STRESSASSESS: Detection Of Individual Perceived Stress and Depression Using Passive Location Tracking Systems At Scale

ABSTRACT

Work-related stress is an increasingly common affliction in all walks of life. When left unmanaged, work stress can inhibit productivity, or even cause psychological distress such as depression. There has been a sharp rise in mHealth initiative to assess and manage stress. However, they usually require participants to install mobile or custom apps or provide regular survey data; making these solutions harder to scale. In this paper, we present *StressAssess* , a community-wide sensing solution, which does not require users to own specific devices or install any software. Instead, *StressAssess* solely uses coarse-grained location data collected from the infrastructure, not client devices directly, and extracts an extensive set of quantitative measures of work routines and group behaviors to automatically detect severe stress, every 6 days, and critical levels of depression, every 15 days, with an ML-based classification method. Our extensive semester-long field studies at an Asian university with 108 active undergraduate participants showed that *StressAssess* achieved an ideal 100% sensitivity (88.79% balanced accuracy) in detecting stress for our main study, and up to 92% sensitivity (91.12% balanced accuracy) in detecting depression on all IS-major students.

1. INTRODUCTION

Work-related factors are one of the primary sources of stress, be it employment among adults [31], or education among students [29]. Without losing its authentic meaning, we refer to 'work' as activities that require skills, time and effort to complete. Cox *et al.* characterizes work-related stress by *work content* (e.g. task, load), and to a large degree, driven by *work context* (e.g. interpersonal relationships) [12, 13]. Stress comes as an immediate response, especially in highly competitive environments, and can motivate us to perform effectively and efficiently. However, overwhelming amounts of stress can compromise work performance and impact health. For example, it can trigger the early onset of depression [12].

There is great value in understanding and tackling work-related stress. Governments, organizations and institutions are increasingly proactive in implementing stress management programs and improving public awareness of mental health problems [26]. Yet, these efforts struggle against the economy of time, common perception of stress as 'normal' [14, 24], and social stigmatization [27]. Thus, health research turns to *mHealth* solutions to provide interventions conveniently and discreetly through one's personal device. Unfortunately, there are drawbacks to existing systems. First, *mHealth* apps bear higher privacy risks as a high amount of personal data is collected for real-time analysis [28]. Active sensing is also resource intensive. Most importantly, the strong focus on sensing personal factors, unintentionally, diminishes the importance of understanding and addressing the unique collective factors that can cause work-related stress (e.g., group interaction at work).

We aim to build a solution that complements existing health measuring systems [18, 40, 41]. *StressAssess* is a solution that allows the *passive detection of perceived stress in individuals using location tracking mechanisms that can be deployed at scale in a work environment*. By passive, we mean that data is not directly or actively sensed from a user's personal device, but collected from the infrastructure directly – in this case we collect RSSI information about every connected device directly from every WiFi access point. By scale, we mean that the solution should work for 10 to hundreds of thousands of individuals across the target environment, such as a university campus, automatically, without needing to install any software on any device. Currently, our solution uses WiFi as it is readily available and widely used across the environments we considered. The key questions addressed are:

1. Can we detect perceived stress in a person using only their location data?
2. Can we detect if someone has the more serious condition of depression using only their location data?

Note: we discuss the privacy implications of our approach and what we do to preserve it in Section 9.3.

While established works [40, 45] inspired our research, our work makes several clear distinctions: First, we build *StressAssess* on the hypothesis that *changes in a person's movement patterns and interaction habits with others are key indicators of perceived stress.* This hypothesis is derived from research that substantiates how behavior changes as a result of mental illness [10]. Our challenge then lies in enabling this hypothesis at scale. Hence, the second characteristic is using passive coarse-grained location tracking sub-systems as the underlying technology to draw behavioral proxies in the work environment; They are (1) a WiFi-fingerprinted indoor localization system that allows tracking of any device (we use mobile phones) regardless of OS, and (2) a group detector system that utilizes the same localization system to derive proxies for human interaction as devices 'move together'. Both systems were successfully deployed, and running in our test environment since 2013.

We conducted two rounds of longitudinal studies, each over 81 days (2 full academic semesters each), and one smaller-scale study over 36 days, with different student populations at an Asian university. These studies extend the *mHealth* research with findings from an Asian country with different norms and attitudes [7, 37]. The primary study, $Study_{SE}$, consisting of 62 active participants, was used to build a stress-depression detection model. We paid particular attention to students enrolled in Software Engineering (SE), a core module that is anecdotally reported to be the most stressful course among our students as they are preassigned to groups with people they do not know and must work with their group to build a fairly complicated web application. The other studies, $Study_{MIX}$ and $Study_{ScE}$, externally validate our models, with 35 and 11 active participants respectively. In all studies, participants provided regular assessments of their mental states using well-established scales (PSS-4 [11] and PHQ-8 [21]), team meeting logs, and MAC addresses of their mobile phones so that we could identify them in our campus deployed WiFi-based location system (by default the location system anonymises all MAC addresses). These score assessments were used as ground truth to validate our model. Our study revealed the following key insights:

1. Course-grained location data is a useful indicator of perceived stress from which we can extract rich sets of features that well represent students' campus routines and interactions. In particular, the only sensor data we collect directly from every access point in the environment is the RSSI values of all connected devices. We do not have access to any cell phone or app data.

2. The strongest predictors of *severe stress* are changes in feature values that individuals make in reference to themselves and their population, and group features that represent interaction habits.

3. Depression requires a longer measurement window to detect accurately, as it intensifies over longer periods, compared to *severe stress*, which can be detected in shorter spans.

Overall, this paper makes the following contributions:

1. This paper is the first to show the feasibility of assessing stress and depression at scale using just coarse location data created from WiFi signal data collected from access points directly. In particular, we do not collect data from a user's device or require the installation of any applications. With users not having to "opt-in", our method enables community-wide assessment and creates an opportunity to provide a safety net for people in greatest need of help with stress.

2. Using only coarse-grained location data, a rich set of mobility features can be extracted to broadly represent routines and group interaction habit. Specifically using "temporal behavior changes" and "interaction patterns" as critical features, we demonstrate their effectiveness in detecting stress at a 6 day interval and depression at a 15 day interval.

3. We rigorously evaluated our stress detection models with 108 participants across three different multi-week studies, and achieved on 100% sensitivity and 88.79% balanced accuracy for our main study. Similarly, we built depression detection models using samples collected at different periods, and achieved up to 92% sensitivity and 91.12% balanced accuracy using personality profile as additional features.

2. MOTIVATION SCENARIOS

StressAssess was designed to offer community-wide mental health assessment as a safety net that could then be used to provide early intervention. Consider the following scenario.

University student counsellors' job would be greatly enhanced were they able to keep track of the stress level of the student population, and thereby help students manage stress more effectively. They would also hope to promote their mindfulness program and encourage students who need help to seek their counsel. These counsellors would prefer to promote their services ideally when the stress level of students' rise significantly (note that students consider frequent e-mails as spam). In addition, they would like to share the current community stress levels with lecturers to increase awareness of students' mental health – lectures can adjust load or deadlines for the current or future curriculums. Finally, the ability to monitor community's overall mental health provides counsellors with invaluable insights on groups or individuals who demonstrate a need for help – the counsellors could then be more proactive in providing necessary support.

StressAssess hopes to provide these capabilities in a scalable manner. By using connectivity data collected from the campus WiFi network (from which coarse-grained device locations are computed), it can assess the level of stress for a sufficiently large subsample of the entire student population. Unlike prior solutions, it does not require installing any data collection applications on client devices or asking for client-generated self-reports; thereby making community-wide deployment and monitoring significantly easier.

