral sensing curves sleeping earlier, getting up later and therefore sleeping longer, visiting more locations on campus during the day, and being around more conversation.

study dataset is relatively small relative to the number of features and quantity of data acquired for each of those features. As such, the results here should be considered relatively exploratory and preliminary. There is a need for the community interested in depression sensing to take the next step and conduct a large scale, longitudinal study with a diverse cohort well beyond students. Only then we can conclusively say something strongly about the application of passive sensing to mental health. While our results show associations between symptom features and depression groundtruth, associations cannot tell us if changes in behavior would benefit or worsen depression severity. Future studies, need to be designed with the aim to find causalities between the proposed symptom features and changes in depression as discussed in [70]. Another limitation is that all of the subjects were Dartmouth undergraduates. While our results are statistically significant and encouraging they are limited to students at Dartmouth College. There are a number of other on-going studies looking into a wide-variety of student health issues as part of the CampusLife Consortium [47]. Results from these studies may shine a light on differences across different campuses (e.g., small Ivy in small town, large research university in city).

We evaluate the symptom features using the PHQ-8 item scores. The item scores, however, might not be a good indicator of symptoms because individual PHQ-8 items have not been validated against depression symptoms defined in DSM-5. Future studies may need to collaborate with clinicians to get better individual symptom measures.

Finally, the use of early wearables (in our case the Microsoft Band 2) present problems for longitudinal studies. The band could only hold a charge for approximately 14 hours while continuously sensing. This meant students would have to charge their bands before bed to get any data from the band during the night. Students had to take the band off while showering. Neglecting to put the band back on after showering would also cause losing data. Newer bands such as the Gamin Vivosmart 3 can run for 4 days and are waterproof potentially increasing compliance for wearables and making them more useful.

10 CONCLUSION

We proposed depression symptom features derived from phone and wearable passive sensor data that proxy 5 out of the 9 major depressive disorder symptoms defined in the DSM-5 for college students. We found students who report higher PHQ-8 scores (i.e., are more depressed) are more likely to use their phones more particularly at study places ($r = 0.391, p < 0.001$) in comparison with all day phone usage ($r = 0.282, p = 0.010$); have irregular sleep schedules (i.e., more variations in bed time ($r = 0.301, p = 0.024$) and wake time ($r = 0.271, p = 0.043$); spend more time being stationary ($r = 0.374, p = 0.009$) and visit fewer places during the day ($r = -0.269, p = 0.023$). We identified a number of symptom features capturing depression dynamics during the term associated with the PHQ-4 depression subscale groundtruth. Specifically, students who report higher PHQ-4 depression subscale scores (i.e., are more depressed) are around fewer conversations ($p = 0.002$), sleep for shorter periods ($p = 0.024$), go to sleep later ($p = 0.001$), wake up later ($p = 0.027$), and visit fewer places ($p = 0.003$) during the last two week period. We showed that the symptom features can predict whether or not if a student is depressed each week with 81.5% recall and 69.1% precision. We believe the methods and results presented in this paper open the way for new forms of depression sensing going forward.

ACKNOWLEDGMENTS

The research reported in this article is supported the National Institute of Mental Health, grant number 5R01MH059282-12.

REFERENCES

[1] Saeed Abdullah, Mark Matthews, Ellen Frank, Gavin Doherty, Geri Gay, and Tanzeem Choudhury. 2016. Automatic detection of social rhythms in bipolar disorder. *Journal of the American Medical Informatics Association* 23, 3 (2016), 538–543.

[2] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. 2016. Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors. *IEEE Transactions on Affective Computing* (2016).

[3] American College Health Association. 2016. American College Health Association-National College Health Assessment II: Reference Group Executive Summary Fall 2016. *Hanover, MD: American College Health Association* (2016).

[4] Apple. 2017. Core Motion. (2017). https://developer.apple.com/reference/coremotion.

