Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

11-2019

StressMon: Scalable detection of perceived stress and depression using passive sensing of changes in work routines and group interactions

Nur Camellia Binte ZAKARIA
Singapore Management University, ncamelliaz.2014@phdis.smu.edu.sg

Rajesh BALAN
Singapore Management University, rajesh@smu.edu.sg

Youngki LEE
Singapore Management University, YOUNGKILEE@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Digital Communications and Networking Commons, Numerical Analysis and Scientific Computing Commons, and the Software Engineering Commons

Citation

ZAKARIA, Nur Camellia Binte; BALAN, Rajesh; and LEE, Youngki. StressMon: Scalable detection of perceived stress and depression using passive sensing of changes in work routines and group interactions. (2019). *Proceedings of the ACM on Human-Computer Interaction*. 3, 37:1-29.
Available at: https://ink.library.smu.edu.sg/sis_research/4862

This Journal Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

StressMon: Scalable Detection of Perceived Stress and Depression Using Passive Sensing of Changes in Work Routines and Group Interactions

CAMELLIA ZAKARIA, Singapore Management University, Singapore
RAJESH BALAN*, Singapore Management University, Singapore
YOUNGKI LEE*, Seoul National University, South Korea

Stress and depression are a common affliction in all walks of life. When left unmanaged, stress can inhibit productivity or cause depression. Depression can occur independently of stress. There has been a sharp rise in mobile health initiatives to monitor stress and depression. However, these initiatives usually require users to install dedicated apps or multiple sensors, making such solutions hard to scale. Moreover, they emphasise sensing individual factors and overlook social interactions, which plays a significant role in influencing stress and depression while being a part of a social system. We present *StressMon*, a stress and depression detection system that leverages single-attribute location data, passively sensed from the WiFi infrastructure. Using the location data, it extracts a detailed set of movement, and physical group interaction pattern features without requiring explicit user actions or software installation on client devices. These features are used in two different machine learning models to detect stress and depression. To validate *StressMon*, we conducted three different longitudinal studies at a university with different groups of students, totalling up to 108 participants. Our evaluation demonstrated *StressMon* detecting severely stressed students with a 96.01% True Positive Rate (TPR), an 80.76% True Negative Rate (TNR), and a 0.97 area under the ROC curve (AUC) score (a score of 1 indicates a perfect binary classifier) using a 6-day prediction window. In addition, *StressMon* was able to detect depression at 91.21% TPR, 66.71% TNR, and 0.88 AUC using a 15-day window. We end by discussing how *StressMon* can expand CSCW research, especially in areas involving collaborative practices for mental health management.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing devices**; **Mobile computing**; **Empirical studies in ubiquitous and mobile computing**; • **Applied computing** → *Health care information systems.*

Additional Key Words and Phrases: stress, depression, small-group, mobility patterns, Wi-Fi indoor localisation

ACM Reference Format:
Camellia Zakaria, Rajesh Balan, and Youngki Lee. 2019. StressMon: Scalable Detection of Perceived Stress and Depression Using Passive Sensing of Changes in Work Routines and Group Interactions. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 37 (November 2019), 29 pages. https://doi.org/10.1145/3359139

*Both authors contributed equally to this research.

Authors' addresses: Camellia Zakaria, Singapore Management University, Singapore, Singapore, nurcamelliaz@smu.edu.sg; Rajesh Balan, Singapore Management University, Singapore, Singapore, rajesh@smu.edu.sg; Youngki Lee, Seoul National University, Seoul, South Korea, youngkilee@snu.ac.kr.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2019 Association for Computing Machinery.
2573-0142/2019/11-ART37 $15.00
https://doi.org/10.1145/3359139

RIGHTSLINK

1 INTRODUCTION

Severe stress and depression are mental health ailments that have been steadily rising across the world [21, 50]. Often associated because of striking similarities in symptoms, severe stress can take a toll on a person's productivity and can result in depression if the stress is left unmanaged. However, depression can occur without any apparent symptoms of stress. Recently, depression is reported by the World Health Organisation (WHO) as one of the leading causes of lost economic productivity, estimated to cost the global economy US $1 trillion each year [49] due to reasons such as absenteeism [26]. It is shown that treating symptoms of depression can reduce work absenteeism and lost productivity [49].

Prior studies have shown that recovering from stress to a healthy state is much easier, in terms of the length of time and treatment required, compared to depression [24]. Further, treating depression early can result in earlier relative recovery times. Thus, there is great merit in detecting individuals' stress and depression early. It is important to note that while our research aims to detect stressed and depressed individuals in a work setting, it does not reveal the underlying reasons for such conditions.

In workplaces, the underlying mechanisms causing stress are much more complex to understand as individuals are frequently working in groups where group dynamics and social interactions can influence the stress levels of group members [13]. For example, being in supportive groups allows individuals to receive peer support, which dramatically reduces their stress levels [28].

There has been considerable research on assessing stress and depression of users in workplaces in the psychology, small groups, and systems domains [26, 28, 61, 65, 69]. Specifically focusing on the Systems domain, there have been numerous efforts to assess mental health through mobile and wearable applications [16, 29, 41, 68]. While promising, these approaches exhibit several limitations. First, many approaches require the user to install an application on their device. This requirement makes these resources hard to scale to large groups of user populations and introduces a strong self-bias where only users who are interested in getting help would receive it. For example, studies have shown that the highly stressed people who are most at risk tend to avoid activities that are not considered "critical" [3, 38]. Further, these apps pose high privacy risk as they collect and analyse a rich set of personal data (e.g., fine-grained location, conversation, sleep patterns, app usage) [52]. The design of these apps usually require semi-frequent user input and drain an excessive amount of battery power to collect and transmit the necessary data – all these factors reduce the willingness to participate in such studies. More importantly, existing methods emphasise sensing individual factors and overlook the *physical social interactions that commonly influence the stress and depression of individuals engaging in social activities such as workgroups*. For example, on the subject of work, Cox *et al.* characterise stress by both *content* (e.g., task, load) and *context* (e.g., relationship tension among colleagues) [13].

In this paper, we propose *StressMon*, a scalable detection solution, envisioned as an environment-wide "safety net" to automatically and non-intrusively identify individuals exhibiting signs of severe stress or depression in a work setting. Specifically, *StressMon* uses the location information, directly sensed from the WiFi infrastructure, to infers both the individual and their physical group interaction patterns. It employs two pieces of prior work: (1) a WiFi fingerprint-based indoor localisation system [43] that uses server-side localisation to track any connected device, and (2) a group detection system [63] that uses location traces to cluster devices together, even in crowded spaces, into logical groups. Our solution complements prior systems in the following ways:

(1) The only input is single-attribute location information, which is significantly different from prior systems utilising multiple fine-grained user inputs, such as accelerometer, voice, galvanic skin response, to detect stress and depression separately [10, 41, 59].

RIGHTSLINK

(2) To overcome the limitations of using single-dimensional data, we augmented the data with inferred individual routine behaviours and physical group interaction patterns.
(3) Additionally, we ascertained the changes in routines and group interaction patterns by comparing against past periods of an individual and their population as key indicators of stress and depression.
(4) *StressMon* accurately detects both stress and depression using the same location data. Prior work detects only depression using location data [69], while addresses stress detection using multiple sensors [17, 59].

StressMon was evaluated using university students working in groups for class projects, as unmanaged stress can result in negative workgroup outcomes [44]. We conducted three rounds of IRB-approved longitudinal studies with different groups of students and across different periods. The primary study, $Study_{SE}$ (81 days), was purposed to conduct hypothesis testing on the changing behaviours of severely stressed individuals and develop detection models for stress and depression. The other two studies, $Study_{ValidA}$ (36 days) and $Study_{ValidB}$ (81 days), were purposed to validate our models. With a total of 108 students, all participants provided background information about their personality (Big-5 [32]), campus routines, regular assessments of mental states (PSS-4 [12] and PHQ-8 [36]), team schedules, and mobile phone MAC address, enabling us to identify them in our deployed WiFi-based location system. Note: by default, the location system anonymises all MAC addresses. Participants attended two semi-structured interview sessions in the middle and end of each study, in order to verify their primary causes of stress, describe their workgroup experiences, and explain how their stresses were managed. The regular assessments and interview findings were used as ground truth to validate our models. Overall, *StressMon* was able to detect stress using changes in individuals' routine behaviours as the key feature, and, separately, it could detect depression, using changes in group interaction patterns as the key feature. In Section 8, we discuss how *StressMon* can be an enabler for a large-scale monitoring solution to support mental health practices in workgroup settings. Overall, this paper makes the following contributions:

(1) We demonstrate the feasibility of *StressMon* to passively detect stress and depression in individuals using just location data extracted from the WiFi network. In particular, *StressMon* uses location data derived from RSSI values reported by the WiFi access points (APs) and does not require any direct user engagement (through apps, portals).
(2) We show how single-attribute location data can be enhanced to produce a rich set of mobility features for individuals operating in collective workspaces. These features have, to the best of our knowledge, not been explored in prior work. In particular,
    - *StressMon* uses location data to extract the interactions of individuals with their peer groups.
    - *StressMon* uses *temporal changes in behaviours* that compare individuals' normal patterns to past periods of their own (*absolute change*), and their population's (*relative change*).
    - The use of these location-driven features allows *StressMon* to detect stress and depression effectively in practice.
(3) We rigorously evaluated our detection models across three different user studies – where training data was strictly separated from test users. Our model detects individuals with severe stress every 6-days at an Area Under the Curve (AUC) score between 0.91-0.94 for all studies. Our model detects depression every 15-days at 0.88 AUC score on all studies. Note: AUC is a statistical measure of how good a binary predictor is [6] with a score of 1 indicating perfect predictor. Our results are similar or better than prior work in detecting stress that uses much finer-grained data [10, 29, 59] and in detecting depression using just location data [69].

RIGHTSLINK

2 RELATED WORK

Significant research suggests that depression is the most likely outcome of exposure to psychological stress [13, 53, 65]. In investigating the relationship between stress and depression, one may argue that a depressive mood could be as a result of a stressor; thus a compound factor to a stressful situation [62]. However, depression could merely be an affective response, associated with personality characteristics [23, 56]. For this reason, depression can occur in a person without them feeling stressed [24], and stress and depression must be treated as two separate entities [27, 62]. Nonetheless, stress and depression share several similarities. Factors such as high work demand, poor social support and relationships, and limited control over situations are stressors that commonly predict depressive symptoms [65]. As a result, substantial evidence observes severely stressed individuals making changes in their usual behaviours [57, 64]; for example, withdrawing from others and the inability to rest. Some of these behavioural symptoms overlap with depression. The struggle to reorient or adapt can bring about more severe consequences [61]. These findings highlight the importance of recovering early from severe stress to a healthy state, compared to when more severe conditions (i.e., depression) have manifested [24].