There are many other work environments (e.g., hospitals) where people experience a high level of stress, and stress management is critical. In these situations, *StressAssess* has the unique capability to conduct scalable and unobtrusive assessment of community-wide stress without requiring any user actions. By using the monitored results, various early interventions can be designed to improve the overall community health. Such early interventions can also increase mental health awareness at both individual and community levels.

Overall, *StressAssess* is designed to be a first-level safety net that provides mental health information about the entire population. It thus nicely complements more fine-grained solutions that require installing active stress trackers etc. that can be used to better understand the health of specific individuals who desire closer monitoring.

3. BACKGROUND AND RELATED WORK

Work Stress

Work shapes individual and social life. The term 'work' is not by any means clearly defined, and therefore reasonable to assume that its meaning and importance vary widely among populations [43]; employees work to gain economic independence, and students work on their academic abilities to objectively pursue their career aspirations. Dohrenwend & Dohrenwend grouped over 500 subjects according to age, gender and ethnicity, and found that most of the highly ranked stressful events were related to work [31]. Work stress can be caused by pressing working conditions, such as work overload and relationship conflicts among colleagues [2], supporting Cox *et al.*'s claim that work stress is driven by *work content* (e.g. task, load), and *work context*. To tackle the problem of assessing work stress, these factors must be equally accounted for.

Health And Behavioral Monitoring

Research in *mHealth* space continuously offers new capabilities to monitor psychological states and mental conditions in real-time [39, 40, 45]. The early works of *StudentLife*, spearheaded by Wang *et al.* assesses mental health and performance of university students using fine-grained sensor data collected directly from mobile devices [40]. Most recently, the authors looked at symptoms features to predict depression scores [41]. In specifically using location data, Canzian *et al.* explores the correlation between GPS-based location features and depressions [10]. While informative, relying on GPS suffers from insufficient indoor precision. Brown *et al.* bridged this gap by using wearable RFID tags to collect indoor location traces of employees interacting with colleagues in different building spaces [8]. Zhou *et al.* proposes *EDUM*, which uses Wi-Fi indoor localisation data to learn about student behaviors [45], and is conceptually closest to ours. However, *EDUM* does not infer the characterizations of workgroups and detect stress

Community Wide Sensing

While much research has been dedicated to developing sensing applications that scale from individuals to target communities [22], many applications of technologies are in the areas of urban planning [3] and security [44], for example, using community-wide video surveillance for purposes of public safety. In contrast, sensing mental health such as stress and depression at scale remains unexplored.

These bodies of work inspire our research, but our proposed solution makes several distinctions. First, our solution can scale easily as we rely solely on sensing data passively from the WiFi infrastructure and not from users's phone. Second, we explore group behaviors on campus environment.

4. SEMESTER LONG FIELD STUDY

To design and validate our system, we conducted three rounds of user studies at an Asian university between 2017 to 2018. Colleagues from our behavioral psychology department as well as psychiatrists from the main local mental health hospital validated our IRB approved study procedure.

4.1 Participants

Table 1 summarizes the demographics of our study. It began with $Study_{SE}$, as the main study conducted for 81 days in Fall AY2017. All 76 participants (Information Systems major, Sophomore) were enrolled in the Software Engineering (SE) module that requires students to work in pre-assigned groups to build a semester-long project. SE groups are made up of 5 pre-assigned members, with each group balanced along multiple dimensions such as gender, GPA, nationality, and programming capability. Consequently, as the groups are pre-assigned and balanced, students find themselves in a situation where they must work with people they do not know and have varying capabilities, personalities, and work styles. Students get to know the project requirements and team grouping on the first day of class, and each team has the full semester to build a fully working software artifact (a fairly complicated web application deployed to a commercial cloud provider) while following the various processes required – these processes require team members to spend equal amounts of time coding, enforce pair programming for coding tasks, require rotation of programming pairs etc. We chose this course for our study as it is continuously cited as a substantially stressful course, mostly due to the pressures of having to work closely with people whom one does not know, and who may not share the same capabilities and aspirations.

$Study_{ScE}$ (13 participants) and $Study_{MIX}$ (51 participants) are validation studies conducted in Spring AY2017 (36 days) and Fall AY2018 (81 days), respectively. Students ranged from Freshmen to Seniors, and were enrolled in different majors and modules. In all these modules, students would form groups to work on either semester-long or short-term

	$\textit{\textbf{Study}}_{SE}$ (main)	$\textit{\textbf{Study}}_{ScE}$ (validation 1)	$\textit{\textbf{Study}}_{MIX}$ (validation 2)
Period	Fall AY2017, 81 days	Spring AY2017, 36 days	Fall AY2018, 81 days
Total	76 students (39 M, 37 F)	13 students (3 M, 10 F)	51 students (24 M, 27 F)
Active	62 students (34 M, 28 F)	11 students (3 M, 8 F)	35 students (15 M, 20 F)
Team	50 students	0 students	25 students
Individual	12 students	11 students	10 students
Age	19 - 25 (22 med)	20 - 24 (22 med)	19 - 26 (22 med)
GPA	1.64 - 3.84 (2.85 med)	2.90 - 3.99 (3.33 med)	0 - 3.78 (2.63 med)
Major	Information Systems (62)	Finance (1) Business management (9) Social Sciences (1)	Information Systems (31) Business Management (1) Economics (2) Accountancy (1)
Study year	Sophomore (62)	Sophomore (3) Junior (3) Senior (5)	Sophomore (12) Junior (9) Senior (1) Freshman (13)
Course	Software Engineering (62)	Social Entrepreneurship (11)	Software Project Management (8) Interaction Design & Prototyping (13) Computational Thinking (1) Information Systems & Innovation (7) Programme in Writing & Reasoning (6)

Table 1: Demographics summary of participants from our main and 2 validation studies. GPA ranges from 0-4, 0 due to Freshmen with no GPA.

projects. None of these courses, unlike SE, required intensive juggling between technical and management practice.

As with other longitudinal studies, managing user retention is critical yet inherently challenging. At the end of our study, 62 ($Study_{SE}$), 35 ($Study_{MIX}$), and 11 ($Study_{ScE}$) active students remained. These are the students who contributed at least 80% of all survey data (used for ground truth validation) and attended at least one interview session. We validated from our mid-semester and end-of-semester interview findings that none of our inactive participants had reported experiencing a highly stressful semester.

4.2 Study Procedure

Each participant filled out a pre-study questionnaire outlining their personality traits using the Big-5 Inventory [20], their current GPA, and their normal campus routines (e.g. meal breaks, sports, extra-curricular activities). During the study, participants reported their perceived stress levels using the PSS-4 [11] survey every three days, and a retrospective assessment to measure depression using the PHQ-8 [21] survey every three weeks. Before this study, we were strongly advised by practicing psychiatrists to use the PHQ-8, instead of the more common PHQ-9, as it omits the last question on suicide, thus avoiding insinuating suicidal thoughts on participants. Additionally, semi-structured interviews were arranged with participants at recess week, and at the end of the semester to better understand their sources of stress, experiences of team-related stress, and how these stresses were managed. Students participated in face-to-face interviews for $Study_{SE}$, and online interviews for $Study_{MIX}$ and $Study_{ScE}$. We used cross-validated survey and interview data as ground truth. Finally, $Study_{SE}$ participants provided access to their SE project schedule (a graded document maintained by all teams to keep track of SE work progress and plans), while $Study_{MIX}$ and $Study_{ScE}$ students provided meeting logs every three days. These schedules and logs included information on meeting dates, duration, and location.

