Advancing the Understanding and Measurement of Workplace Stress in Remote Information Workers from Passive Sensors and Behavioral Data

Mehrab Bin Morshed[*1,2], Javier Hernandez[1], Daniel McDuff[1], Jina Suh[1], Esther Howe[1,3], Kael Rowan[1], Marah Abdin[1], Gonzalo Ramos[1], Tracy Tran[1], and Mary Czerwinski[1]

[1]Microsoft Research, Microsoft, Redmond, USA
[2]Georgia Institute of Technology, Atlanta, USA
[3]University of California, Berkeley, USA

***Abstract*—Workplace stress has been increasing in recent decades and has worsened by the unique demands imposed by COVID-19 and the new remote/hybrid work settings. High-stress working conditions can be detrimental to the health and wellness of workers and can lead to significant business costs in terms of productivity loss and medical expenses. An essential step toward managing stress involves finding comfortable ways to sense workers and recognizing stress as soon as it happens. This work explores the potential value of using pervasive sensors such as keyboards, webcams, and behavioral data such as calendar and e-mail activity to passively assess individual stress levels of work in real-life. In particular, we collected a large corpus of such data from 46 remote information workers over one month and asked them to self-report their stress levels and other relevant factors several times a day. Analysis of the data demonstrates that passive sensors can effectively detect both triggers and manifestations of workplace stress and that having access to prior data of the worker is critical for developing well-performing stress recognition models. Furthermore, we provide qualitative feedback capturing workers' preferences in workplace stress monitoring.**

***Index Terms*—Workplace Stress, Sensing, Emotion, Resilience, Modeling, Early Detection, Demands, Stressors, Resources**

I. Introduction

Stress is a significant and growing issue in our modern society. Prolonged high levels of stress have been shown to contribute to a wide variety of physical and psychological health issues such as high blood pressure [40], depression [30], mood disorders [2], and suicidal ideation [20]. One of the major sources of daily stress is workplace stress which can be defined as the reaction that people may experience when they are subject to high demands and pressures at work that do not correspond to their past experience and/or coping capabilities [5]. Some of the main contributing factors include, but are not limited to, juggling between professional and personal life, a perceived lack of job security, interpersonal issues with colleagues, and a high workload. When experienced over long periods, workplace stress has been shown to impair decision making, negatively affect productivity, and decrease job satisfaction as well as lead to significant business costs, which is approximately $300 billion per year in the U.S. alone [2].

[+]Corresponding author: mehrab.morshed@gatech.edu

Given the importance of workplace stress, researchers have explored the potential use of technology to help better manage it [5]. One popular method involves the delivery of stress-reduction interventions when someone is experiencing stress to reduce its potential harm [5]. This approach usually involves the development of real-time stress detection systems [15], [36] that create a *digital phenotype*[1] of the individual in order to infer their stress levels [38]. A large body of this work, however, requires instrumenting individuals with custom hardware such as wearables, which limits the potential scalability of the technology to a broad population. In addition, many of the studies have been performed in controlled laboratory environments [14], [27], which tend to offer limited generalization to real-life settings, or real-life settings are used, but different sample populations are considered (e.g., call-center employees [13], graduate students [35]) which may not necessarily generalize to remote information workers, as was our current focus.

Our work addresses these limitations and extends stress research by conducting a month-long, naturalistic, unconstrained, observational, and multimodal sensing study with 46 remote information workers at a large technology company. We leverage pervasive sensors such as keyboards and webcams, available to most remote information workers, and connect them with passively sensed behavioral data such as e-mail and calendar activity. In addition, we used the experience sampling method to collect ground truth self-reported measures of stress as well as other relevant factors, such as the level of demands and resources during the workday. We then used the collected data to address the following questions:

- RQ1: What is the *digital phenotype* of remote information workers' stress?
- RQ2: Can we accurately recognize self-reported stress from passively sensed data?
- RQ3: What are key end-user considerations when deploying stress sensing systems?