Stress and Other Mental Health Monitoring

There has been considerable research in developing stress monitoring mobile applications such as *UStress* [16], *cStress* [29], *StressSense* [41], and *AutoSense* [17], among many others [10, 59]. *Mobilyze* [8] and *Big Black Dog* [15] are examples of smartphone-based applications to detect depression. However, all of these context-aware applications make use of fine-grained sensor data such as electrodermal activity (EDA), electrocardiogram (ECG), and device activity data from wearable sensors and/or smartphones to detect mental conditions in real-time. *StudentLife* [67] from Wang *et al.* analysed behavioural changes related to stress and the same authors also analysed symptoms features to predict depression scores [68]. However, these solutions require installing a custom application which demands much higher user attention (resulting in both fairly low user participation rates and high attrition rates), increases privacy threats, and increases power consumption on their mobile device. Moreover, some of these solutions require specific mobile sensors which would automatically exclude users without those sensors.

Location-enabled Technologies. More recently, researchers have explored the use of location data for mental health monitoring. For example, Canzian *et al.* [9] and Lu *et al.* [41] found a correlation between GPS-based location features and depression. These solutions, unfortunately, cannot be used indoors (for example, inside campus buildings) where GPS is unavailable. Brown *et al.* [7] bridged this gap by using wearable RFID tags to collect indoor location traces of employees interacting with colleagues in different building spaces. However, this technique requires providing custom devices to every individual and thus greatly limits scalability. Ware *et al.* [69] used location data collected from the WiFi infrastructure to detect depression. Similarly, Zhou *et al.* [74] used WiFi indoor localisation data to learn about student behaviour. However, [69] did not consider group behaviours in detecting depression, while [74] neither detected stress nor depression detection.

Large-scale Sensing Solutions

Much research is devoted to developing sensing applications that scale from individuals to entire communities [37]. These applications, however, are mostly in the areas of urban planning [1] and security [73]; for example, using community-wide video surveillance for purposes of public safety. Specific to mental health, Ware *et al.* [69] utilised WiFi association data from the university's infrastructure to detect depression. Our data collection mirrors [69], except we derived group activities from mobile phones connected to the same APs. *StressMon* differs in the following ways:

RIGHTSLINK

(1) it is a full system that works by pulling data directly from the WiFi infrastructure in real-time while prior work used data provided in an offline fashion from campus IT [69], (2) it incorporates features representative of physical group interactions into its models which prior work did not, (3) it maximises the value of single-attribute location data by ascertaining changes in behaviours comparing against past periods of individuals and their population as reliable indicators of stress and depression, and (4) it detects stress and depression using the same sets of features compared to [69] whose approach was to only detect depression.

Overall, *StressMon* is designed to be a first-level safety net providing mental health support for large groups of users, either scholastically or professionally; in that, any user whose device is connected to WiFi in the environment can leverage *StressMon*, without requiring additional device or application. Thus, it nicely complements more fine-grained solutions which require installing active stress trackers for users who desire closer monitoring.

3 SYSTEM OVERVIEW

Figure 1 shows an overview of *StressMon*. It is comprised of three components: (1) Location and Group Tracking Sub-systems, (2) Feature Extractor and (3) Stress-Depression Engine.

3.1 Location and Group Tracking Sub-Systems

We leveraged an existing passive WiFi-based localisation system [43] that uses real-time location services (RTLS) to extract Receiver Signal Strength Information (RSSI). This is the signal strength of each device connected to the AP as measured by the AP. These signal strengths decay as the device moves further away from the AP. Hence, by using RSSI observed by multiple APs, we can compute the position of each device using a method known as *reverse triangulation*. This approach uses data collected solely from the infrastructure (each WiFi AP) and thus can work across any mobile device (e.g. iOS, Android) and does not require installing any client software. The solution we are using has been deployed at several public spaces and with accuracies between 6 to 8 meters in

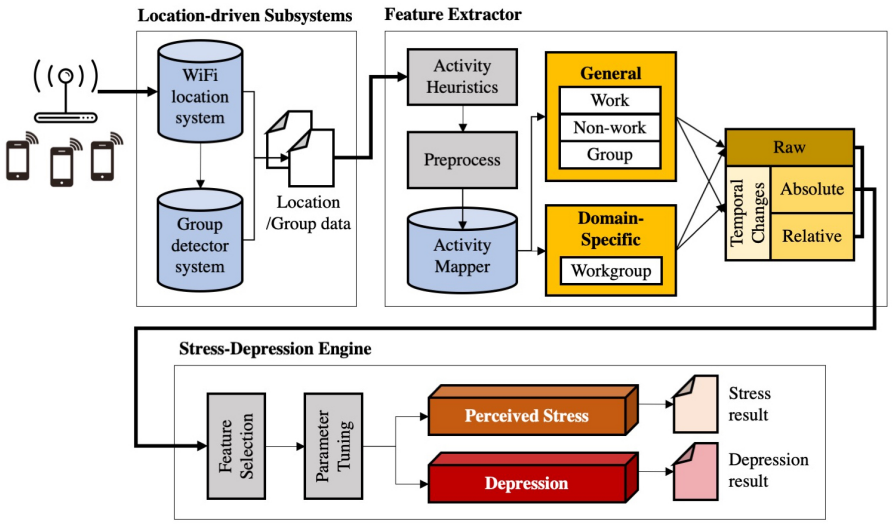


Fig. 1. *StressMon* system consists of three components; existing key location-driven sub-systems, feature extractor and stress-depression detection engine.

RIGHTSLINK

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 1. Group sizes are defined to extract interaction patterns on campus.

most places; sufficient to localise a device to a specific room. Additionally, the system anonymises MAC addresses of all connected devices using a 1-way hash function. Hence, users will need to provide their device MAC address so the same hash function can be applied to identify their devices from these location traces. Note that this step is only to evaluate the validity of our detection mechanisms; in real-world operation, we do not expect to collect the MAC addresses unless users consent. For our user study, only location traces generated by mobile phones were collected.

Next, we utilised an existing group detector, called GruMon [63], which extracts group information from the localisation system. Specifically, the system processes location information to cluster devices located in the same vicinity that move together using the Markov Cluster algorithm (MCL). GruMon was shown to be highly accurate, detecting over 80% of the groups, with 97% precision, within 10 minutes of observing location data. Most of its errors arose from detecting large groups, defined as 7 or more individuals. It was much more accurate in detecting small and medium groups. Note: for this work, we used group sizes (Table 1) based on insights from Jayarajah *et al.* [**?**]. In all our trials, student groups never exceeded five members in size. Thus they fell within the optimal detection capabilities of GruMon.

Overall, *StressMon* was built on top of two existing, accurate, and mature technology solutions. However, both the localisation system and the GruMon group detector can still produce erroneous results that could impact the performance of *StressMon*. In this paper, we did not correct for any errors produced by these components and all results shown include all sources of errors.

WiFi Location Data and Group Data. Each WiFi location entry corresponds to a connection made from the mobile phone to an AP every 5 minutes. Each tuple consists of $[d_i, u_i, l_i, a_i]$, where d is date-time stamp, u is the hashed MAC address of connected devices (representing the users), l is the location code at which the device is localised, a is the accuracy of the localisation and i is the number of entries in the dataset. Each location code, l_i, is mapped to a location name (in the format of <building name>_<level>_<room name>), the location's maximum capacity and current occupancy. Thus, each tuple informs us of the amount of time a user is detected to be at a room-level location and how 'busy' the location is between 97-99% accuracy.

Each group data extends a location entry with $[d_i, g_i, ct_i, ll_i, tt_i, lh_i, s_i]$, where d is datetime stamp, g is a concatenation of hashed MAC addresses connected to the same AP over a period of time, ct is the last datetime the devices were detected as a group, ll is the last location code the devices were detected as a group, tt is the total time detected as a group, lh is the location history which provides a concatenation of location and the detected time, and s is size of group or number of devices concatenated in g. Consequently, group data gives information pertaining to users who make up the group, the various locations and amount of time spent at different locations over a period of time.

3.2 Activity Mapper

The activity mapper assigns the most likely activity to occur at a particular location based on two heuristics; (A) places of activity and time thresholds of students' routines created from demographic

RIGHTSLINK

surveys, and (B) students' average estimation of 15 minutes to *transit* between activities. With (A), we combined our everyday knowledge of how and when different campus spaces are being utilised. For example, lectures are conducted on three fixed time slots. The survey estimates determine time thresholds for students' daily activities such as capping instructor consultation at 1 hour, gym at 2 hours, eating at 30 minutes, otherwise treated as work engagements. With (B), a nomadic device that jumps from one AP to another only considers an activity if the connection lasts for at least 15 minutes, otherwise labelled as 'in transition'. Fortunately, missing location data points were not in long time intervals (no more than a few days), hence, treated with AKIMA interpolation. AKIMA spline affects only the curve of neighbouring data points, minimising the error of its estimates [2].

3.3 Feature Extractor

Our features can be categorised into four broad categories; *Work* (W), *Non-work* (NW), *Group* (G) and *Workgroup* (WG). Work-features are events that take place in locations such as open-study areas, seminar rooms, and group meeting rooms. In contrast, Non-work features are events that take place in locations such as the campus gym, dance studio and cafeterias. Group-features capture properties in the group data. Workgroup-features are Work-features verified against students' project schedule to represent project-specific events. Note that each schedule entry captures location, date, duration, types of task, and attendees. Off-campus and contradicting entries, for example, a detected location which did not match the logged location (for a particular time of the day) were identified as 'unique' task. Except for Workgroup-related features, all features were generated purely with heuristics mentioned in Section 3.2; therefore, considered as a *General* set. Workgroup-features make up a *Domain-specific* set. We extracted these raw features (***raw***) based on:

(1) ***Number of unique visits per day*** records the number of different buildings visited. Our university campus is comprised of seven buildings (five storeys each) and has an underground concourse that connects most buildings.
(2) ***Total time spent on <activity type> & Number of times engaged in <activity type> per day*** consist of the following activities with 15 minutes unit time per activity: campus (W), studying (W), attending lecture (W), group meetings (W), study consultation (W), transiting (NW), eating (NW), exercising (NW), visiting the clinic (NW). Domain-specific workgroup (WG) activities include the types of tasks declared in the project schedule such as pair-programming, knowledge sharing, application design, and milestone preparation and *unique* events.
(3) ***Total time spent being in <group type> & Number of times being in <group type> per day*** consist of the various group types listed in Table 1.

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

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

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

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

RIGHTSLINK

Change Features

We hypothesised that ***changes in a person's movement patterns and interaction habits in reference to themselves are key indicators of perceived stress***. This hypothesis is based on prior research that showed how changes in behaviour occur due to stress [4, 57], and struggles to reorient could bring about serious consequences [61]. An individual's work routine or group interaction behaviour, j, is compared against their own, x^u, or their population, x^o, from an earlier period, $x_{i+1,j} - x_{i,j}$. The interval between periods is calculated in multiples of three days, w, as our ground truth data were collected every three days. The changes calculated against an individual's prior behaviour is denoted by $abs^u_{i,j}$ for absolute change (***abs***). The changes calculated against an individual's population (users who were enrolled in the same course) is denoted by $rel^u_{i,j}$ for relative change (***rel***). In summary, our features consist of *General* and *Domain-specific* ***raw***, ***abs*** and ***rel*** features. We present the top 10 features used in building our stress detection models (Table 6c) and depression detection models (Table 8a) in Section 7 using ROC curve analysis to quantify the diagnostic ability of each feature.