To enhance our data collection process, we administered all surveys using Qualtrics [30] and embedded the surveys in a custom online portal, built using the October-CMS open-source PHP platform [25]. Our portal, as illustrated in Figure 1, served as the primary platform for survey access, facilitating reminders for participation and updates of earnings. Finally, we compensated the participants with a maximum amount of USD 30 in two ways, 1) for entering the study, and 2) for remaining active. The amount of compensation varied depending on how actively they provided self-reported surveys. Active participants were also eligible for a lucky draw to win a USD 76 cash prize. Besides, we offered USD 37 as a bonus to participants, whose entire project group joined and completed the study.

4.3 Location Data Collection

The majority of our data consisted of location and group data, derived from connection logs between participants' mobile phones and the campus WiFi network. We used a WiFi location system, developed previously [1] and deployed across the entire campus, that uses RTLS (real-time location services) to extract RSSI information about every connected WiFi device directly from every access point in the environment. These raw RSSI values are then processed, using

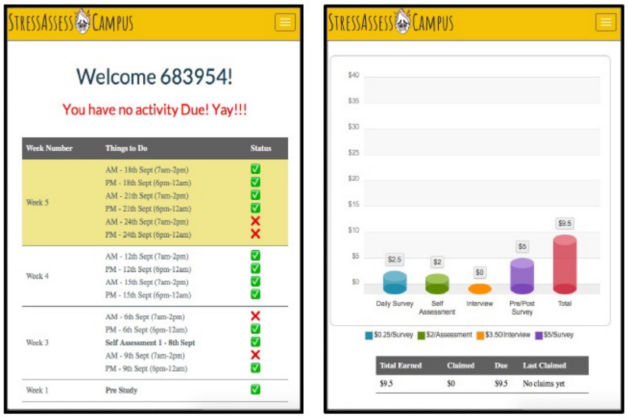


Figure 1: *StressAssess* survey portal used for data collection.

reverse triangulation with many heuristics and filters, to produce the final location of every connected WiFi device.

For this study, we extracted three months worth of WiFi localization data for $Study_{SE}$ and $Study_{MIX}$ and one-month data from $Study_{ScE}$. This localization data was then processed by a group detector, using techniques similar to Grumon [36], to produce group information. For $Study_{SE}$, this resulted in 310,000 processed location data points and 89,280 processed group data points for all 62 participants each day, which is equivalent to an average of 7 hours worth of location and 2 hours of group interaction data per participant (i.e., 310,000/62 * 5 seconds and 89,280/62 * 5 seconds). Participants were also detected to have visited, on average, 96 unique locations on campus each month. We extracted mobility features from location and group data points on a per day basis and mapped them to stress (PSS-4) and depression (PHQ-8) labels. In summary, the final dataset consists of 5,022 data points (62 * 81-days) from $Study_{SE}$, 2835 data points from $Study_{MIX}$, and 468 data points from $Study_{ScE}$.

4.4 Creating Ground Truth of Stress and Depression

Self-reports of PSS-4 and PHQ-8, coupled with verification of students' experiences from two interview findings constituted our main mechanisms for collecting ground truth. We now describe how we grouped the assessment scores.

Converting Stress Scale

PSS-4 is a well-established scale, ranging from 0 to 16, used among students and employees [33, 42]. While a score close to 16 suggests high perceived stress, it is not designed as a diagnostic tool. Hence, we referenced work by Wartig *et al.* that provides norms for an English sample (N>1500, with various ethnicity: White, Mixed, Black African, and Asian) for PSS-4 [42]. In computing a 95% prediction interval, the upper bound ends up being 6.11 + 3.14(1.96) = 12.26. This implies that a PSS-4 score over 12.26 would place someone above the 95th percentile of the population, supporting our interpretation of 12+ as *severe stress*. Moreover, the norms

	normal stress	***severe stress***
$Study_{SE}$	1529 (91%)	145 (9%)
$Study_{MIX}$	944 (99%)	1 (1%)
$Study_{ScE}$	129 (98%)	3 (2%)
	non-critical depression	***critical depression***
$Study_{SE}$	1395 (83%)	279 (17%)
$Study_{MIX}$	722 (76%)	223 (24%)
$Study_{ScE}$	122 (92%)	10 (8%)

Table 2: Distribution of stress and depression labels for all studies.

in the study place score of PSS-4 = 9+ in the moderate group that is ≈1 SD above the mean (6.11 + 3.14). A score of PSS-4 = 12+ places someone in the severe category, 2 SD above the mean. Accordingly, scores of 12 and above are grouped as **1:*severe stress*** (positive class), otherwise **0:*normal stress***.

Converting Depression Scale

The use of PHQ-8 is more straightforward as the scale is a diagnostic tool with clear cutoffs – minimal: 0-4, mild: 5-9, moderate: 10-14, moderate-severe: 15-19, and severe: 20-24. We are particularly concerned with students falling in the "moderate-severe" and "severe" depression categories, and consider these occurrences as "critical". Accordingly, PHQ-8 scores of 15 and above are grouped as **1:*critical depression*** (positive class), otherwise **0:*non-critical depression***.

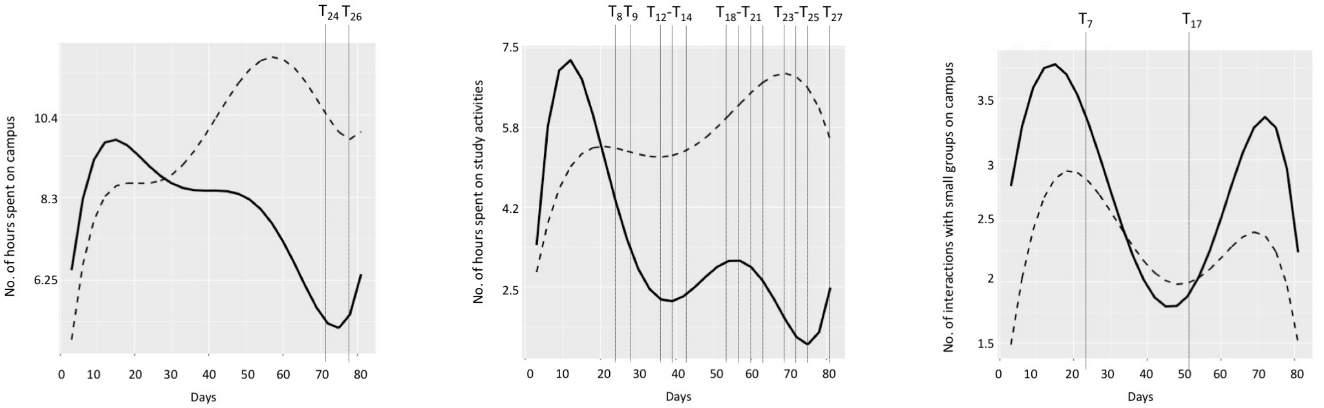
Distribution of Labels

Table 2 lists the distribution of labels. The PSS-4 conversion resulted in a distribution of more than 90% *normal stress* labels for all studies. Prior work [9] suggests that this imbalance in labels, seen in our ground truth, is to be expected as individuals overwhelmed by stress tend to be outliers. Such skewed datasets can lead to poor prediction performance if not corrected [9]. Thus, we addressed this problem by applying SMOTE [6] to synthetically oversample training set data in the *severe stress* and *critical depression* classes.

5. FEASIBILITY STUDY

We explored the feasibility of detecting stress using only coarse-grained location data collected from the campus WiFi network. We developed a set of mobility-driven features and hypotheses based on interview studies with students. Hypothesis testing was performed to statistically validate features that can differentiate students experiencing *severe stress*. For the ground-truth labels, we placed the students in $Study_{SE}$ into *severe stress* (n=4) and *normal stress* (n=58) categories, based on their self-reported average PSS-4 scores.