[1]In the context of precision medicine, digital phenotypes represent an individual's interactions with digital technologies (e.g., smartphones, wearable devices, etc.) that can generate a longitudinal health profile [38].

II. Related Work

While stress is more frequently associated with a negative response, specifically known as "distress" it is also well known that there are positive forms of stress, known as "eustress." In the context of work, distress might arise from a "toxic work environment, negative workload, isolation, number of hours worked, role conflicts, role ambiguity, lack of autonomy, career development barriers, difficult relationships with administrators and/or coworkers, managerial bullying, harassment, and organizational climate" [7]. Eustress is a "force that stimulates us to productively work through challenging situations and tasks" [7]. Examples might include a promotion, a successful project presentation, a deadline, or a positive but intense meeting.

While technology may not be the sole solution to helping people manage all of these sources of stress, researchers have explored its potential use to better understand and address some of the sources. For instance, studies of information workers have found that distractions can lead to higher reported stress and lower productivity [23], [24] and there are promising opportunities for technology to support workers' well-being through reflection [29], and interventions [33]. These solutions need to be designed carefully [41].

In a relevant study, Mark et al. [23], [24] presented a framework for how engagement and challenge at work were related to focus, boredom, stress, and rote work. Overall, they found more focused attention was present in the workplace than boredom. They also found that focus peaks in mid-morning and mid-afternoon, while boredom was highest in the morning. People were happiest doing "rote", or easy work, showing that focused work *can* involve stress. Their study was the first to show that rhythms of attentional states are associated with context and time, even in a dynamic workplace environment. A subsequent empirical study [25] in the workplace found, using physiological sensors (heart rate monitors), computer log data, and ethnographic methods, that stress (as measured by heart rate variability (HRV)) was lower when not using email. Both qualitative and quantitative data corroborated the stress findings around email use [26]. More recently, McDuff et al. [28] analyzed information workers' facial expressions longitudinally to reveal that passive sensors could pick up similar diurnal patterns in affective experience with displays of negative affect increasing monotonically on average over the course of the day.

In a separate effort, Lopez et al. [21] have looked at real-time automatic stress detection for information workers but within a controlled setting. Using physiological data gathered by an Empatica E4 wristband for registering EDA, they examined an arousal-based statistical approach, and they compared their stress detection model to self-reported stress in quiet office environments versus when their participants were exposed to different kinds of emotional triggers. Though they had some success with this approach to detecting stress, it was still not studied in a realistic, natural setting. Similarly, Ide et al. [16] utilized multiple physiological signals to predict stress in daily life. However, this research also used a laboratory method to induce stress in various ways. Using electrocardiogram, pulse wave, breathing rate, and skin temperature, the authors predicted four psychological states: relaxed, normal stress, monotonous stress, and nervous. They used the integration of nine physiological features identified as related to stress leading to 87% accuracy for stress detection and 63% accuracy for stress type.

In an extensive survey examining stress detection in daily life using mostly wearable sensors, Can et al. [5] reviewed the reported accuracy of various combinations of sensors across different environments, including office workplaces. They found that office environments provide a nice bridge between controlled laboratory settings and more mobile settings since office workers tend to be seated at their desks and quiet more often. In their review, research that employed EDA and HR had the highest performance in the office setting. They discussed problems with identifying key contextual features and the artifacts that result from physical movement. Finally, the authors discussed the problems involved in getting subjective ground truth from users who might exhibit the same physiological markers but rate their stress levels in dissimilar ways. We acknowledge these challenging issues and will also need to address them in our research.

III. Methodology

A. Study Design and Data Collection

The study duration was four weeks, in which participants had to install a data logging software and respond to several surveys focused on stress and other relevant factors. In the following, we describe the surveys gathered at various stages of the study and the data logging tool. The study was reviewed and pre-approved by the institutional Ethics Review Board.