3.4 Stress-Depression Analysis Engine

At its core, *StressMon* uses a standard machine learning pipeline of feature selection and classification. The pipeline includes a recursive feature elimination (RFE) process to use a small subset of ***raw***, ***abs*** and ***rel*** features. We compared the Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) classifiers as similar prior work [9, 18] and are commonly used for binary classification problems [35]. Section 7.1 shows the accuracy of each predictor for stress with the Random Forest algorithm achieving the best performance.

Input and Output. Dataset to build and evaluate our stress-depression detection model is made up of location and group features averaged over three days throughout the whole study. Survey results from each stress assessment (see Section 5.1) were mapped to each feature vector as labels. Similarly, the retrospective depression assessment, which totalled up to 5 surveys, were mapped to each feature vector corresponding to the sampling period. The output of this classifier will be the predicted outcome of single *severe stress* or *depressed* instances.

4 LONGITUDINAL USER STUDIES

To validate *StressMon*, we conducted three longitudinal user studies from 2017 to 2018 using different student populations. In particular, the primary user study was used to build the detection models, while the other two studies were used to validate the models. Colleagues from our behavioural psychology department as well as practising psychiatrists from a local mental health hospital evaluated the entire study procedure. Our studies were approved by our university's Institutional Review Board (IRB).

4.1 Participants

Table 2 summarises the demographics of the three studies. $Study_{SE}$ (primary study) was conducted for 81 days using 76 second-year students who were all enrolled in the Software Engineering (SE) core module. We chose to study the SE module as students often cite it for having a highly stressful group project. Specifically, students work in pre-assigned groups of 5 members balanced across gender, nationality and skills to build a cloud-based web application while following strict team processes. For example, members must spend equal amounts on different types of tasks, pair program, and rotate their programming partner. Thus, students work with people they do not know who have varying abilities, personalities, and work styles.

RIGHTSLINK

	$Study_{SE}$ (primary)	$Study_{ValidA}$ (validation 1)	$Study_{ValidB}$ (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	1. Software Engineering (62)	1. Social Entrepreneurship (11)	1. Software Project Management (8) 2. Interaction Design & Prototyping (13) 3. Computational Thinking (1) 4. Information Systems & Innovation (7) 5. Programme in Writing & Reasoning (6)

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

Two validation studies, $Study_{ValidA}$ (N=13) and $Study_{ValidB}$ (N=51), were conducted for 36 days and 81 days, respectively. Students enrolled in different majors and courses, ranged from Freshmen to Seniors, and formed their groups to work on a mix of semester-long or small projects. None of these courses, unlike SE, require strict team scheduling and complex technical implementations. Managing user retention was challenging in all user studies. At study end, only 62 $Study_{SE}$, 11 $Study_{ValidA}$ and 35 $Study_{ValidB}$ students remained participative. That is, students contributed at least 80% of all survey data and attended at least one interview session. We verified from the interview findings that removed students had not reported a highly stressful semester. None of our participants resided on campus vicinity as university residence is not part of our city-centre campus.

4.2 Procedure

Each participant filled out a pre-study questionnaire outlining their personality traits (Big-5 [32]), current GPA, and regular campus routines (e.g., meal breaks, sports, frequented workspaces). During the study, participants reported their stress levels using PSS-4 [12] survey every three days, and a retrospective assessment for depression using the PHQ-8 [36] questionnaire approximately every two weeks. Note that the practising psychiatrists who evaluated our entire study strongly advised

RIGHTSLINK

Event Description	Period	Event Description	Period
Collect assessment #1	3	Conduct interview #1	39 – 43
M1: Release of proj. specs.	04 - 08	M3: User Acceptance Test (UAT)	53 – 56
Collect assessment #2	15	Collect assessment #4	57
M2: Team Goal	25 - 29	M4: Final deliverable	74-78
Collect assessment #3	36	Collect assessment #5	75
1-Week Recess	39 – 43	Conduct interview #2	77-81

Table 3. Data collection periods (in days) were timed before and after critical SE milestones (shaded rows and indicated as M#).

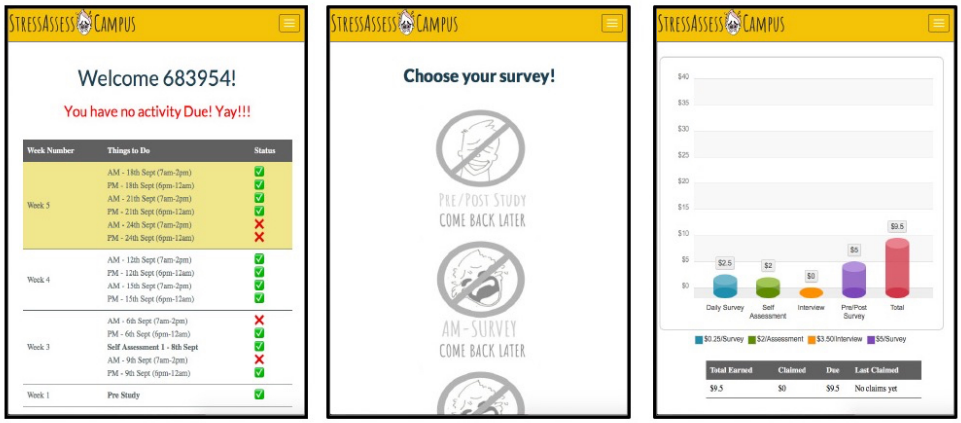


Fig. 2. *StressMon* online portal used for ground truth collection – reporting surveys every three days and retrospective assessments every two weeks. Students were reminded of times to participate and notified of their earnings at the end of each participation.

us to use the PHQ-8 (and not the PHQ-9 version) to avoid the ninth question related to suicidal thoughts, as our research team was not trained to handle a definite answer to that specific question.

To strike a balance between frequency of surveying and reducing user burdens, administering surveys every three days allowed us to collect sufficient samples for every day of the week. Additionally, students attended two semi-structured interview sessions at the midpoint (i.e., 1-week recess, see Table 3) and study end to share about their primary sources of stress, experiences of work-related issues, and ways of managing stressful work situations. We used the survey responses and interview data as ground truth. $Study_{SE}$ participants provided access to their SE project schedule (a graded document maintained by all teams to keep track of project plans). These records included information about meeting dates, duration, and location. Finally, all students provided their mobile phone MAC address so that they could be identified on the WiFi localisation system (see Section 3.1). Note that these data were collected to build and validate our detection models - once *StressMon* is operational, it evaluates stress and depression based on passive location data.

We administered all surveys using Qualtrics [55], embedded in a custom online portal that provides reminders and compensation updates built on October-CMS [48] (see Figure 2). All participants were compensated with a maximum amount of USD 30 in two ways; 1) for entering the study and 2) for remaining participative throughout the study. Participants were also eligible

RIGHTSLINK

for a lucky draw to win a USD 76 cash prize and a USD 37 bonus to participants, provided their entire project group joined and completed the study.

5 USER STUDY DATA PROCESSING

The survey data collected from our user studies was to serve as ground truth for *StressMon*'s detection results. *StressMon* does not use any of the survey data to make its predictions. For this paper, the survey and interview data provided final validation of the severity of student stress levels and the leading causes of that stress.

5.1 Binning Stress and Depression Scores

Self-reports of PSS-4 and PHQ-8, coupled with verification of students' experiences from their interviews constituted to ground truth. Table 3 provides the periods (in days) at which data was collected and corresponded with critical milestone periods for SE. We now describe how assessment scores were grouped. Note that $Study_{ValidA}$ started on Day 45, after the 1-week recess.

Stress, PSS-4. PSS-4 is a well-established scale, ranging from 0 to 16, used among students and employees [58, 70]. PSS-4 accounts for negative and positive typed stresses through reverse scoring. Having a score close to 16 suggests severe amounts of negatively perceived stress [12]. Severe negative stress has been found to result in adverse cognitive and emotional consequences and vulnerability to depression [54]. In our study, PSS-4 scores have the following distribution: min=1, max=16, median=8, mean=7.66, SD=2.35. We divided them into two groups: ***severe stress*** (1, positive class) for the scores of 12 and above, two standard deviations away from the mean, otherwise ***normal stress*** (0, negative class). Since PSS-4 is not designed as a diagnostic tool for severe stress, 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 [70].

Depression, PHQ-8. The use of PHQ-8 is more straightforward as the scale is a diagnostic tool with clear cutoffs – 0-4 (no/minimal depression), 5-9 (mild depression), 10-14 (moderate depression), 15-19 (moderate-severe depression), and 20-24 (severe depression). Based on related work [34, 69], assessments with PHQ-8 score >= 10 are treated as clinically significant depression. Accordingly, scores of 10 and above (min=0, max=24, median=8, mean=8.23, SD=4.77) were grouped as ***depressed*** (1, positive class), otherwise ***non-depressed*** (0, negative class).

5.2 Label Data Distribution

Table 4 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 [**?**] suggests that the imbalance in labels, seen in our ground truth, is to be expected as individuals overwhelmed by stress tend to be outliers. Skewed datasets could lead to poor prediction performance if not corrected [**?**]. We addressed the problem of the imbalanced dataset by applying SMOTE [**?**] to synthetically oversample training set

	$Study_{SE}$	$Study_{ValidA}$	$Study_{ValidB}$
severe stress	145 (9%)	3 (2%)	1 (1%)
normal stress	1529 (91%)	129 (98%)	944 (99%)
depressed	534 (32%)	28 (21%)	330 (35%)
non-depressed	1140 (68%)	104 (79%)	615 (65%)

Table 4. Distribution of stress and depression labels for all studies; 27 samples per $Study_{SE}$ and $Study_{ValidB}$ participants, and 12 samples per $Study_{ValidA}$ participants.

RIGHTSLINK

data in the *severe stress* and *depressed* classes. SMOTE is widely applied in similar dataset problems as ours, and have shown to improve over other re-sampling techniques, including modifying loss ratio and class weights [**?**]. Note that upsampling was strictly contained in the training set; thus, our presented results reflect the true performance of an unaltered test set.

5.3 Interview Response

The semi-structured interview sessions were guided by questions on students' primary sources of stress and experiences of stressful workgroup situations. To ensure stability, accuracy and reproducibility, we used the same two coders for all interviews, with both coders using a standard coding scheme to reflect critical categories such as the primary source of stress and any experiences of critical (positive or negative, constructive or emotional) team experiences. We now present several insights from the ground truth data.