The analysis was conducted as follows; First, we visually examined the changes in mobility features over time, between the two groups, as shown in Figure 2 by averaging features every three days, and plotting them over the study duration of 81 days. We define *Time Point*, T_x as a sample made every 3 days – i.e., T_{24} = 24 * 3 = day 72 of the study.



(a) Total time spent on campus - T24 and T26 are significantly different between groups (p=0.04).

(b) Time spent on work-typed activities - Multiple time points are significantly different between groups (p<0.001).

(c) Participation in small-sized groups - T7 and T17 are significantly different time point between groups (p=0.04).

For each figure, the dashed line is the data for students with *normal stress*, and the solid line is the data for students with *severe stress*. We highlight the time points where the two distributions were statistically significantly different. Note: all other points, even where they appear to be different, were not significantly different. Each time point is computed across a 3 day window. so T_{24} = day 72 (24 * 3) of the study.

Figure 2: Mean plots of different mobility features between the *normal stress*and *severe stress*groups.

Second, we performed one-way MANOVA to investigate the significance of the multivariate mean effects on different features, and ran individual t-tests with Bonferroni correction to check for specific mean differences across time periods.

Campus Routines

We formulated hypotheses beginning with a conjecture that students with *severe stress* are more likely to reduce their interactions with working peers on campus.

H1: Students with *severe stress* spend fewer hours on campus

Overall, we observed that students with *normal stress* incrementally spent more time on campus, especially towards the second half of the semester. This is an expected trend because students typically spend more time working on projects and preparing for examinations. Yet, students with *severe stress* were found to spend significantly less time on campus (p=0.04), specifically on T_{24} and T_{26} (see Figure 2a).

H2: Students with *severe stress* participate more in work activities on campus

Interestingly, we observed significantly less involvement in work activities for students with *severe stress* (p<0.001) than students with *normal stress* (see Figure 2b). While students with *severe stress* began with more work participation, this pattern did not persist. In addition, two large dips occurred around the recess week (T_{12}-T_{14}) and the end of the semester (T_{23}-T_{25}). These time points closely corresponded to important project milestones for the software engineering (SE) course and were significantly different between both groups.

Group Interaction

With SE emphasising on teamwork, we believe group interaction is a crucial indicator for stress. We hypothesize:

H3: Students with *severe stress* are likely to be more engaged in group work activities on campus

From Figure 2c, we found that students with *severe stress* spent significantly more time with small groups (p=0.04), with a significant difference at T_7. However, a noticeable dip happened during the recess week, with a significant difference at T_{17}. Additionally, we found that students with *severe stress* spent significantly more time with their groups (p=0.04). Note: this chart is not presented in the interest of space.

Overall, these results suggest that features generated by coarse-grained location data can be strong indicators of stress levels. We use these insights to build our solution.

6. SYSTEM OVERVIEW

We designed *StressAssess* based on our feasibility study findings. Figure 3 shows an overview of the entire system. It comprises of five main components: (1) Passive Location Tracking sub-system: this is a pre-existing component that tracks all devices (when connected to campus WiFi) every 5 seconds from WiFi signal strengths measured by the WiFi access points, (2) Group Detection sub-system: this is a pre-existing component that identifies groups by looking at correlations in devices' movement trajectories, (3) Activity Mapper: a new component that applies *in-situ* heuristics to location and group data to map activities that are commonly engaged by our studied population, and take place in designated spaces in the environment, (4) Feature Extractor: a new component that generates statistical and temporal

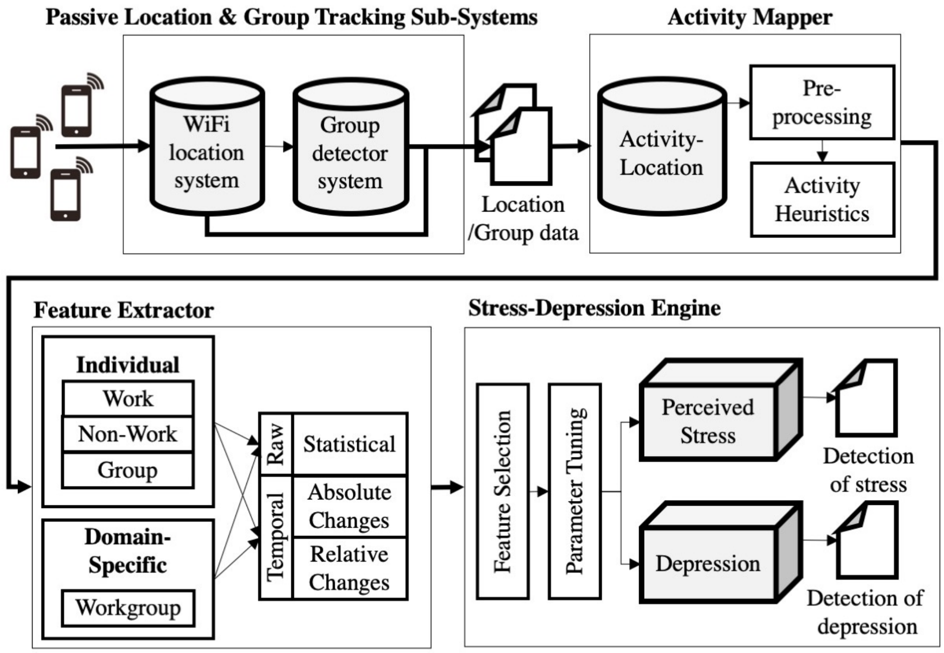


Figure 3: *StressAssess* system overview.

change from mapped location traces that are representative of routines and group interactions as mobility features, and (5) Stress-Depression Engine: a new component that systematically builds detection models for stress and depression and outputs predictions.

6.1 Passive Location & Group Tracking Sub-Systems

StressAssess adopts a passive WiFi-based localization system similar to the one described in [redacted][1]. It uses data collected directly from WiFi access points without requiring the installation of client software. For our study, location traces of non-personal devices were filtered, leaving only traces generated by mobile phones. The accuracy of the system is between 6 to 8 meters in most places – sufficient to localize a device to a specific room or location (refer to [redacted] for the detailed explanation, algorithm and validation results). Additionally, the system anonymises MAC addresses of all connected devices using a 1-way hash function. Hence, our study participants provided their mobile phone MAC address so we could apply the same hash function and identify their devices from these location traces. Note: this step is to evaluate the validity of our stress assess mechanisms - in real world operation, we do not expect to collect the MAC addresses, unless users consent.

Our second sub-system is a group detector, which extracts group information from the localization system. Specifically, the group detector processes location information to cluster devices located in the same vicinity and move together using similar techniques described in Grumon [36]. We defined group sizes (Table 3) with insights from Jayarajah *et al.*[19]. Notably, the "small" group size is similar to a normal project group as student groups are rarely more than five-members strong.

Size	No. of devices	Interaction Type
Solo	1	Alone
Small	2 <= Device => 5	With close/work group
Medium	6 <= Device => 20	With medium-sized group
Large	Device > 20	Mass participation

Table 3: Group sizes are defined to extract interaction patterns on campus.

6.2 Activity Mapper

Next, we assigned various types of activities to the location data using two different two conditions; 1) based on the activity most associated with each location (for that day and time. e.g. seminar room is marked as "Work" during office hours), and 2) if the participant was at that location for at least 15 minutes. Otherwise the activity is marked as *in transition*. In the process, we interpolate missing values using AKIMA interpolation [4]. Note that many of these locations may be multipurpose open spaces and a location can be mapped to different activities depending on the specific days and time of the day.

6.3 Feature Extractor

Figure 4 summarises the features extracted. We identified four main types of campus activities, primarily "Work", "Workgroup", "Non-work", and "Group", which are common across all students."Work" refers to events in locations designated for studying, including seminar rooms for lectures, and group study rooms for workgroup-related events.