[5] American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®).* American Psychiatric Pub.

[6] Min Aung, Faisal Alquaddoomi, Cheng-Kang Hsieh, Mashfiqui Rabbi, Longqi Yang, JP Pollak, Deborah Estrin, and Tanzeem Choudhury. 2016. Leveraging Multi-Modal Sensing for Mobile Health: a Case Review in Chronic Pain. *IEEE Journal of Selected Topics in Signal Processing* 10, 5 (2016), 1–13.

[7] P Bech, N-A Rasmussen, L Raabæk Olsen, V Noerholm, and W Abildgaard. 2001. The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. *Journal of affective disorders* 66, 2 (2001), 159–164.
[8] Aaron T Beck, David Guth, Robert A Steer, and Roberta Ball. 1997. Screening for major depression disorders in medical inpatients with the Beck Depression Inventory for Primary Care. *Behaviour research and therapy* 35, 8 (1997), 785–791.
[9] Aaron T Beck, Robert A Steer, Gregory K Brown, et al. 1996. Beck depression inventory. (1996).
[10] 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* (2014), sbu033.
[11] Dror Ben-Zeev, Rui Wang, Saeed Abdullah, Rachel Brian, Emily A Scherer, Lisa A Mistler, Marta Hauser, John M Kane, Andrew Campbell, and Tanzeem Choudhury. 2015. Mobile behavioral sensing for outpatients and inpatients with schizophrenia. *Psychiatric services* 67, 5 (2015), 558–561.
[12] Dror Ben-Zeev, Michael A Young, and Patrick W Corrigan. 2010. DSM-V and the stigma of mental illness. *Journal of Mental Health* 19, 4 (2010), 318–327.
[13] Alison L Calear and Helen Christensen. 2010. Systematic review of school-based prevention and early intervention programs for depression. *Journal of adolescence* 33, 3 (2010), 429–438.
[14] 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.
[15] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tonmoy Choudhury, and Andrew T Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE, 145–152.
[16] Tanzeem Choudhury, Sunny Consolvo, Beverly Harrison, Jeffrey Hightower, Anthony LaMarca, Louis LeGrand, Ali Rahimi, Adam Rea, G Borriello, Bruce Hemingway, et al. 2008. The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Computing* 7, 2 (2008).
[17] 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).
[18] Patrick Corrigan and Alicia Matthews. 2003. Stigma and disclosure: Implications for coming out of the closet. *Journal of mental health* 12, 3 (2003), 235–248.
[19] Dartmouth College Office of Institutional Research. 2016. Dartmouth Student Health Survey. (2016). http://www.dartmouth.edu/ oir/2016-dartmouth-health-survey-final-web-version.pdf.
[20] Kadir Demirci, Mehmet Akgönül, and Abdullah Akpinar. 2015. Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students. *Journal of behavioral addictions* 4, 2 (2015), 85–92.
[21] Olive Jean Dunn. 1961. Multiple comparisons among means. *J. Amer. Statist. Assoc.* 56, 293 (1961), 52–64.
[22] Daniel Eisenberg, Ezra Golberstein, and Sarah E Gollust. 2007. Help-seeking and access to mental health care in a university student population. *Medical care* 45, 7 (2007), 594–601.
[23] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD '96*. AAAI Press, 226–231.
[24] Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. 2016. Behavior vs. Introspection: Refining prediction of clinical depression via smartphone sensing data. In *7th Conference on Wireless Health, WH*.
[25] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232. http://www.jstor.org/stable/2699986
[26] Susan R Furr, John S Westefeld, Gaye N McConnell, and J Marshall Jenkins. 2001. Suicide and depression among college students: A decade later. *Professional Psychology: Research and Practice* 32, 1 (2001), 97.
[27] Steven J Garlow, Jill Rosenberg, J David Moore, Ann P Haas, Bethany Koestner, Herbert Hendin, and Charles B Nemeroff. 2008. Depression, desperation, and suicidal ideation in college students: results from the American Foundation for Suicide Prevention College Screening Project at Emory University. *Depression and anxiety* 25, 6 (2008), 482–488.
[28] Ginger.io. 2017. Ginger.io. (2017). https://ginger.io/.
[29] Google Activity Recognition Api. 2017. Google Activity Recognition Api. https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognitionApi. (2017).
[30] Max Hamilton. 1960. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry* 23, 1 (1960), 56.
[31] James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36.
[32] C Haring, R Banzer, A Gruenerbl, S Oehler, G Bahle, P Lukowicz, and O Mayora. 2015. Utilizing Smartphones as an Effective Way to Support Patients with Bipolar Disorder: Results of the Monarca Study. *European Psychiatry* 30 (2015), 558.