5.4 Ground Truth Insights

Figure 3 charts the percentage of our students in each of three user studies reporting *severe stress* over a 3-days interval amounting to 27 samples. 7% of our $Study_{SE}$ students reported *severe stress* at the beginning of the semester and peaked at 17% on Day 69 before the final project deliverable. In contrast, only one student from $Study_{ValidB}$ reported *severe stress* on Day 45 when the semester resumed. We sampled $Study_{ValidA}$ students from the second half of the semester (Day 45 onwards) and received the first reports of *severe stress* on Day 51 and towards semester end. Note: $Study_{ValidA}$ was intended as a small-scale study of participants whose behaviours were only monitored for a limited time. Most $Study_{SE}$ students (33) reported SE as their primary source of stress with 14 students attributing the stress to negative emotional interactions leading to relationship tension with their team members. Experiences of detrimental emotional disagreements were commonly attributed to feeling devalued for their efforts or believing others did not make a concerted effort to meet their standards. 17 students did not explicitly state their primary sources of stress but expressed a mix of personal and academic factors, while the remaining 12 students attributed their stress to other academic courses.

In general, we received a higher percentage of depression reports (see Figure 4). Further, the analysis revealed a concerning trend of 40 student participants who reported feeling depressed

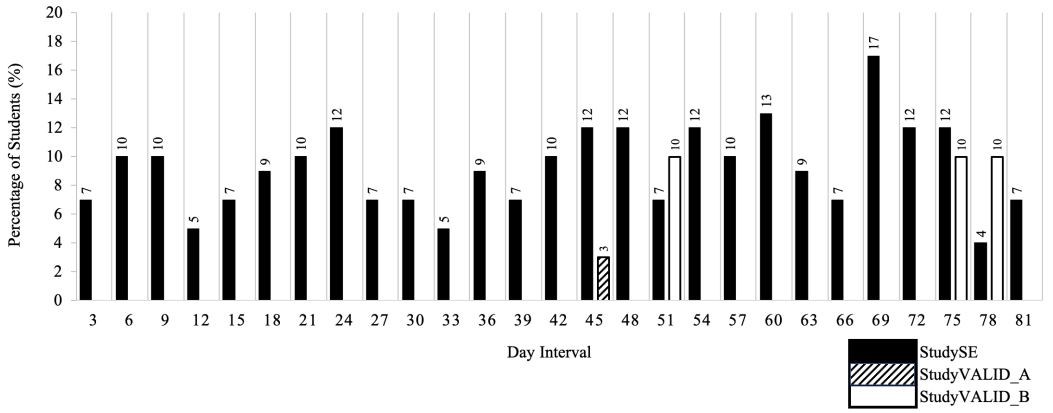


Fig. 3. Histogram of percentage of students from different user studies reporting *severe stress* (PSS-4 score more than 12) every 3 days. Samples for $Study_{ValidA}$ students were only collected from day 45 onwards.

RIGHTSLINK

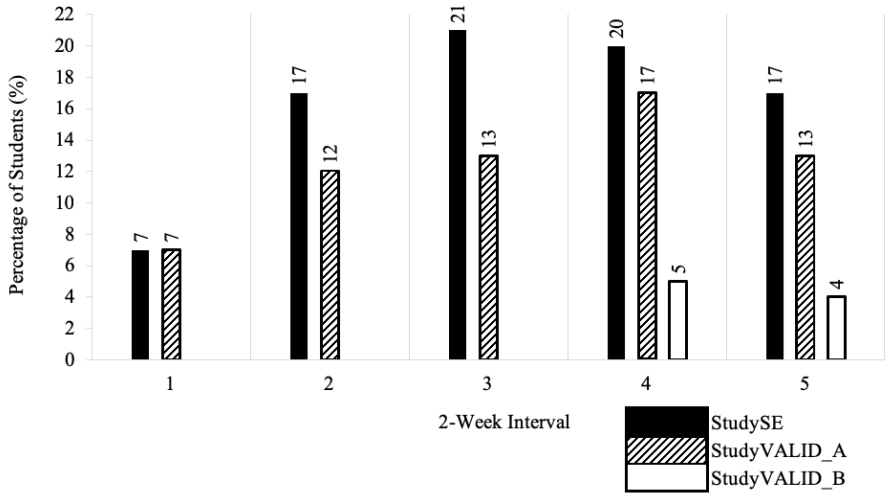


Fig. 4. Histogram of percentage of students from different user studies reporting feeling *depressed* (PHQ-8 score more than 9) approximately every 2 weeks. Samples for $Study_{ValidA}$ students were only collected from day 45 onwards, corresponding to sample 4 and 5.

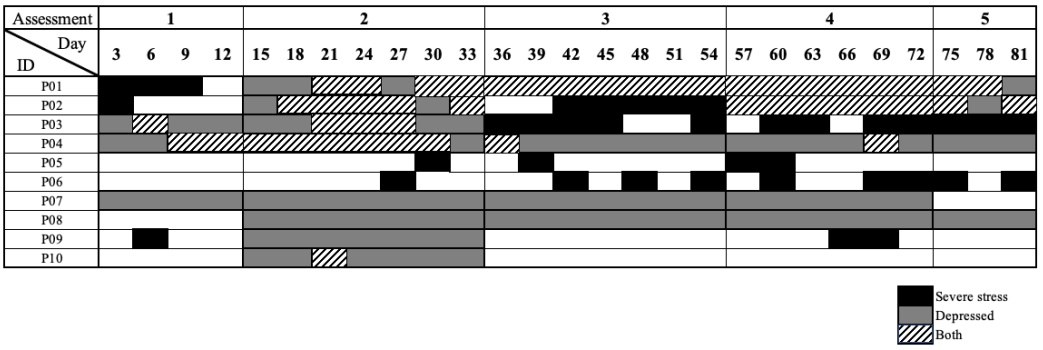


Fig. 5. Reports of *severe stress* and/or *depression* of 15 students from Study_SE charted selectively to illustrate four different patterns; 1) frequent instances of *severe stress* and *depression* (P01-P04), 2) only *severe stress* (P05,P06), 3) only *depression* (P07,P08) and 4) occasional *severe stress* and *depression* (P09,P10).

continuously for approximately four weeks (for all studies). Among these 40 were four $Study_{SE}$ students, who simultaneously experienced frequent severe stress from SE (see Figure 5, P01-P04). In real-world operation, students who are concurrently depressed and severely stressed and frequently depressed but not severely stressed are those that *StressMon* detects as "red-flags" so that interventions can take place as early as possible.

5.5 Location Data

The bulk of our data collection comprised of WiFi signals sensed directly from every AP to generate location and group information. This data collection was in collaboration with our university IT department who were already using an existing localisation solution [43] and allowed us access

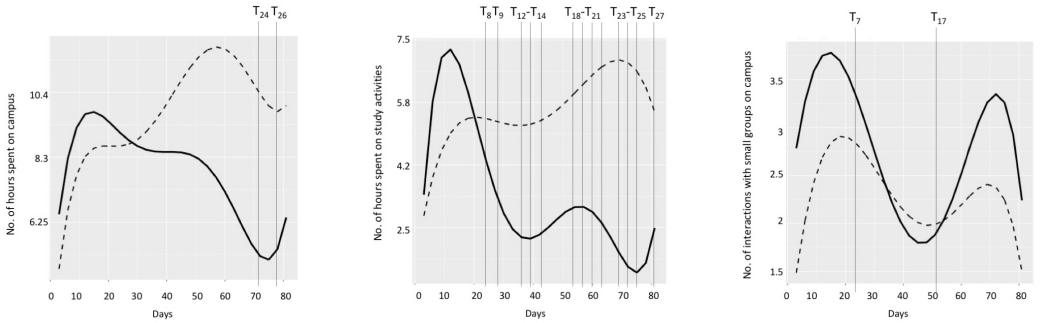
RIGHTSLINK

to that data. We extracted three months worth of WiFi signal data for $Study_{SE}$ and $Study_{ValidB}$ and one-month for $Study_{ValidA}$. These records amounted to an average of 7 hours location and 2 hours of group interaction data per student. Students were detected to have visited, on average, 96 unique locations on campus each month. We extracted mobility features from location and group data points for each day and mapped these features to stress (PSS-4) and depression (PHQ-8) labels. In summary, the final dataset consists of 1674 data points (62 users * 81-days averaged over three days) for $Study_{SE}$, 132 data points for $Study_{ValidA}$ (11 users over 36-days), and 945 data points for $Study_{ValidB}$ (35 users over 81-days).

6 EFFICACY OF LOCATION & GROUP FEATURES

We explored the feasibility of detecting stress using only coarse-grained location data collected from the campus WiFi network. First, we developed a set of mobility-driven features and hypotheses based on interview studies with $Study_{SE}$ students from our main study. Hypothesis testing was performed to validate features which statistically 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 6 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. 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 periods.



(a) Total time spent on campus - T24 and T26 are significantly different between groups (p<0.05).

(b) Time spent on work-typed activities - Multiple time points are significantly different between groups (p<0.01).

(c) Participation in small-sized groups - T7 and T17 are significantly different time point between groups (p<0.05).

Fig. 6. Mean plots of different mobility features between the *normal stress* and *severe stress* groups. We highlight the time points where the two distributions were statistically significantly different. 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. Note: dashed line represent students with *normal stress*, solid line represents students with *severe stress.*

RIGHTSLINK

6.1 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 occurrence is an expected trend because students typically spend more time working on projects and preparing for examinations, as the semester ends. Yet, students with *severe stress* were found to spend significantly less time on campus ($p<0.05$), specifically on T_{24} and T_{26} (see Figure 6a). Interestingly, students with *severe stress* exhibited declining participation during the same time their stress level peaked on Day 21.

H2: Students with severe stress participate less in work activities on campus. Students with *severe stress* were significantly less involved ($p<0.01$) in work-related activities (i.e., seminar attendance, self-study and group project activities) than students with *normal stress* (see Figure 6b). A more interesting observation is how severely stressed students began the semester displaying more participation in these activities but decreased over time. 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.

6.2 Group Interaction

As the SE course demands technical rigour and emphasises on project management, we believe group interaction factors are crucial indicators of stress. We hypothesised:

H3: Students with severe stress participate more in group work activities on campus. From Figure 6c, we found students with *severe stress* spent more time with small groups ($p<0.05$), with a significant difference at T_7. However, a noticeable (and significant) dip happened during the recess week at T_{17}. Additionally, we found that students with *severe stress* spent significantly more time with their SE groups ($p<0.05$). 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 reliable indicators of stress levels. We used these insights to build *StressMon* detection solution.

7 EVALUATION OF SYSTEM

Table 5 summarises our results with different sets of features used for the stress models, $\boldsymbol{Model_{S\_SE}}$ and $\boldsymbol{Model_{S}}$, and depression model, $\boldsymbol{Model_{D+}}$. $\boldsymbol{Model_{S\_SE}}$ is specific to $Study_{SE}$ student population as it uses *Domain-specific* (SE-related) features. $\boldsymbol{Model_{S}}$ is a generalised version which excludes all *Domain-specific* features. Finally, $\boldsymbol{Model_{D+}}$ is the depression model which additionally uses *Neuroticism*, one of the Big-5 personality traits, as a feature.