1) Surveys Instruments: During the study, participants responded to various surveys that can be grouped into five main categories based on their delivery time.

Study Intake. At the beginning of the study, we gathered baseline information from our participants about their demographics (e.g., age, gender, job type) and baseline mental well-being. In particular, we collected validated surveys such as the DASS-21 [22] that captures stress, depression, and anxiety, and the Perceived Stress Scale [6].

Experience Sampling. During each workday, participants received multiple prompts (around every hour $\pm$ 15 min) containing several 5-Likert scale questions to rate the level of perceived work demands and resources [9], their valence and arousal levels [4], [34], and their stress levels during the 30 minutes preceding the prompt. Participants were given the following options for demands and resources: *very low, low, moderate, high, and very high.* These definitions and options were consistent with the previous literature looking at workplace demands and resources [9]. For valence and arousal, participants were provided with the following options: *very unpleasant, unpleasant, neutral, pleasant, and very pleasant* and *very low, low, moderate, high, and very high* respectively. For stress, participants were provided with the following options: *not at all, slightly stressed, moderately stressed, very stressed, and extremely stressed.*

Daily Check-In. At the beginning of each workday, participants were asked to answer questions about their previous night of sleep. In particular, they were asked to report the time when they started to try to fall asleep, the time they got out of bed, the number and duration of awakenings, and the overall quality of sleep as it has been shown to influence stress [18].

Daily Check-Out. At the end of each workday, we asked participants to rate their daily stress, valence, arousal, demands, and resources with the same 5-Likert scale questions described above. In addition, participants were asked to report their food and caffeinated drink intake episodes during the workday as having a stressful workday might often lead to more snacking [39], irregular meal patterns [31], and drinking more caffeinated drinks [8], among other things. Finally, participants were asked to indicate the presence and potential intensity of the following stressors: 1) a high pace workday, 2) too many meetings, 3) too much emails, 4) overly packed day, 5) too many ongoing activities, 6) sitting for too long, 7) lack of breaks, 8) missing exercise due to work/personal life demands, 9) loss of sleep due to longer working hours or deadlines, and 10) unable to separate work life demands. The answers ranged from 0 (did not apply today) to 5 (applied with a lot of impact). These stressors emerged as being some of the most relevant ones in an exploratory survey.

End of Study. At the end of the study, participants shared their potential expectations when having a stress sensing system deployed on their work machine. In particular, participants answered the following questions: 1) what would be your comfort level with different sensing modalities (e.g., wearable device, computer usage, webcam, etc.)? and 2) how would you prefer your data to be stored (e.g., local vs. cloud)?

2) Passive Sensing: To capture the digital manifestations of stress, we developed a custom multimodal logging software that recorded information about the participants' activities, behaviors, and physiological states. The main components are as follows:

Email, Calendar, and Application Usage Information. The software ran on the participant's desktop computer and logged the number of emails received in their inbox, number of calendar appointments, and applications used (including when they were opened and closed, in the foreground, etc.). We logged events that were informative about an individual's stress level based on prior work. For example, we know that email is one of the most significant signals of work-related communications for information worker [25]. In addition, calendar information contains vital workday-related information for information workers. Especially during COVID-19, the frequency of remote meetings has increased significantly, and individuals spend more time in meetings, often leading to stress and fatigue [1]. Application usage, keyboard, and mouse activity are direct proxies of how much interaction an information worker has with their work environment. They are often investigated in the stress literature as meaningful signals for detecting the stress levels of participants [14].