"Workgroup" is thus a subset of 'Work'. Workgroup-related events were also verified with $Study_{SE}$ participants' project schedules. "Non-work" refers to events in recreational spaces such as the campus gym, dance studios, and food courts. Finally, "Group" events occur in locations where multiple devices were found to be connected together (using the logic of the group detector). These activities can be categorised into *individual* and *domain-specific* sets and characteristics such as activity frequency, duration, averages and proportions of different activities were extracted as ***raw*** features.

	Feature Name	Description
Individual	campus_totalTime	Total time spent on campus daily
	transitions_count/time	No./amount of time in transition
	locations_num	No. of unique locations visited daily
	buiding_num	No. of unique school buildings visited daily
	<**work**Type>_count/time	No./amount of time spent on `Work' activity - seminar, consultation, study
	<**nonwork**Type>_count/ time	No./amount of time spent on `Non-work' activity - gym, cca, having meals
	<**group**Type>_count/time	No./amount of time spent with different groups - Solo (1), small (<=5), medium (<=20), large (>20)
Domain-Specific	**Workgroup**_totalTime	Total time spent on SE-related events
	<**Workgroup**>_count/time	No./amount of time spent on various SE activities - Pair programming, meetings, knowledge sharing, project management/milestones
	WorkgroupMember_count	No. of time spent in groups with SE member

Figure 4: *StressAssess* List of features grouped in sets.

Change Features

We hypothesized that ***changes in a person's movement patterns and interaction habits in reference to themselves are key indicators of perceived stress***. This was based on prior research that showed that changes in behavior could occur due to stress [5, 32, 38], and that struggles to reorient or adapt could bring about serious consequences [35]. An individual's behavior is compared against their own from an earlier period, $x^u_{i,j}$, as ***absolute change (*****abs** (Equation 1)). Additionally, we compare how an individual is changing against their peer population, $x^o_{i,j}$, as ***relative change (*****rel** (Equation 2)). These change values are calculated over different time windows in multiples of 3 days (as our ground truth data was collected every 3 days). In summary, we extracted approximately 250 ***raw***, ***abs*** and ***rel*** features.

$$x^o_{i,j} = \sum_{V \in [1..N] \setminus u} x^u_{i,j} / N - 1$$

$$\hat{x}^*_{i,j} = \sum_{k=i}^{i+w} x^*_{i,j} \quad i \in [1..K-w], [w] := \{3, 6, 9, \ldots, w\}$$

$$\boldsymbol{abs}^u_{\mathbf{i,j}} = \hat{x}^u_{i+1,j} - \hat{x}^u_{i,j} \tag{1}$$

$$\boldsymbol{rel}^u_{\mathbf{i,j}} = (\hat{x}^u_{i+1,j} - \hat{x}^o_{i+1,j}) - (\hat{x}^u_{i,j} - \hat{x}^o_{i,j}) \tag{2}$$

Model	Data Sample/split	Final Feature Settings
$Models_{S\_SE}$	$Study_{SE}$ (by user) 80& train, 20% test	Time window: 6 day Type: Change features Set: Individual+Domain
$Models_S$	$Study_{SE}$ (by user) 80& train, 20% test	Time window: 6 day Type: Change features Set: Individual
$Models_{S\_IS}$	$Study_{SE}$,$Study_{MIX}$ IS-students only (stratified split) 80& train, 20% test	Time window: 6 day Type: Change features Set: Individual

Table 4: Summary of data sample and split for training to achieve our final feature settings for each model version.

6.4 Stress-Depression Analysis Engine

At its core, *StressAssess* uses a standard machine learning pipeline of feature selection and classification. This includes a recursive feature elimination (RFE) process to search through 250 different features for the most optimal ones. In deciding the final model, we compared the following three classifiers (as suggested by similar prior work [10]): Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF). As described in Section 7, the classifier was re-calibrated to correct for lower predicted probabilities for the positive class.

7. STRESS ASSESSMENT

We present findings on evaluating three different stress models, namely (1) $Models_{S\_SE}$, a highly specific model which uses Domain-specific features,(2) $Models_S$, which excludes Domain-specific features, and (3) $Models_{S\_IS}$, a trade-off on generalizability as it considers demographic profile.

7.1 Experiment Setup

Table 4 lists details of how our models were set up. Selection of various model settings were determined through a standard procedure of 10-fold and Group K-fold cross validation. As explained in Section 4.4, we applied SMOTE to correct the highly imbalanced training dataset. Correspondingly, samples from $Study_{MIX}$ and $Study_{ScE}$ were used to externally validate the final model.

Performance Metrics

Performance metrics were defined to cater for class imbalance in our test/validation sets, mainly using *sensitivity* (true positive rate), *specificity* (true negative rate), *misclassification rate*, and *balanced accuracy*. In an ideal scenario, our model should achieve both high sensitivity and specificity to correctly detect *severe stress* instances. Balanced accuracy is the based on recall of both classes, proportionate to class distribution, fit for anomaly detection, such as ours [23].

7.2 Cross Validation

Table 5 summarizes results of our cross validation at different stages.

	(***raw***)	(***rel+abs***)	(***raw+rel+abs***)
10-fold CV			
Sensitivity	60.56	**64.32**	63.77
Specificity	84.95	**85.56**	82.23
Misclassification	23.83	**22.08**	23.78
Bl. Accuracy	72.75	**74.94**	73.50
Group 5-fold CV			
Sensitivity	58.07	**62.27**	62.27
Specificity	35.88	**34.61**	33.92
Misclassification	62.37	**63.11**	63.70
Bl. Accuracy	46.97	**48.44**	48.10

(a) LR classifier with different feature sets, ***raw***, ***rel+ abs***, ***raw+ rel+ abs***

Feature Name	Set	Type	varImp
study_count	Individual	***abs***	100
building_num	Individual	***abs***	39.78
campus_totalTime	Individual	***abs***	39.58
med_count	Individual	***rel***	29.48
transitions_count	Individual	***abs***	23.09
group_time	Individual	***rel***	22.75
small_count	Individual	***rel***	18.41
locations_num	Individual	***abs***	16.76
count groups	Individual	***rel***	14.75
SEmember_count	Domain-specific	***rel***	13.86

(b) List of features in order of variable importance.

	LR	SVM	**RF**
10-fold CV			
Sensitivity	64.32	62.05	**94.60**
Specificity	85.56	88.03	**99.12**
Misclassification	22.08	21.32	**02.51**
Bl. Accuracy	74.94	75.04	**96.86**
Group 5-fold CV			
Sensitivity	62.27	69.27	**98.02**
Specificity	34.61	70.02	**54.02**
Misclassification	63.11	30.15	**41.26**
Bl. Accuracy	48.44	69.65	**76.02**

(c) Random Forest yielded high sensitivity in both CV.

	3	**6**	9	12
10-fold CV				
Sensitivity	94.60	**94.29**	88.11	86.23
Specificity	99.12	**98.90**	99.34	99.30
Misclassification	02.51	**02.76**	04.70	05.41
Bl. Accuracy	96.86	**96.59**	93.72	92.76
Group 5-fold CV				
Sensitivity	98.02	**97.92**	93.91	91.99
Specificity	54.02	**69.21**	76.62	74.44
Misclassification	41.26	**27.93**	21.70	24.15
Bl. Accuracy	76.02	**83.56**	85.26	83.22

(d) Change features calculated incrementally at 3 days interval.

Table 5: CV yields best settings with change features (***rel+abs***) calculated at 6 days interval using Random Forest classifier

7.2.1 *Feature Set Experiment*

We first checked if change features (Section 6.3) make stronger predictors of stress. Using Logistic Regression (LR) as our base classifier, we achieved the best results using only change features (***rel+abs***) (Table 5a). Table 5b lists the ten most important features (and their sets), in order of variable importance. This experiment led us to only use change features, moving forward.