Experiment Setup. We conducted our evaluation in three parts: First, we performed a *Group K-fold* cross-validation (CV), splitting 80% of the dataset for training and 20% for testing. That is, 12-13 distinct students from $Study_{SE}$ make up each test fold to determine various model settings for $\boldsymbol{Model_{S\_SE}}$. Next, we conducted *Train-Test* by training on the whole $Study_{SE}$ dataset, and validated individually on $Study_{ValidA}$ and $Study_{ValidB}$. Finally, we built an *All-population* model by combining all users from three populations and performed a Group K-fold CV. Note: We did not use a Leave-one-out validation due to the highly imbalanced dataset, which might result in no one *severe stress* sample in a user. Across 12-13 students each contributing 27 samples, at least one sample in each group reported *severe stress*.

RIGHTSLINK

				Feature Settings			Performance	
Sec.	**Model**	**Study**	**Method**	**Set**	**Type**	**Interval**	**AUC**	**TPR (%)**
7.1	***Model*$_{S\_SE}$**	*Study*$_{SE}$	Group 5-fold 5-fold	*General*+ *Domain-specific*	***rel*+*abs***	6-days	0.96	98.93
7.2	***Model*$_S$**	All	Group 5-fold	*General*	***rel*+*abs***	6-days	0.97	96.01
7.3	***Model*$_{D+}$**	All	Group 5-fold	*General*+ Neuroticism	***rel*+*abs***	15-days	0.88	91.21

Table 5. Summary of stress and depression models configuration achieving best performances. Sections 7.1, 7.2 and 7.3 provide detailed results for each experiment. ***Model*$_{S\_SE}$** is a highly specific stress model that uses *Domain-specific* (SE-related) features, while ***Model*$_S$** excludes all *Domain-specific* features. ***Model*$_{D+}$** adds Neuroticism score of Big-5 as a feature to detect depression.

Performance Metrics: We computed the *area under the ROC curve* (AUC) score to cater to the class imbalance of *severe stress* in all population sets, and also calculated the *True Positive Rate* (TPR), *True Negative Rate* (TNR), and *Misclassification Rate* (MR) for each experiment. In an ideal scenario, our model should achieve an AUC score close to 1 (indicating a perfect predictor), and high TPR as the small number of positive *severe stress* occurrences must be correctly identified. We now present the results for each evaluation beginning with the Group 5-fold CV results on *Study*$_{SE}$ students.

7.1 Stress Detection: Main Study

Choice of Algorithm. First, we determined the use of algorithms, comparing Support Vector Machine (SVM) and Random Forest (RF) against Logistic Regression (LR) and all features, as the base classifier. Tuning all classifiers to achieve good performance on the positive class (high AUC), we empirically determine the cutoff of the classifier, which is typically set at 0.5 to 0.45 (thus, at the cost of a high false negatives rate). This change led to our results in all other experiments to obtain higher AUC, prioritising TPR. As shown in Table 6a, RF yielded significantly better AUC=0.97 (at p=0.01 level) than LR and SVM (0.57 and 0.86 respectively); subsequently, became our choice algorithm.

Feature Set. Next, we investigated our hypothesis that *change features make the strongest predictors of stress* (see Section 3.3). We achieved the highest AUC score of 0.97, using a combination of raw and change features (***raw*+*rel*+*abs***). However, the addition of ***raw*** set did not lead to significantly better performance. Hence, we retained only the change set as a smaller set of features to avoid developing an overfitted or computationally expensive model. Using change features, we were able to achieve an AUC of 0.95 (see Table 6b). In addition, we used recursive feature elimination (RFE), with backward elimination of 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. Table 6c lists the top 10 features sorted in the order of variable importance; that is, the ROC curve analysis conducted on each predictor is used as the measure of importance. Most top features used were extracted from the location data (unrelated to *Group*).

Individual (Work + Non-work) vs. Group Interaction Features. To better understand the best set of features used in our stress detection model, the next step was to compare each model performance using only individual routine features (these were Work and Non-work features extracted from location data) and social interaction features (these were Group features extracted from group data). As summarised in Table 6d, the use of group-related features alone did not yield high performance.

RIGHTSLINK

	LR	SVM	RF
AUC	0.57	0.86	0.97
TPR (%)	69.77	79.41	99.60
TNR (%)	35.35	68.27	72.00
MR (%)	62.65	30.73	25.56 (*)

(a) Results from using all features on different algorithms; Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF). (*) indicates significantly lower rate of misclassification at p=0.01 level.

	raw*+*rel*+*abs	***raw***	***rel*+*abs***
AUC	0.97	0.91	0.95
TPR (%)	99.60	98.93	99.33
TNR (%)	72.00	59.65	69.49
MR (%)	25.56	36.82	27.92

(b) Results from using different combination of feature types on chosen Random Forest algorithm; All (***raw*+*rel*+*abs***), Raw (***raw***) and Change (consists of relative change, ***rel*** and absolute change, ***abs***).

Description	Type	varImp
Number of times engaged in *studying*	***abs*** (*Work*)	100.0
Number of times being in *solo group*	***abs*** (*Group*)	48.57
Number of times being in *solo group*	***rel*** (*Group*)	40.22
Number of times engaged in *eating*	***abs*** (*Non-work*)	33.68
Number of times engaged in *studying*	***rel*** (*Work*)	32.38
Number of times engaged in *exercising*	***rel*** (*Non-work*)	31.83
Number of unique building visits	***abs*** (*Non-work*)	29.60
Number of times engaged in *transiting*	***abs*** (*Non-work*)	29.21
Total time spent with *small+medium groups*	***abs*** (*Group*)	25.34
Number of times engaged in *attending lectures*	***rel*** (*Work*)	24.67

(c) Top 10 features for detecting *severe stress*, using ROC curve analysis, and sorted by variable importance (varImp)

	rel*+*abs(W+NW+G)	Individual (W+NW)	Group (G)
AUC	0.95	0.85	0.69
TPR (%)	99.33	95.7	97.99
TNR (%)	69.49	60.44	22.03
MR (%)	27.92	36.63	71.62

(d) Results from separating the changes in group-related features from individual (routine) features; Change (***rel*+*abs***), Individual (includes Work and Non-work routines) and Group (consists of solo, small and medium groups) to detect stress at 3-days interval.

	3-days	6-days	9-days	12-days	15-days	18-days
AUC	0.95	0.96	0.89	0.82	0.89	0.83
TPR (%)	99.33	98.93	88.55	77.91	67.35	56.93
TNR (%)	69.49	75.54	84.62	87.86	90.27	91.52
MR (%)	27.92	22.41 (*)	14.52	12.07	10.59	10.17

(e) Results from calculating chosen Change type (***rel*+*abs***) features on different time intervals; from 3 to 18-days. (*) indicates significantly lower rate of misclassification at p=0.1 level.

Table 6. Results from conducting Group 5-fold CV on three different experiments to derive best model settings using Random Forest algorithm and Change feature type calculated every 6-days for $\boldsymbol{Models_{S\_SE}}$. $\boldsymbol{Models_{S\_SE}}$ is highly specific to the $Study_{SE}$ student population as it uses *Domain-specific* (SE-related) features.

RIGHTSLINK

Instead, a large portion of its inaccuracies was attributed by the signification reduction in TNR. In contrast, the changes in individual routines make stronger predictors from correctly classifying more negative cases. However, the combination of all change features proved to significantly improve overall accuracy to 72.08% (MR=27.92%). Accordingly, we retained all change features.

Time Window Experiment. Finally, we sought to determine the time at which our model would detect severe stress most accurately. With gradual time increase every 3 days (corresponding to the frequency of PSS-4 samples collected), we observed the highest AUC=0.96 on a 6-days interval. The reduced misclassification rate of 5.51% with 6-days interval (22.41%) is an improvement (at p=0.1 level) than the 3-days interval (see Table 6e). As the interval increases to 9-days, both TPR and TNR achieved comparable results, leading to lower misclassification rate of 14.52%. At this point, it is important to consider time as a key factor of intervention. That is, while stress is an everyday experience and chronic stress evolves over a longer period, prolonging detection of severe stress by more than a week might result in students missing out on vital help, leading to depression. Since timeliness should be prioritised over the accuracy, similarly, prioritising true positives over true negatives, we concluded the best model settings for detecting severe stress using Random Forest (RF) algorithm, all change set features (***rel+abs***) calculated at a 6-days interval – $\boldsymbol{Model_{S\_SE}}$.

7.2 Stress Detection: Validation Study

Recall $\boldsymbol{Model_{S\_SE}}$ includes *Domain-specific* (SE) features, which are highly tailored to SE students in the $Study_{SE}$ sample. Accordingly, we generalised the model to exclude all *Domain-specific* features as $\boldsymbol{Model_S}$. First, we trained on $Study_{SE}$ sample and tested on different populations. Then, we performed a Group 5-fold CV on all three populations. Table 7 lists our results in detail.

Our solution successfully yielded an AUC=0.94, 100% TPR as it correctly detected 1 *severe stress* instance in $Study_{ValidB}$. While the misclassification rate dropped to 18.73% (81.25% TNR), the false detection, unfortunately, affected most students in the sample. Out of 35 students, 14 had reported feeling *depressed* despite not experiencing *severe stress*. Our test on $Study_{ValidA}$ students achieved reduced AUC=0.91 and 66.67% TPR. That is, out of 3 *severe stress* reported by two students (1 student reported two instances of *severe stress*), the student with one report of *severe stress* was misclassified. Approximately 10% misclassification (90.70% TNR) was as a result of 6 students, 2 of whom did not report *severe stress* but felt *depressed*.

The final step combined all students from three user studies to build an all-population stress model, evaluated using a Group 5-fold CV. We achieved an average AUC=0.97 and 96.01% TPR (4 out of 149 *severe stress* instances would go unnoticed). Unfortunately, the misclassification rate of 18.20% (80.76% TNR) continued to affect most participants by identifying them as severely stressed

Method	**Train**	**Test**	**AUC**	**TPR** (%)	**TNR** (%)	**MR** (%)
Train-Test	$Study_{SE}$	$Study_{ValidA}$	0.91	66.67	90.70	09.85
	$Study_{SE}$	$Study_{ValidB}$	0.94	100.0	81.25	18.73
Group 5-fold (All population)	Folds 2-5	Fold 1	0.98	94.44	86.81	12.94
	Folds 1,3-5	Fold 2	0.96	88.88	84.26	15.66
	Folds 1-2,4,5	Fold 3	0.97	98.64	80.88	17.91
	Folds 1-3,5	Fold 4	0.96	98.07	75.46	22.35
	Folds 1-4	Fold 5	0.96	100.0	76.38	22.35
		Average	0.97	96.01	80.76	18.24

Table 7. Summarised results for stress model, $\boldsymbol{Model_S}$, on three different validations. $\boldsymbol{Model_S}$ is a generalised stress model that excludes all *Domain-specific* features.

RIGHTSLINK

at some point in the study. We verified that at least 35 students with misclassified instances had reported multiple accounts of *depression* during the same time. The higher percentage of students who reported *depression* as compared to *severe stress* raised a serious need for our technique to detect *depressed* students successfully.

7.3 Depression Detection

7.3.1 Main Study. We continued to investigate how our model could be used to detect *depressed* users, simultaneously drawing comparisons to a recently published work by Ware *et al.* [69]. Note: The authors used WiFi indoor localisation data to extract building-level features such as ours (i.e., *General* ***raw*** typed features). We set the time window interval to calculate changes in routines and group interactions to 15-days, close to the PHQ-8 assessment interval, although [69] was set at 13-days.