[33] Treniece Lewis Harris and Sherry Davis Molock. 2000. Cultural orientation, family cohesion, and family support in suicide ideation and depression among African American college students. *Suicide and Life-Threatening Behavior* 30, 4 (2000), 341–353.
[34] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[35] Joel W Hughes and Catherine M Stoney. 2000. Depressed mood is related to high-frequency heart rate variability during stressors. *Psychosomatic medicine* 62, 6 (2000), 796–803.
[36] Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. (2010).
[37] Richard Kadison and Theresa Foy DiGeronimo. 2004. *College of the overwhelmed: The campus mental health crisis and what to do about it.* Jossey-Bass.
[38] Andrew H Kemp and Daniel S Quintana. 2013. The relationship between mental and physical health: insights from the study of heart rate variability. *International Journal of Psychophysiology* 89, 3 (2013), 288–296.
[39] Michael J Kozak and Bruce N Cuthbert. 2016. The NIMH research domain criteria initiative: background, issues, and pragmatics. *Psychophysiology* 53, 3 (2016), 286–297.
[40] Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric Annals* 32, 9 (2002), 509–515.
[41] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The PHQ-9. *Journal of general internal medicine* 16, 9 (2001), 606–613.
[42] Kurt Kroenke, Robert L Spitzer, Janet BW Williams, and Bernd Löwe. 2009. An ultra-brief screening scale for anxiety and depression: the PHQ–4. *Psychosomatics* 50, 6 (2009), 613–621.
[43] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders* 114, 1 (2009), 163–173.
[44] Min Kwon, Dai-Jin Kim, Hyun Cho, and Soo Yang. 2013. The smartphone addiction scale: development and validation of a short version for adolescents. *PloS one* 8, 12 (2013), e83558.
[45] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. 2010. A survey of mobile phone sensing. *Communications Magazine, IEEE* 48, 9 (2010), 140–150.
[46] Nicholas D Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. 2011. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare*. 23–26.
[47] Georgia Tech Campus Life. 2017. Campus Life | Optimizing the Student Environment. (2017). http://www.quantifiedcampus.gatech.edu/.
[48] Marek Malik. 1996. Heart rate variability. *Annals of Noninvasive Electrocardiology* 1, 2 (1996), 151–181.
[49] Alban Maxhuni, Angélica Muñoz-Meléndez, Venet Osmani, Humberto Perez, Oscar Mayora, and Eduardo F Morales. 2016. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing* (2016).
[50] Charles E McCulloch and John M Neuhaus. 2001. *Generalized linear mixed models.* Wiley Online Library.
[51] 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.* ACM, 1132–1138.
[52] Microsoft. 2016. Microsoft Band. (2016). https://www.microsoft.com/microsoft-band/en-us.
[53] Megan A Moreno, Lauren A Jelenchick, Katie G Egan, Elizabeth Cox, Henry Young, Kerry E Gannon, and Tara Becker. 2011. Feeling bad on Facebook: Depression disclosures by college students on a social networking site. *Depression and anxiety* 28, 6 (2011), 447–455.
[54] Christopher JL Murray, Jerry Abraham, Mohammed K Ali, Miriam Alvarado, Charles Atkinson, Larry M Baddour, David H Bartels, Emelia J Benjamin, Kavi Bhalla, Gretchen Birbeck, et al. 2013. The state of US health, 1990-2010: burden of diseases, injuries, and risk factors. *JAMA* 310, 6 (2013), 591–606.
[55] Christopher JL Murray, Theo Vos, Rafael Lozano, Mohsen Naghavi, Abraham D Flaxman, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, et al. 2013. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 380, 9859 (2013), 2197–2223.
[56] Venet Osmani. 2015. Smartphones in mental health: detecting depressive and manic episodes. *IEEE Pervasive Computing* 14, 3 (2015), 10–13.
[57] Venet Osmani, Alban Maxhuni, Agnes Grünerbl, Paul Lukowicz, Christian Haring, and Oscar Mayora. 2013. Monitoring activity of patients with bipolar disorder using smart phones. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia.* ACM, 85.
[58] Skyler Place, Danielle Blanch-Hartigan, Channah Rubin, Cristina Gorrostieta, Caroline Mead, John Kane, Brian P Marx, Joshua Feast, Thilo Deckersbach, et al. 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).
[59] 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.