7.2.2 *Choice of Classifier*

Next, we sought to investigate which classifier, Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) works best. As shown in Table 5c, RF performed significantly better ($p<0.001$) than the rest and was chosen. As part of the feature selection process, we used recursive feature elimination (RFE), with a backward elimination with a step size 1, on all change features. However, no change features were completely redundant and the performance of the RF classifier peaked with all change features considered.

7.2.3 *Time Window Experiment*

Finally, we determined the time window, in increments of 3 days (corresponding to the frequency of PSS-4 samples), that provides the best stress detection. From Table 5d, we observed that sensitivity remained high (>90%), for both cross validations, using 3 or 6 day intervals. We selected 6 day intervals, compared to 3, as the misclassification rate was significantly lower for 5-fold CV.

7.3 Validation Results

Overall, we used a Random Forest with change type features calculated at 6 day interval as our main stress model, $Model_{S\_SE}$. Table 6a summarizes our test and validation results of the main stress model, $Model_{S\_SE}$ and the variant $Model_{S}$. Recall (Section 7), the only difference between them is the exclusion of Domain-specific features, which are highly tailored to SE-students in the $Study_{SE}$ sample. Generalizing the model to exclude Domain-specific features also allows for validation on other samples.

The $Model_{S\_SE}$ had an ideal 100% specificity when used on the $Study_{SE}$ test set, although it had a 22.22% misclassification rate (72 false positives out of 321 negative cases (*normal stress*)). Notably, the removal of Domain-specific features did not significantly affect the results on the $Study_{SE}$ test set. Specifically, the false positives slightly increased with 5 more instances (77) but this did not affect the detection of *severe stress*.

Our solution yielded high sensitivity, correctly detecting the 1 *severe stress* instance for $Study_{MIX}$, although the specificity decreased to 63.35% (explained in Section 7.5). Despite the increase in misclassification rate, the balanced accuracy remained above 80%. The sensitivity reduced to 66.67% in $Study_{ScE}$ with 1 out of 3 *severe stress* instances misclassified. While more *normal stress* instances were correctly detected, the balanced accuracy ultimately dropped to 77.91%.

7.4 Major-Specific Stress Model

One way to demonstrate scalability is to build an effective stress model using just the major of study. That is, the model can be used effectively on other students taking the same major, regardless of quantity of samples, year of study or exact courses taken. To test this, we built a major-specific model, $Model_{S\_IS}$, for all IS-major students using data from $Study_{SE}$ and $Study_{MIX}$. This collection resulted in 2366 *normal stress* and 145 *severe stress* labels from 93 students.

As shown in Table 6b, our test results were less satisfactory with only 65.52% sensitivity as a result of 10 (out of 29) *severe stress* instances being misclassified. Upon investigation, we found that the stratification method produced a split of only 22 $Study_{SE}$ participants used for training. Further, only 3 out of the 22 students were reportedly under *severe stress*. Because stratification kept classes equally balanced in both training and test sets, the result of SMOTE sampling was largely based on these 3 participants, and thus, the drastic reduction of *severe stress* labels (originally from 50 participants) led to poorer detection results.

7.5 Error Analysis

Testing $Model_S$ on $Study_{MIX}$ revealed 346 misclassified out of 944 *normal stress* labels. We sought to understand this – Specifically, these misclassifications were as a result of 13 participants. Interestingly, these participants reported *critical depression* from their PHQ-8 assessment, but their PSS-4 scores were not high enough (over 12) to be recognised as *severe stress*. Removing these cases reduced the false positives to 26.8%. In an operational setting, we would want *StressAssess* to detect and respond as early as possible to cases of high stress; However, as we quickly realised, students may not necessarily perceive themselves as being under severe stress, even when they were feeling severely depressed.

A manual error analysis on $Model_{S\_IS}$ revealed more concerning observations. More than half our training sample reported experiencing *critical depression* (these were different participants from the above 13), but made none to few reports of *severe stress*. To understand this issue, we checked if all (93) participants experienced anything more concerning than “severe stress” by cross-examining the PHQ-8 and interview reports in detail. We found that among the IS students, 9 reported *severe stress*, 28 reported *critical depression*, 9 reported *severe stress* and *critical depression*, and the rest reported nothing. This large amount of *critical depression* without a corresponding *severe stress* indication reaffirmed our decision to extend the problem assessment to also detect *critical depression*.

8. DEPRESSION ASSESSMENT

We realised, from the previous section, that more students were reporting *critical depression* than *severe stress*, which was affecting the performance of our stress detection. We now show how we modified our stress model to also detect depression – $Model_{D\_IS}$.

8.1 Stress Model on Depression

We first tried to detect *critical depression* using $Model_{S\_IS}$. However, this had a low sensitivity of 54.55%. Digging deeper, we found that students only reported *critical depression* (PHQ-8 score above 15) after 39 days on average, with the earliest being after 12 days. This matches clinical wisdom as, typically, people develop quicker responses to stressful experiences, whereas depressive episodes often intensify over longer time periods. Thus, we hypothesized that a longer time interval ($Model_{S\_IS}$ uses 6 day intervals) was needed to accurately detect depression.

8.2 Cross Validation: Effect of Time Window

Subsequently, we repeated the time window experiment to determine an appropriate interval for detecting depression. We maintained a 3 day window size to match the original collection of PSS-4 labels, and used windows sizes ranging from 3 up to 30 days in length – as shown in Table 7. We found that the Random Forest classification algorithm using change features and a 15 day interval yielded highest sensitivity on both cross validations methods. We call this model $Model_{D\_IS}$.

8.3 Validation Results

We validated $Model_{D\_IS}$ on the IS-major test set and achieved a 74.75% sensitivity. That is, 25 out of 99 *critical depression* instances remained misclassified as *non-critical depression*. We expanded our analysis to review the profiles of our students such as their gender, academic year, GPA, and Big-5 personality test [20] filled out as part of the demographics survey. Our training set has an approximately fair distribution of Freshmen, Sophomores, and Juniors and we noticed that a high percentage of students (90% of sample) had low scores on Extraversion and high scores on Agreeableness and Neuroticism (more than 60%) on their Big-5.

8.3.1 *Personality Profile*

We thus decided to add Big-5 factors as additional features in our $Model_{D\_IS+}$ model. Previous studies have established Big-5 as compelling measures of behavior in real situations [15]. However, traits can still vary from moment-to-moment and become a potential limitation to our implementation. Instead of using raw Big-5 scores, we compared the scores of each factor in the personality test to their average on all factors to generate trait standings; e.g. high on Extraversion is when the average Extraversion score is higher than the total average [15]. This addition resulted in high sensitivity of 91.92% with 8 *critical depression* instances missed from being detected. The misclassification

	Models$_{S_SE}$	***Models$_S$***		
Population	***Study$_{SE}$***	***Study$_{SE}$***	***Study$_{MIX}$***	***Study$_{ScE}$***
Sensitivity	100	100	100	66.67
Specificity	77.57	76.01	63.35	89.15
Misclassification	22.22	23.77	36.61	11.36
Bl. Accuracy	88.79	88.01	81.67	77.91
Confusion Matrix	(A)	(B)	(C)	(D)

(A)	**0**	**1**
0	249	0
1	72	3

(B)	**0**	**1**
0	244	0
1	77	3

(C)	**0**	**1**
0	598	0
1	346	1

(D)	**0**	**1**
0	115	1
1	14	2

(a) Test results on stress model. *Models$_{S_SE}$* includes Domain-specific features, and tested on *Study$_{SE}$*.*Models$_S$* excludes Domain-specific features, and tested *Study$_{SE}$*, *Study$_{MIX}$*, and *Study$_{ScE}$* respectively.