We built $\boldsymbol{Model_{D\_raw}}$ with *General* ***raw*** features, similar to Ware's *Day Monitoring* scenario. Note that their experiments ran on two phases, resulting in an average TPR of 77.00% and TNR of 59.50%. Unfortunately, a Group 5-fold CV of $\boldsymbol{Model_{D\_raw}}$ did not achieve comparable results (AUC score = 0.57). Using all features (***raw***+***rel***+***abs***) reduced MR to 41.51% and increased AUC score to 0.68. In contrast, $\boldsymbol{Model_{D\_chg}}$, built with change features (***rel***+***abs***) calculated over a 15-days interval, achieved better performance of AUC=0.72 and improving TNR by 53.39%. Table 8a lists the top 5 most important features being used. In comparison to detecting stress, most top features were extracted from group data.

Accordingly, we repeated the process of separating individual routine features (these were Work and Non-work features extracted from location data) and social interaction features (these were Group features extracted from group data) to investigate the differences in using these features for detecting depression. As summarised in Table 8b, the removal of group features only led to our model correctly classifying the positive cases at random chance. In comparison, group features achieved 65.24% TPR and reduced AUC score of 0.67. While group features were stronger predictors for depression, using these features alone did not yield high performance; individual features improved TNR from 54.88% to 63.53%. Accordingly, we retained all change features.

To better improve our model performance, we analysed the classification results by reviewing profiles of misclassified students, including their gender, academic year, GPA and Big-5 personality assessment [32]. Our manually-driven analysis revealed only one case of depression by a student who scored low on neuroticism (score <= 2.25 out of 5). The most significant portion of depression reports was by students whose neuroticism scores were 3.75 and 4. Indeed, many studies draw correlations between high neuroticism scores and depression [23, 51]. Thus, we revised the model to include neuroticism score (denoted as 'N' in Table 8) as a feature for $\boldsymbol{Model_{D+}}$. The added feature helped boost TPR up to 90.21% and TNR to 69.45% (but not significant).

7.3.2 Validation Study. Further validation of $\boldsymbol{Model_{D+}}$ (trained on $Study_{SE}$ and tested on $Study_{ValidA}$ and $Study_{ValidB}$) did not achieve favourable results in detecting *depressed* students from $Study_{ValidA}$. However, a Group 5-fold CV on an all-population model yielded an average AUC=0.88, 91.21% TPR and 66.71% TNR; 9 out of 55 students who reported depression had several instances of depression misclassified. Unfortunately, one student was completely undetected by the model.

7.4 Summary

The best results obtained by our models are summarised in Table 5. Overall, the evaluation demonstrated the strength of *StressMon* in two different ways. First, our approach does not require training a new stress model to detect *severe stress* in different groups of students. For example, even when using a model trained solely from students enrolled in Software Engineering ($Study_{SE}$), the $\boldsymbol{Model_S}$

RIGHTSLINK

still achieved a high 0.94 and 0.91 AUC score when used on two different populations, $Study_{ValidA}$ and $Study_{ValidB}$, respectively. In addition, the removal of *Domain-specific* features (i.e., features related to Software Engineering project) also increased the TNR for our all-population model. However, this had the benefit of increasing the TPR and thus improving the likelihood of *StressMon*

Description	Type	varImp
Total time spent with *all groups*	***abs*** (*Group*)	100.0
Number of times being in *solo group*	***abs*** (*Group*)	96.66
Number of times being in *all groups*	***abs*** (*Group*)	91.31
Number of times engaged in *studying*	***rel*** (*Work*)	85.66
Number of times engaged in *transiting*	***abs*** (*Non-work*)	82.51

(a) Top 5 features for detecting *depression*, using ROC curve analysis, and sorted by variable importance (varImp)

	rel+***abs***(W+NW+G)	Individual (W+NW)	Group (G)
AUC	0.72	0.63	0.67
TPR (%)	70.25	51.83	65.24
TNR (%)	63.53	64.72	54.88
MR (%)	34.42	36.68	43.09

(b) Results from separating the changes in group-related features from individual (routine) features; Change (***rel***+***abs***), Individual (includes Work and Non-work routines) and Group (consists of solo, small and medium groups) to detect depression at 15-days interval.

Model	Feature Settings	Method	Train/Test	AUC	TPR(%)	TNR(%)	MR(%)
$\boldsymbol{Model_{D\_raw}}$	Set: General Type: ***raw*** Interval: 3-days	Group 5-fold	$Study_{SE}$	0.57	93.50	10.14	63.59
$\boldsymbol{Model_{D\_chg}}$	Set: General Type: ***rel***+***abs*** Interval: 15-days	Group 5-fold	$Study_{SE}$	0.72	70.25	63.53	34.42
$\boldsymbol{Model_{D+}}$	Set: General Type: ***rel***+***abs***+N Interval: 15-days	Group 5-fold	$Study_{SE}$	0.87	90.21	69.45	22.84
		Train: Test:	$Study_{SE}$ $Study_{ValidA}$	0.41	57.14	47.12	50.76
		Train: Test:	$Study_{SE}$ $Study_{ValidB}$	0.86	81.52	72.68	24.23
		Group 5-fold	*All population*	0.88	91.21	66.71	24.94

(c) Summary of depression model performances achieved from different feature settings.

Table 8. Results from conducting different experiments to derive best feature settings for depression detection. $\boldsymbol{Model_{D+}}$ uses Random Forest algorithm, Change feature type calculated every 15-days and single Big-5 personality score, Neuroticism, denoted as N. $\boldsymbol{Model_{D\_raw}}$ uses only *General* ***raw*** features and $\boldsymbol{Model_{D\_chg}}$ uses *General* change features.

RIGHTSLINK

detecting the small number of *severe stress* cases even at the cost of misclassifying more cases of no stress as *normal stress*. We believe that a model that prioritises accurate detection of the more critical cases is the right tradeoff for an early warning solution. Second, our evaluation of $\boldsymbol{Model_{D+}}$ showed that *StressMon* could detect depression with an AUC=0.88 (91.21% TPR and 66.71% TNR) on all studies. Overall, *StressMon* is better at detecting stress than prior work [10, 29, 59] that use much finer-grained app-collected data. While our approach achieved relatively similar performance to prior work [69] when using just mobility-driven features, we found that including personality traits, specifically 'neuroticism', improved the prediction accuracy. Our findings support prior work that found personal characteristics [23, 56] could largely attribute depression.

8 *STRESSMON* IN THE CSCW DOMAIN

We now discuss the application of *StressMon* to CSCW domains; especially in detecting mental health issues in individuals participating in various social and workgroups.

8.1 Naturalistic Evaluation of Workgroup Processes

In substantiating a process-oriented approach to evaluate collaborative technologies, Neale and Carroll highlight the importance of naturalistic methods to study various factors of group processes [45], and how these studies must be cognizant of social group dynamics shaping user behaviours, and how these factors should be applied to the design of collaborative solutions [20, 45].

8.1.1 Mechanism to Measure the Dynamism of Cooperations. According to Johansen, *cooperative work* is distributed physically in time and space. Further, membership of cooperative ensembles is often non-determinable with transient formations, emerging to handle different work situations [31]. The physical participation of individuals within their workgroup collectives has also been found to be negatively impacted by emotional intra-group conflict and spilt over to their CMC-based interactions [71].

We believe that *StressMon* can enable the monitoring of natural user behaviours in small groups, where group members are co-located, and characterise work group dynamics. Referencing prior work that shows working with severely stressed individuals can manifest into various emotional outcomes, *StressMon* can be used to conduct non-invasive longitudinal evaluations to understand how individual health conditions influence the dynamics of group processes and the use of tools. At present, most of the longitudinal evaluations focus on analysing rich data of online behaviours to study how technologies are being used in teams. However, we believe that *StressMon* can enable more comprehensive studies of the underlying relationships in physical groups by providing a seamless non-invasive yet accurate lens into the physiological states of the group members.

8.1.2 Support for Individual and Group Level Interventions. These longitudinal evaluations can similarly support the studies of interventions intended to improve team effectiveness. For example, we learn from prior work that conflicting situations can be differentiated by how individuals physically interact within their workplace settings, and additionally, leading to the emotionally-charged individuals pulling themselves out of online communicative space [71]. With *StressMon* ascertaining behavioural differences of individuals in their workgroups, we believe such behavioural analyses can help researchers explore ways in which interventions may be targeted at individuals (e.g., to manage their stresses) or at teams as performing units (e.g., take a proactive approach towards demonstrating peer support to members affected by stressful situations). Essentially, Berg *et al.* suggested that systems for collectives must enhance the teams' competencies and responsibilities through means of engaging individuals in sense-making [5].

RIGHTSLINK

8.2 Passively Supporting Non-responders in Healthcare

In reviewing systems for healthcare, Fitzpatrick suggested the contribution for the broader research community is understanding how technologies can support the everyday collaborative practices between diverse professional groups in direct contact with the individuals (receiving health support) [20, 33]. A large amount of prior work in the CHI and related domains has explored creative and engaging ways of supporting interactions between the patient and professional [47]. Despite healthcare technologies progressing to offer real-time measurements of individuals and responses of professionals [39], these technologies operate on the assumption of an active user role from both parties. In reality, professionals remain overwhelmed by the sheer number of individuals to "look after", albeit technologies are providing better workflow management. Furthermore, many passive health seekers want to be healthy but are not actively participating to receive health support.

StressMon can enable researchers to study new forms of interaction and engagement between passive stakeholders. In the case of monitoring stressed/depressed students on campus, past studies have found that educators do not believe it is their place to intervene in students' mental health unless they are being approached [19]. Even if students seek assistance from a university counselling service, the service is usually understaffed and under-resourced to support large groups of students [22]. We believe *StressMon* can operate effectively in this space by delivering group-level interventions anonymously. That is, upon identifying and localising users who are severely stressed/depressed (and whose identity remains unknown), *StressMon* can assist educators with group-level strategies of improving (the entire class of) students' awareness in mental health issues. These strategies can be decided and recommended by counsellors as online mental health resources. The realisation of such technology plays a two-fold part in research; (1) designing effective and ethical interventions for passive users and (2) developing regulatory frameworks for real-world practice.

8.3 *StressMon*'s Ethical Practice

With increasing progress in enabling technologies for large-scale behavioural research, ethical concerns remain a challenge within the research community. Since *StressMon* considers the influence of social relationships to detect severe cases of mental health issues, the process of collecting and deriving behavioural patterns in groups of user data must abide by compelling ethical principles. Referencing the CSCW guidelines for social computing [66], we argue that the mechanisms operationalising *StressMon* comply with principles of the Belmont Report [14] in the following ways:

1. Respect for Persons: Having an informed consent form (with IRB-approval) is the most straightforward way of respecting and protecting users from harm. For *StressMon*, data collection (of individual location and group location) is enabled by a single-sourced, Wi-Fi indoor localisation system. This mechanism allows us to bypass direct communications with the user's device to collect RSSI values from the Wi-Fi APs (of devices being connected to the campus Wi-Fi). Additionally, the localisation mechanism used [43] maintains anonymity by applying a one-way hash function, which prevents a user's device from being easily identified. Unless a user consents to disclose their device's MAC address, they will be tracked anonymously.