Face position and Facial Action Units. Based on the relevance of facial expressions in the context of emotion understanding, we captured a participant's facial position and facial action units [11]. In particular, the software used a pipeline to process the video frames in real-time (i.e., without storing video frames in the cloud for privacy) at 1 frame per second. Using a convolutional neural network (CNN) facial detector, it extracted the bounding box corresponding to the user's face[2] and then processed this region of interest using another CNN model to extract the probabilities of 12 facial action units (AU01, AU02, AU04, AU05, AU06, AU09, AU12, AU15, AU17, AU20, AU25, and AU26) [12], based on the standard Facial Action Coding System (FACS) [11]. These FAUs were selected as they are generally most frequently observed and most accurate at predicting projected emotional actions. These actions are associated with expressions of both positive (e.g., AU12/zygomatic major/smiling) and negative (e.g., AU04/corrugator/brow furrowing) effects. It is important to note, however, that no action unit maps uniquely to anyone's expression or emotional state, but they may still capture a rich array of users' behaviors [32].

Non-Contact Physiological Sensing. Physiological signals usually sensed with contact-based sensors have been frequently used for monitoring stress but they usually require wearing additional devices, which can be cumbersome and socially stigmatizing [5]. To minimize these challenges while maximizing the benefits, we leveraged the previous computer vision pipeline to extract heart rate, inter-beat intervals, and breathing rate from subtle color and motion changes of the face using a non-contact video-based approach [19].

B. Analysis

To help capture both stressors and potential manifestations of stress, we extracted nine different types of features from the different modalities (see Table I). As can be seen, two of the feature groups were extracted from survey data, and the rest from passively sensed information. While the original goal was to develop a purely passive sensing method, we decided to include sleep and eating/drinking habits as these were previously found to be relevant in the context of stress and could help inform future sensing efforts.

In terms of ground truth, this work considers workplace stress at two temporal granularities: 1) instantaneous stress, which was experienced during a period of 30 minutes and was self-reported during experience sampling (N: 4747), and 2) daily stress, which was experienced during the workday and self-reported during the daily check out (N: 803). The features were extracted at these two temporal resolutions to model momentary and daily stress. On some occasions, however, a particular feature could not be computed due to missing data (e.g., camera not working) so we had to implement a strategy to impute the features. If a participant had at least 3 days of data for a missing data stream, we imputed the missing data with the median values of the feature of the participant. If the participant did not have at least 3 days of data, we imputed the missing features by taking the median of all participants.

For the machine learning analysis, we treated the problem as a binary classification problem with low and high-stress levels. To ensure a balanced split within each individual, we used an adaptive baseline approach that separated the samples of each person in a way that would minimize the difference in sample size across the two classes. In terms of models, we explored the predictive performance of different classifiers that are commonly used in the context of stress prediction [5].

[2]https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB

TABLE I
FEATURES EXTRACTED FROM EACH DATA MODALITY.

Signal Modality	Type	Features
Sleep	Survey	Self-reported sleep quality, time participants went to bed, time participants tried to fall asleep, number of awakenings during the sleep, time participants got out of bed, total sleep time, and difference of total sleep time compared to the mean sleep time for each participant.
Eating and Drinking	Survey	Total number of eating events, total number of caffeinated drinking episodes, difference between breakfast and lunch time compared to the average breakfast and lunch time within individuals.
Email	Passive	Total number of unique email threads, average number of CC'ed contacts, and total number of emails that an individual received until the time that they reported stress.
Calendar	Passive	Total number of various meetings, total allocated time (in minutes) for different meetings that an individual had on their calendar up to the time they provided the stress report, total count and total duration (in minutes) for accepted meetings, number of cancelled meetings, number of tentative meetings, number of self-meetings, and number of recurring meetings.
Application Usage, Keyboard, and Mouse	Passive	Total number of minutes, number of different computer applications were running in the foreground, total number of minutes different applications were being actively used by the participant, total number of key press events from the keyboard, total number of mouse clicks, and total number of mouse wheel rotations.
Physiological Sensing	Passive	Mean, median, and standard deviation of Root mean squared difference between successive inter-beat intervals (RMSSD), beats per minute, and breaths per minute.
Facial Action Units	Passive	Mean, max, median, and standard deviation of action units AU01, AU02, AU04, AU05, AU06, AU09, AU12, AU15, AU17, AU20, AU25, and AU26.
Day Specific Features	Passive	Hour of the day and day of the week.
Face Position Features	Passive	Standard deviation of face rectangle's length, width, and standard deviation of top and left coordinates.