	Models$_{S_IS}$		
	10-CV	Grp-CV	Test
Sensitivity	95.21	87.25	65.52
Specificity	98.06	72.29	79.49
Misclassification	2.97	26.42	21.31
Bl. Accuracy	96.64	84.48	72.50
Confusion Matrix	-	-	(E)

(E)	**0**	**1**
0	376	10
1	97	19

(b) CV and test results of our major-specific model, *Models$_{S_IS}$*, which also excludes Domain-specific features. Model is built and tested on strictly IS-major students from two different studies, *Study$_{SE}$* and *Study$_{MIX}$*.

Table 6: Validation results of 3 different stress models.

	3	6	9	12	**15**	18	21	24	27	30
10-fold CV										
Sensitivity(%)	72.92	79.58	79.85	83.12	**86.39**	79.34	77.95	74.30	72.04	69.25
Specificity(%)	80.23	79.84	82.00	86.15	**89.29**	89.80	92.06	91.69	93.45	93.70
Misclassification(%)	23.42	20.27	19.07	15.36	**12.15**	15.43	14.99	17.00	17.26	18.52
Bl. Accuracy(%)	76.58	79.71	80.92	84.64	**87.84**	84.57	85.01	83.00	82.74	81.48
Group 5-fold CV										
Sensitivity(%)	48.15	50.37	50.54	53.46	**54.79**	54.96	51.42	50.92	47.97	45.43
Specificity(%)	61.94	59.56	62.31	64.86	**69.30**	70.99	74.04	75.41	78.07	79.84
Misclassification(%)	40.69	41.97	40.05	37.45	**33.65**	32.04	29.75	28.90	27.19	26.16
Bl. Accuracy(%)	58.84	58.53	60.03	63.15	**66.05**	66.82	66.15	66.72	66.24	65.41

Table 7: CV results on Random Forest classifier with ***chg - rel+ abs*** to detect depression on different time interval yielded best sensitivity at 15 days.

	Model$_{D_IS}$	*Model$_{D_IS+}$*
Sensitivity	74.75	91.92
Specificity	80.15	90.32
Misclassification	20.92	09.36
Bl. Accuracy	77.45	91.12
Confusion Matrix	(F)	(G)

(F)	**0**	**1**
0	323	25
1	80	74

(G)	**0**	**1**
0	364	8
1	39	91

Table 8: Test results on depression model, *Model$_{D_IS}$* and *Model$_{D_IS+}$*, which includes personality traits as additional features. Both models were tested on IS-major students.

rate dropped to 9.36%, and the balanced accuracy yielded 91.12%.

9. DISCUSSION

9.1 Real-Time Fine-Grained Detection vs. Post Coarse-Grained Analysis

Many *mHealth* solutions offer real-time analysis to support just-in-time interventions [34]. We concede that our solution is based on post-analysis; *StressAssess* detects *severe stress* and *critical depression* at 6 and 15 day intervals, respectively. However, we believe that careful monitoring must be in place for a reasonable period of time before making decisions to intervene. Without downplaying the significance of stress/depression, these responses can be common experiences. It is when stress/depression are prolonged or become unmanaged that they are "red flagged". Second, using coarse-grained information limits our system from identifying issues that more precise and personalized information would be able to provide. In addition, more precise information is needed before diagnosis can be made [16]. However, our primary goal is to create a first level safety net to "filter" highly concerning cases *at scale* – individual users can decide to go "all-in", at which time solutions that use fine-grained data from multiple sensors with app support can be used to offer personalized interventions.

9.2 Early Identification and Intervention

We were quite careful to only address the problem of detecting stress and depression. The next challenge is to provide appropriate help to individuals with *severe stress* and *critical depression* before they reach more serious consequences; stress and depression being different effects that require different interventions. For example, Figure 5 charts students with *severe stress* and *critical depression*. 2 reported depression as early as the first week of semester and continued till the end. These “red flags” are nearly impossible for peers and lecturers to immediately spot, but were detected by *StressAssess*. There is a pressing need for early identification and intervention, though they must be introduced in medically-approved and controlled ways. We are working with medical professionals to develop interventions that complement the early detection capabilities of *StressAssess*.

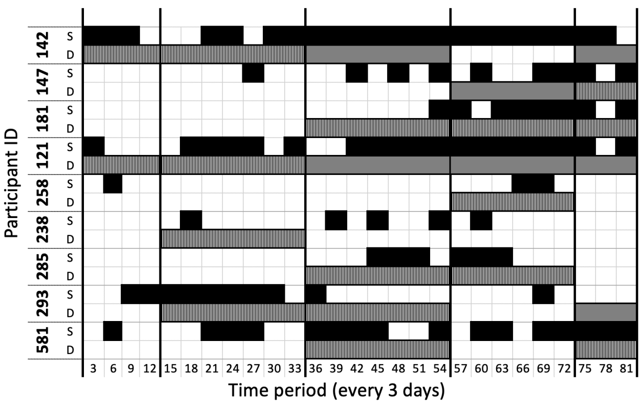


Figure 5: 9 students reported *severe stress* (in black), and *critical depression* (‘moderately severe’ (shaded), ‘severe’ (grey)).

9.3 Privacy Considerations

While our method enables community-wide sensing without any active involvement for users to be monitored, it naturally leads to privacy concerns that data could be used against users, or simply, users being tracked against their will. In the case of our initial studies on a university campus, it was possible to obtain IRB approval to deploy this solution as students are not paid employees, thus detecting students with *severe stress*/*critical depression* has no lasting negative implications. Moreover, we anonymised all location data and student self-reported information. However, if *StressAssess* is deployed at scale to also detect mental health of university employees, the concern becomes that these predictions could be misused in performance evaluations and other review-type processes. This problem remains unclear to us, but we believe our solution must be coupled with appropriate policies protecting employee rights including the exercise of adequate anonymity. We are currently working with experts in privacy and ethics laws to develop appropriate policies and procedures for these types of community wide health sensing systems.

9.4 Limitations of Study

We have shown that temporal changes in an individual’s routine and their group interactions, extracted from coarse-grained location data make useful features in detecting *severe stress* and *critical depression*. While our experiments were tested using three user studies, sampled at different times across different student groups, the demographic of our users does not fully represent the diversity of university students. Second, our models for stress and depression were built on highly imbalanced positive samples, thus requiring oversampling using the well-established SMOTE method. Unfortunately, the exceptionally low occurrence of *severe stress* is reflective of university students, where less than 9% of the population would potentially raise serious concerns from *severe stress* [17]. Finally, we have only deployed and tested *StressAssess* on university students on campus and it’s unclear how it would work in other environments.

10. CONCLUSION

We presented *StressAssess*, a system to detect *severe stress* and *critical depression* episodes in individuals. *StressAssess* was designed to be an easy to deploy solution that can offer community-wide sensing on a large numbers of users without needing to install specific applications or owning specific devices. Our system achieves this by using coarse-grained location data, collected passively from the WiFi infrastructure, to calculate changes in an individual’s routines and group interactions. *StressAssess* uses these changes to detect stress at 6 day intervals, and depression at 15 day intervals. We described the various mechanisms used by *StressAssess* and showed, via three rounds of semester long user studies at an Asian university campus, that it achieves an ideal 100% sensitivity (88.79% balanced accuracy) in detecting stress on $Study_{SE}$ sample using a highly specific stress model, and up to 92% sensitivity (91.12% balanced accuracy) in detecting depression on IS-major students using a more generalised depression model with personality profile as additional features. Currently, we are engaged with the research arm of a premier mental health hospital who see *StressAssess* as an appropriate mechanism to automatically detect highly stressed hospital staff in need of support – the concern being that hospital functions, e.g. the oncology wards, can cause extreme stress, which needs to be managed. However, many administrative and policy issues need to be addressed along with testing and improving the performance of *StressAssess* . We will also be trialing *StressAssess* at 3 other local educational institutions in the following months.