2. Beneficence: Beneficence is weighing the risks over the benefits of any research. The risks of exposing users' identity are minimised because we are making no direct communication with users' devices to collect location information and no requirement for personally identifiable information to conduct behavioural analyses. Additionally, while WiFi operates in an unlicensed spectrum, the RTLS data is encrypted by the AP before transmission and is thus hard to decode by a hacker. Even so, it could be possible for the RTLS data to be retrieved by a determined hacker and even

RIGHTSLINK

anonymised location information could still be used to identify a small subset of the user population. Despite these risks, we believe *StressMon* provides more significant social benefit for both individual and collective levels. With approximately 66% of college students suffering from either depression or stress [30], and campus service providers facing resource crisis [22], *StressMon* can be deployed as a campus-wide "safety net" for those in greatest need of help. *StressMon* can be an enabler for students to receive help via external methods such as interventions moderated by counsellors.

3. Justice: Justice requires fair user participation. Fairness is true for *StressMon*, as its data collection is not influenced by factors such as the socioeconomic status or technical experience of the user. Instead, *StressMon* leverages Wi-Fi, which is readily available in public spaces (e.g., offices, campuses and shopping malls) and commodity devices (e.g., laptops and mobile phones). Since monitoring with *StressMon* does not require any explicit user interaction of installing/running a dedicated application on their phone, the resource is readily available to all users in the environment.

The evolution of social sensing enables the measuring of large-scale human behaviour. Technologies, such as *StressMon*, provide foundational mechanisms for interdisciplinary research communities to explore new methods of facilitating mental health benefits/interventions and studying the natural processes of small group phenomena while ensuring such studies stay within the boundaries of ethical practices.

9 LIMITATIONS AND FUTURE WORK

In this section, we clarify the barriers of scaling *StressMon* to all users and different environments, and describe our ongoing efforts to improve the system.

9.1 Indoor Location Sub-system Requirement

StressMon fundamentally requires the availability of an indoor positioning system that can generate location information for every device in the environment using data collected solely from the infrastructure. This data is then processed by our software to generate group information and predictions. Currently, we use WiFi as it is the predominant solution deployed and used on our campus, and we believe it is also readily available on other campuses worldwide as well. If the WiFi deployment in a particular environment is sparse, then the accuracy of the location tracking will decrease, and this could affect the performance of *StressMon*. The indoor location solution used by *StressMon* [43] currently works with WiFi networks that use equipment from Aruba [40], Cisco [11], Zebra [72], or Ubiquiti [46]. Moreover, *StressMon* can leverage other techniques such as Bluetooth if it is deployed generally; for example, at hospitals to help staff find their way to departments or wards [42]. In the future, if new technologies, such as 5G, replace WiFi in indoor environments, *StressMon* will be modified to use these technologies for its base sensing needs.

9.2 Applicability to Other Workplace Settings

In this paper, we showed how *StressMon* could accurately detect stress and depression among students in a university setting. It seems likely that *StressMon* would work on other campuses as well as there is nothing in our solution that is explicitly tied to our campus. But how easy is it to deploy *StressMon* in other work environments? Fundamentally, *StressMon* uses deviations in work routines and interactions to produce its output; thus, it will not work well in highly regimented work environments where the location of an individual does not change significantly across time. For example, factories where each worker is assigned to a dedicated point in the assembly process and stays there the entire day with minimal interaction with their peers (except during brief breaks) or elementary level education where students are in the same classroom the entire day. Instead, *StressMon* works best in work environments where monitoring deviations in work schedules and collaborative practices is possible. For example, on university campuses, hospitals, or military bases

RIGHTSLINK

where students, nurses, and officers frequently move, daily, to different parts of the environment, and have ample opportunities to interact with different people. Moreover, offsite work behaviours and online work collaborations are not yet supported. We can extend *StressMon* to use more sensors, such as GPS, if necessary. However, such extensions reduce the scalability (as these sensors will require apps or other mechanisms) and increases the privacy concerns.

9.3 Latency of Predictions

StressMon currently detects stress every 6 days and depression every 15 days. Thus, detection is not real-time, even though it collects and processes real-time data. Health monitoring solutions are offering real-time stress analysis [60] for real-time interventions. In contrast, we designed *StressMon* to detect large and significant swings in mental health, which require a sufficiently long measurement period. For example, *StressMon* differentiate a severely stressed individual, who is more likely to struggle with managing stress, from one that is just instantaneously stressed and then recovers. Additionally, detecting depression requires a longer observation window as this is a fairly fundamental change in mental health that needs to be carefully assessed. As stated previously, we strived to design a first level safety net that flags egregious changes in mental health *at scale*.

9.4 Correlation Between Stress and Depression

One of the more interesting takeaways from our studies was a validation of prior findings that stress and depression are only somewhat correlated [25]. In particular, we found cases (see Figure 5), where students who were detected as depressed (and who indicated as such on their PHQ-8 surveys) did not report being stressed. Understanding depression from the medical perspective, we learned that depression is a complex process associated with personality characteristics [23, 51, 56] and can occur without an individual feeling stressed [24]. In our study, we were able to quickly establish these students having relatively high negative emotions from their Big-5 personality (neuroticism) score. The downside to this observation is, unlike other features used, this personality trait is not mobility-driven. Separately, we found cases of highly stressed students, who reported being stressed over multiple consecutive reporting periods, who do not report themselves as being depressed. Our findings confirmed that stress and depression should be treated as two separate entities [27, 62], and thus, reinforced our decision to use separate models for stress and depression, although the models share many features. Note: there are sufficient differences to make the models or stress and depression unique and distinct. Conclusively, it is noteworthy that further research is needed to conveniently assess personality types as they characteristically reveal reliable indicators to a wide range of mental disorders, including depression.

9.5 Measuring Group Phenomena

We learned from our qualitative analysis that it is common for our students to feel severely stressed over negative emotional interactions with their team members. We hypothesise that these critical events are highly likely to cause a sudden change in group norms and could be observed through individuals' mobility patterns and interactions, as represented by our feature sets. To build on these findings, we are currently studying how group measures such as *conflict* and *social identification* can be distinguished through differences in mobility patterns. We believe this body of work will simultaneously add on to the CSCW community in understanding human behaviours within small groups and demonstrate the feasibility of utilising *StressMon* to conduct in-the-field longitudinal studies for group dynamics.

RIGHTSLINK

9.6 Extending to Different Populations

We have shown that changes in an individual's routine and their group interactions, extracted from coarse-grained location data, make useful features in detecting *severe stress* and *depression*. However, even though our experiments were tested on three different and separate student populations, sampled at different times, further studies will be required to determine the efficacy of *StressMon* in other work settings (both scholastic or professional). Currently, we are engaged with a different local educational institution who see *StressMon* as an appropriate mechanism to automatically detect severely stressed students in need of support. The operationalisation of *StressMon* includes deploying the same sensing mechanisms used in this study. We believe trialling *StressMon* in a different population of students will help us understand the nuances of workgroup dynamics across different learning cultures and how social factors affect individuals' stress and depression.

9.7 Privacy and Ethical Concerns

As discussed in Section 8, we believe *StressMon* can offer promising applications to provide mental health benefits, especially to users who are actively interacting in their social groups. While the use of our solution may be an unfamiliar practice to the Institutional Review Boards (IRBs), we have argued that *StressMon* follows the same basic ethical principles in terms of the data it collects and analyses it conducts. Nonetheless, *StressMon* must have appropriate policies and mechanisms protecting user privacy rights before it can be widely deployed. Policy solutions are especially crucial as *StressMon* is likely to be used, due to its inherent mechanisms, without explicit user consent. A more pressing concern that requires interdisciplinary research attention is on ways *StressMon* can provide actual benefits to users in the real world. For example, "What are the appropriate privacy and ethical policies to ensure that no individual feels unfairly targeted or discriminated against while ensuring that anyone who needs help (even if unaware of their need) receives it?" We believe a feasible way could be to enable group-level interventions between school counsellors and educators to support students passively. Even more importantly, sending interventions to individuals or groups with problems needs to be even more carefully monitored – to avoid the intervention accidentally worsening the condition. The next phase of *StressMon* includes (1) working with our psychology colleagues and student counsellors to design and evaluate various interventions that can be sent by a system that uses triggers generated by *StressMon*, and (2) working with our colleagues in Privacy and Ethics Law to develop appropriate policies and procedures for community-wide health monitoring systems such as *StressMon* that balance the privacy of individuals with the ability to provide help to those who most need it (and may not realise it).

10 CONCLUSION

We presented *StressMon*, a system to detect *severe stress* and *depressive* episodes in individuals. *StressMon* is designed to be a scalable solution that does not require installing specific applications or owning specific devices. Using coarse-grained location data collected directly from the WiFi infrastructure, we extracted features of individuals' routine behaviours and features that sufficiently describe an individuals' physical interaction patterns. These features were used in two different models to detect stress and depression, respectively. We demonstrated, via three different semester-long user studies involving 108 students at a university campus, that *StressMon* has an Area Under the Curve (AUC) score of 0.97 (96.01% TPR and 80.76% TNR) at detecting stress using 6-days interval, and an AUC of 0.88 (91.21% TPR and 66.71% TNR) at detecting depression using 15-day intervals.

RIGHTSLINK

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative.

REFERENCES

[1] Karl Aberer, Saket Sathe, Dipanjan Chakraborty, Alcherio Martinoli, Guillermo Barrenetxea, Boi Faltings, and Lothar Thiele. 2010. OpenSense: open community driven sensing of environment. In *Proceedings of the ACM SIGSPATIAL International Workshop on GeoStreaming*. ACM, 39–42.

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

[3] Lisa J Barney, Kathleen M Griffiths, Helen Christensen, and Anthony F Jorm. 2009. Exploring the nature of stigmatising beliefs about depression and help-seeking: implications for reducing stigma. *BMC public health* 9, 1 (2009), 61.

[4] Meghan Baruth, Duck-Chul Lee, Xuemei Sui, Timothy S Church, Bess H Marcus, Sara Wilcox, and Steven N Blair. 2011. Emotional outlook on life predicts increases in physical activity among initially inactive men. *Health Education & Behavior* 38, 2 (2011), 150–158.

[5] Marc Berg. 1999. Accumulating and coordinating: occasions for information technologies in medical work. *Computer Supported Cooperative Work (CSCW)* 8, 4 (1999), 373–401.

[6] Andrew P Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, 7 (1997), 1145–1159.

[7] Chloë Brown, Christos Efstratiou, Ilias Leontiadis, Daniele Quercia, Cecilia Mascolo, James Scott, and Peter Key. 2014. The architecture of innovation: Tracking face-to-face interactions with ubicomp technologies. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 811–822.

[8] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, and David C Mohr. 2011. Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research* 13, 3 (2011), e55.

[9] 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.