[60] Matthia Sabatelli, Venet Osmani, Oscar Mayora, Agnes Gruenerbl, and Paul Lukowicz. 2014. Correlation of significant places with self-reported state of bipolar disorder patients. In *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*. IEEE, 116–119.

[61] Sohrab Saeb, Emily G Lattie, Stephen M Schueller, Konrad P Kording, and David C Mohr. 2016. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ* 4 (2016), e2537.

[62] 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).

[63] SAMHSA. 2015. Key Substance Use and Mental Health Indicators in the United States: Results from the 2015 National Survey on Drug Use and Health. (2015). https://www.samhsa.gov/data/sites/default/files/NSDUH-FFR1-2015/NSDUH-FFR1-2015/NSDUH-FFR1-2015.htm.

[64] Henry Scheffe. 1999. *The analysis of variance*. Vol. 72. John Wiley & Sons.

[65] Suzanne C Segerstrom and Lise Solberg Nes. 2007. Heart rate variability reflects self-regulatory strength, effort, and fatigue. *Psychological science* 18, 3 (2007), 275–281.

[66] John Shawe-Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge university press.

[67] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, Patient Health Questionnaire Primary Care Study Group, et al. 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama* 282, 18 (1999), 1737–1744.

[68] Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. 2017. DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 715–724.

[69] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[70] Fani Tsapeli and Mirco Musolesi. 2015. Investigating causality in human behavior from smartphone sensor data: a quasi-experimental approach. *EPJ Data Science* 4, 1 (2015), 24.

[71] Verily. 2017. Tackling Mental Health at Verily. (2017). https://blog.verily.com/2017/05/tackling-mental-health-at-verily.html.

[72] Theo Vos, Abraham D Flaxman, Mohsen Naghavi, Rafael Lozano, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, Victor Aboyans, et al. 2013. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 380, 9859 (2013), 2163–2196.

[73] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. 2016. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth* 4, 3 (2016).

[74] Rui Wang, Min S. H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A. Scherer, Vincent W. S. Tseng, and Dror Ben-Zeev. 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 (UbiComp '16)*. ACM, New York, NY, USA, 886–897. https://doi.org/10.1145/2971648.2971740

[75] 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 (UbiComp '14)*. ACM, New York, NY, USA, 3–14. https://doi.org/10.1145/2632048.2632054

[76] 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.

[77] Rachel Yehuda. 2002. Post-traumatic stress disorder. *New England journal of medicine* 346, 2 (2002), 108–114.

[78] Yosef Hochberg Yoav Benjamini. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300. http://www.jstor.org/stable/2346101

Received May 2017; revised November 2017; and accepted January 2018.