11. REFERENCES

[1] Omitted for double blind review.

[2] 2017 work and well-being survey, 2017.

[3] Aberer, K., Sathe, S., Chakraborty, D., Martinoli, A., Barrenetxea, G., Faltings, B., and Thiele, L. Opensense: open community driven sensing of environment. *Proceedings of the ACM SIGSPATIAL International Workshop on GeoStreaming*, pages 39–42. ACM, 2010.

[4] Akima, H., Gebhardt, A., Petzoldt, T., and Maechler, M. Akima: Interpolation of irregularly spaced data. r package version 0.5-4, 2006.

[5] Baruth, M., Lee, D.-C., Sui, X., Church, T. S., Marcus, B. H., Wilcox, S., and Blair, S. N. Emotional outlook on life predicts increases in physical activity among initially inactive men. *Health Education & Behavior*, 38(2):150–158, 2011.

[6] Batista, G. E., Prati, R. C., and Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29, 2004.

[7] Bromet, E., Andrade, L. H., Hwang, I., Sampson, N. A., Alonso, J., De Girolamo, G., De Graaf, R., Demyttenaere, K., Hu, C., Iwata, N., et al. Cross-national epidemiology of dsm-iv major depressive episode. *BMC medicine*, 9(1):90, 2011.

[8] Brown, C., Efstratiou, C., Leontiadis, I., Quercia, D., Mascolo, C., Scott, J., and Key, P. The architecture of innovation: Tracking face-to-face interactions with ubicomp technologies. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 811–822. ACM, 2014.

[9] Burnaev, E., Erofeev, P., and Papanov, A. Influence of resampling on accuracy of imbalanced classification. *Eighth International Conference on Machine Vision (ICMV 2015)*, volume 9875, page 987521. International Society for Optics and Photonics, 2015.

[10] Canzian, L. and Musolesi, M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1293–1304. ACM, 2015.

[11] Cohen, S., Kamarck, T., Mermelstein, R., et al. Perceived stress scale. *Measuring stress: A guide for health and social scientists*, 1994.

[12] Cox, T. *Stress research and stress management: Putting theory to work*, volume 61. Hse Books Sudbury, 1993.

[13] Cox, T., Griffiths, A., and Rial-González, E. *Research on work-related stress*. European Communities, 2000.

[14] Eisenberg, D., Downs, M. F., Golberstein, E., and Zivin, K. Stigma and help seeking for mental health among college students. *Medical Care Research and Review*, 66(5):522–541, 2009.

[15] Fleeson, W. and Gallagher, P. The implications of big five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of personality and social psychology*, 97(6):1097, 2009.

[16] Harrington, R. and Clark, A. Prevention and early intervention for depression in adolescence and early adult life. *European archives of psychiatry and clinical neuroscience*, 248(1):32–45, 1998.

[17] Hefner, J. and Eisenberg, D. Social support and mental health among college students. *American Journal of Orthopsychiatry*, 79(4):491, 2009.

[18] Jaques, N., Taylor, S., Azaria, A., Ghandeharioun, A., Sano, A., and Picard, R. Predicting students' happiness from physiology, phone, mobility, and behavioral data. *Affective computing and intelligent interaction (ACII), 2015 international conference on*, pages 222–228. IEEE, 2015.

[19] Jayarajah, K., Lee, Y., Misra, A., and Balan, R. K. Need accurate user behaviour?: pay attention to groups! *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 855–866. ACM, 2015.

[20] John, O. P., Donahue, E. M., and Kentle, R. L. The big five inventory—versions 4a and 54, 1991.

[21] Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., and Mokdad, A. H. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1):163–173, 2009.

[22] Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9), 2010.

[23] Li, S. and Tsang, I. W. Maximum margin/volume outlier detection. *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*, pages 385–392. IEEE, 2011.

[24] Mowbray, C. T., Megivern, D., Mandiberg, J. M., Strauss, S., Stein, C. H., Collins, K., Kopels, S., Curlin, C., and Lett, R. Campus mental health services: recommendations for change. *American Journal of Orthopsychiatry*, 76(2):226, 2006.

[25] OctoberCMS. October cms: Simple yet flexible, 2016.

[26] Organization, I. L. Workplace stress: A collective challenge, 2016.

[27] Organization, W. H. et al. Promoting mental health: Concepts, emerging evidence, practice: Summary report. 2004.

[28] Pantelopoulos, A. and Bourbakis, N. G. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1):1–12, 2010.

[29] Pariat, M., Rynjah, M., Joplin, M., and Kharjana, M. Stress levels of college students: interrelationship between stressors and coping strategies. *IOSR J. Humanit. Soc. Sci*, 19:40–45, 2014.

[30] Qualtrics. Qualtrics, 2017.

[31] Rahe, R. H. Stressful life events—their nature and effects. *Psychosomatic Medicine*, 39(1):64–65, 1977.

[32] Roshanaei-Moghaddam, B., Katon, W. J., and Russo, J. The longitudinal effects of depression on

physical activity. *General hospital psychiatry*, 31(4):306–315, 2009.

[33] Sacco, P., Bucholz, K. K., and Harrington, D. Gender differences in stressful life events, social support, perceived stress, and alcohol use among older adults: results from a national survey. *Substance use & misuse*, 49(4):456–465, 2014.

[34] Sarker, H., Tyburski, M., Rahman, M. M., Hovsepian, K., Sharmin, M., Epstein, D. H., Preston, K. L., Furr-Holden, C. D., Milam, A., Nahum-Shani, I., et al. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4489–4501. ACM, 2016.

[35] Schabracq, M. J. and Cooper, C. L. The changing nature of work and stress. *Journal of Managerial Psychology*, 15(3):227–241, 2000.

[36] Sen, R., Lee, Y., Jayarajah, K., Misra, A., and Balan, R. K. Grumon: Fast and accurate group monitoring for heterogeneous urban spaces. *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, pages 46–60. ACM, 2014.

[37] Tan, J. B. and Yates, S. Academic expectations as sources of stress in asian students. *Social Psychology of Education*, 14(3):389–407, 2011.

[38] Van Uffelen, J. G., van Gellecum, Y. R., Burton, N. W., Peeters, G., Heesch, K. C., and Brown, W. J. Sitting-time, physical activity, and depressive symptoms in mid-aged women. *American journal of preventive medicine*, 45(3):276–281, 2013.

[39] Wang, R., Aung, M. S., Abdullah, S., Brian, R., Campbell, A. T., Choudhury, T., Hauser, M., Kane, J., Merrill, M., Scherer, E. A., et al. Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 886–897. ACM, 2016.

[40] Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., and Campbell, A. T. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14. ACM, 2014.

[41] Wang, R., Wang, W., daSilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., and Campbell, A. T. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):43, 2018.

[42] Warttig, S. L., Forshaw, M. J., South, J., and White, A. K. New, normative, english-sample data for the short form perceived stress scale (pss-4). *Journal of health psychology*, 18(12):1617–1628, 2013.

[43] Weiss, R. S. and Kahn, R. L. Definitions of work and occupation. *Soc. Probs.*, 8:142, 1960.

[44] Zhang, D., Guo, B., and Yu, Z. The emergence of social and community intelligence. *Computer*, 44(7):21–28, 2011.

[45] Zhou, M., Ma, M., Zhang, Y., SuiA, K., Pei, D., and Moscibroda, T. Edum: classroom education measurements via large-scale wifi networks. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 316–327. ACM, 2016.