[10] Jongyoon Choi and Ricardo Gutierrez-Osuna. 2009. Using heart rate monitors to detect mental stress. In *2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks*. IEEE, 219–223.

[11] CISCO. 2019. *Cisco Digital Network Architecture (Cisco DNA)*. https://www.cisco.com/c/en/us/solutions/enterprise-networks/index.html.

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

[13] Tom Cox. 1993. *Stress research and stress management: Putting theory to work*. Vol. 61. Hse Books Sudbury.

[14] Education Department of Health et al. 2014. The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research. *The Journal of the American College of Dentists* 81, 3 (2014), 4.

[15] Afsaneh Doryab, Jun Ki Min, Jason Wiese, John Zimmerman, and Jason Hong. 2014. Detection of behavior change in people with depression. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

[16] Begum Egilmez, Emirhan Poyraz, Wenting Zhou, Gokhan Memik, Peter Dinda, and Nabil Alshurafa. 2017. UStress: Understanding college student subjective stress using wrist-based passive sensing. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 673–678.

[17] Emre Ertin, Nathan Stohs, Santosh Kumar, Andrew Raij, Mustafa al'Absi, and Siddharth Shah. 2011. AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. ACM, 274–287.

[18] Anja Exler, Marcel Braith, Kristina Mincheva, Andrea Schankin, and Michael Beigl. 2018. Smartphone-Based Estimation of a User Being in Company or Alone Based on Place, Time, and Activity. In *International Conference on Mobile Computing, Applications, and Services*. Springer, 74–89.

[19] Louise Farrer, Amelia Gulliver, Kylie Bennett, and Kathleen M Griffiths. 2015. Exploring the acceptability of online mental health interventions among university teaching staff: Implications for intervention dissemination and uptake. *Internet Interventions* 2, 3 (2015), 359–365.

[20] Geraldine Fitzpatrick and Gunnar Ellingsen. 2013. A review of 25 years of CSCW research in healthcare: contributions, challenges and future agendas. *Computer Supported Cooperative Work (CSCW)* 22, 4-6 (2013), 609–665.

[21] Mental Health Foundation. 2019. *Mental health statistics: stress*. https://www.mentalhealth.org.uk/statistics/mental-health-statistics-stress.

RIGHTSLINK

[22] Leah Goodman. 2017. Mental Health on University Campuses and the Needs of Students They Seek to Serve. *Building Healthy Academic Communities Journal* 1, 2 (2017), 31–44.
[23] AMGF Griens, K Jonker, PH Spinhoven, and MBJ Blom. 2002. The influence of depressive state features on trait measurement. *Journal of Affective Disorders* 70, 1 (2002), 95–99.
[24] Constance Hammen. 2005. Stress and depression. *Annu. Rev. Clin. Psychol.* 1 (2005), 293–319.
[25] Constance L Hammen and Susan D Cochran. 1981. Cognitive correlates of life stress and depression in college students. *Journal of Abnormal Psychology* 90, 1 (1981), 23.
[26] Gaston Harnois, Phyllis Gabriel, World Health Organization, et al. 2000. Mental health and work: impact, issues and good practices. (2000).
[27] Harvard Medical School Harvard Health Publishing. 2017. *What causes depression?* https://www.health.harvard.edu/mind-and-mood/what-causes-depression.
[28] Olivier Herrbach. 2006. A matter of feeling? The affective tone of organizational commitment and identification. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 27, 5 (2006), 629–643.
[29] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 493–504.
[30] Justin Hunt and Daniel Eisenberg. 2010. Mental health problems and help-seeking behavior among college students. *Journal of adolescent health* 46, 1 (2010), 3–10.
[31] Robert Johansen. 1988. *Groupware: Computer support for business teams*. The Free Press.
[32] Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. The big five inventory—versions 4a and 54. (1991).
[33] Bonnie Kaplan and Kimberly D Harris-Salamone. 2009. Health IT success and failure: recommendations from literature and an AMIA workshop. *Journal of the American Medical Informatics Association* 16, 3 (2009), 291–299.
[34] David J Katzelnick, Farifteh Firoozmand Duffy, Henry Chung, Darrel A Regier, Donald S Rae, and Madhukar H Trivedi. 2011. Depression outcomes in psychiatric clinical practice: using a self-rated measure of depression severity. *Psychiatric services* 62, 8 (2011), 929–935.
[35] Kaitlin Kirasich, Trace Smith, and Bivin Sadler. 2018. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review* 1, 3 (2018), 9.
[36] 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.
[37] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications magazine* 48, 9 (2010).
[38] Christoph Lauber, Carlos Nordt, Luis Falcato, and Wulf Rössler. 2001. Lay recommendations on how to treat mental disorders. *Social psychiatry and psychiatric epidemiology* 36, 11 (2001), 553–556.
[39] Matthew L Lee and Anind K Dey. 2011. Reflecting on pills and phone use: supporting awareness of functional abilities for older adults. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2095–2104.
[40] Hewlett Packard Enterprise Development LP. 2019. *Aruba Enterprise Networking and Security Solutions*. https://www.arubanetworks.com/me/.
[41] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 351–360.
[42] Ehud Mendelson. 2015. Indoor and outdoor mapping and navigation utilizing RF bluetooth beacons. (Dec. 1 2015). US Patent 9,204,257.
[43] Archan Misra and Rajesh Krishna Balan. 2013. LiveLabs: initial reflections on building a large-scale mobile behavioral experimentation testbed. *ACM SIGMOBILE Mobile Computing and Communications Review* 17, 4 (2013), 47–59.
[44] Susan Mohammed and Linda C Angell. 2004. Surface-and deep-level diversity in workgroups: Examining the moderating effects of team orientation and team process on relationship conflict. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 25, 8 (2004), 1015–1039.
[45] Dennis C Neale, John M Carroll, and Mary Beth Rosson. 2004. Evaluating computer-supported cooperative work: models and frameworks. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. ACM, 112–121.
[46] Ubiquiti Networks. 2019. *Ubiquiti Networks - Software*. https://www.ui.com/software/.
[47] Tao Ni, Amy K Karlson, and Daniel Wigdor. 2011. AnatOnMe: facilitating doctor-patient communication using a projection-based handheld device. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3333–3342.
[48] OctoberCMS. 2016. *2017 WORK AND WELL-BEING SURVEY*. http://octobercms.com/.

RIGHTSLINK

[49] World Health Organisation. 2017. *Mental health in the workplace.* https://www.who.int/mental_health/in_the_workplace/en/.
[50] World Health Organisation. 2019. *Depression.* https://www.who.int/news-room/fact-sheets/detail/depression.
[51] Johan Ormel, Albertine J Oldehinkel, and Wilma Vollebergh. 2004. Vulnerability before, during, and after a major depressive episode: a 3-wave population-based study. *Archives of general psychiatry* 61, 10 (2004), 990–996.
[52] Alexandros Pantelopoulos and Nikolaos G Bourbakis. 2010. 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 (2010), 1–12.
[53] Hynek Pikhart, Martin Bobak, Andrzej Pajak, Sofia Malyutina, Ruzena Kubinova, Roman Topor, Helena Sebakova, Yuri Nikitin, and Michael Marmot. 2004. Psychosocial factors at work and depression in three countries of Central and Eastern Europe. *Social science & medicine* 58, 8 (2004), 1475–1482.
[54] Diego A Pizzagalli, Ryan Bogdan, Kyle G Ratner, and Allison L Jahn. 2007. Increased perceived stress is associated with blunted hedonic capacity: potential implications for depression research. *Behaviour research and therapy* 45, 11 (2007), 2742–2753.
[55] Qualtrics. 2019. *Qualtrics.* http://www.qualtrics.com.
[56] Sakina J Rizvi, Diego A Pizzagalli, Beth A Sproule, and Sidney H Kennedy. 2016. Assessing anhedonia in depression: potentials and pitfalls. *Neuroscience & Biobehavioral Reviews* 65 (2016), 21–35.
[57] Babak Roshanaei-Moghaddam, Wayne J Katon, and Joan Russo. 2009. The longitudinal effects of depression on physical activity. *General hospital psychiatry* 31, 4 (2009), 306–315.
[58] Paul Sacco, Kathleen K Bucholz, and Donna Harrington. 2014. 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 (2014), 456–465.
[59] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction.* IEEE, 671–676.
[60] Hillol Sarker, Matthew Tyburski, Md Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H Epstein, Kenzie L Preston, C Debra Furr-Holden, Adam Milam, Inbal Nahum-Shani, et al. 2016. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI conference on human factors in computing systems.* ACM, 4489–4501.
[61] Marc J Schabracq and Cary L Cooper. 2000. The changing nature of work and stress. *Journal of Managerial Psychology* 15, 3 (2000), 227–241.
[62] Sergio Luis Schmidt and Juliojulio Cesar Tolentino. 2018. DSM-5 criteria and depression severity: implications for clinical practice. *Frontiers in psychiatry* 9 (2018), 450.
[63] Rijurekha Sen, Youngki Lee, Kasthuri Jayarajah, Archan Misra, and Rajesh Krishna Balan. 2014. Grumon: Fast and accurate group monitoring for heterogeneous urban spaces. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems.* ACM, 46–60.
[64] Grant S Shields, Loren L Toussaint, and George M Slavich. 2016. Stress-related changes in personality: A longitudinal study of perceived stress and trait pessimism. *Journal of research in personality* 64 (2016), 61–68.
[65] Christopher Tennant. 2001. Work-related stress and depressive disorders. *Journal of psychosomatic research* 51, 5 (2001), 697–704.
[66] Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* ACM, 941–953.
[67] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM, 3–14.
[68] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 43.
[69] Shweta Ware, Chaoqun Yue, Reynaldo Morillo, Jin Lu, Chao Shang, Jayesh Kamath, Athanasios Bamis, Jinbo Bi, Alexander Russell, and Bing Wang. 2018. Large-scale Automatic Depression Screening Using Meta-data from WiFi Infrastructure. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 195.
[70] Sheryl L Warttig, Mark J Forshaw, Jane South, and Alan K White. 2013. New, normative, English-sample data for the short form perceived stress scale (PSS-4). *Journal of health psychology* 18, 12 (2013), 1617–1628.
[71] Camellia Zakaria, Kenneth Goh, Youngki Lee, and Rajesh Balan. 2019. Exploratory Analysis of Individuals' Mobility Patterns and Experienced Conflicts in Workgroups. In *Proceedings of the 5th ACM Workshop on Mobile Systems for*

RIGHTSLINK

Computational Social Science. ACM, 27–31.

[72] Zebra. 2019. *Location Technologies*. https://www.zebra.com/ap/en/products/location-technologies.html.

[73] Daqing Zhang, Bin Guo, and Zhiwen Yu. 2011. The emergence of social and community intelligence. *Computer* 44, 7 (2011), 21–28.

[74] Mengyu Zhou, Minghua Ma, Yangkun Zhang, Kaixin SuiA, Dan Pei, and Thomas Moscibroda. 2016. EDUM: classroom education measurements via large-scale WiFi networks. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 316–327.

Received April 2019; revised June 2019; accepted August 2019

RIGHTSLINK