In particular, we evaluated Random Forest (RF), Multilayer Perceptron (MLP), Gradient Boosting (XGB), Support Vector Machines with RBF Kernel (SVM), and Ridge Classifier (RC). For a baseline comparison, we included two other classifiers: 1) a *Majority* classifier that always predicted the majority class based on the training set, and 2) a *Random* classifier that made random predictions.

C. Participant Recruitment and Compensation

The study participants were recruited through an email sent to a random set of information workers at a large technology company. As part of the eligibility criteria, participants needed to mostly use a single work computer, avoid working in virtual machines, be connected to the Internet, have a webcam, and use Outlook desktop as their default email and calendar software. In addition, their computer had to be able to run the custom sensing software that ran on Windows OS and used around 2 GB RAM.

We recruited a total of 50 information workers. However, one participant dropped out during the second week of the study due to software problems. In addition, we had to exclude three participants from the analysis as they showed little variance in their self-reports ($<$ 2 std), indicating that they provided similar responses most of the time. Considering the final set of participants, 25 self-identified as male, 19 as female, and 2 as non-binary/gender diverse. The large majority of participants (71.74%) reported being in the age range of 26-45 years old. The majority of participants (60%) described their job function to be in engineering and development, but there were also participants working as administrative assistants, sales, and human resources. Finally, the majority of participants (71.74%) reported not having any direct reports. Each participant could receive up to $300 in the form of a gift card. To promote completion of the study, we provided a scalable monetary reward in which each participant received $50 after the 1st week, $50 after the 2nd week, $100 after the 3rd week, and $100 after the 4th week.

IV. RESULTS

A. RQ1: What is the digital phenotype of remote information workers' stress?

To better understand workplace stress and how it manifested in the collected data, we separately considered the two types of stress (instantaneous and daily) and their correlation with different variables. In particular, we used Pearson's correlation coefficient, reported correlation coefficients, and Bonferroni-adjusted p-values.

1) How does instantaneous stress manifest on the data?: Considering the experience sampling responses, we found that self-reported stress was positively correlated with demands (r=0.55, p$<$0.01) and negatively correlated with resources (r=-0.19, p$<$0.01), which is consistent with prior work examining these factors in the context of workplace stress [3]. Self-reported stress was also negatively correlated with valence (r=-0.33, p$<$0.01) and less strongly but still significant with arousal (r=-0.13, p$<$0.01). Self-reported level of demands was negatively correlated with valence (r=-0.14, p$<$0.01) and positively correlated with arousal (r=0.11, p$<$0.01), suggesting that a high level of demands was generally associated with negative and aroused feelings. Self-reported level of resources was positively correlated with valence (r=0.31, p$<$0.01) and arousal (r=0.41, p$<$0.01), suggesting that a high level of resources was associated with positive and arousing feelings.

We found that lower quality of sleep was negatively correlated (r=-0.22, p$<$0.01) with higher stress levels. Late bedtime was associated (r=0.07, p$<$0.01) with a higher stress level.

These observations are aligned with previous literature on stress [17]. More keyboard activity and less facial movement were positively correlated with stress (r=0.05, p<0.05 and r=0.09, p<0.05, respectively), which could be interpreted as a proxy for high work demands and the need for sustained attention for long periods. AU06 or cheek raiser was negatively (r=-0.10, p<0.01) correlated with stress, indicating that our participants seem to smile less when stressed. On the other hand, AU07 or "lid tightener" was positively correlated (r=0.08, p<0.01) with stress, which may similarly indicate prolonged sustained attention.

2) How does daily stress manifest on the data?: When considering the stress reported at the end of the workday, we found that the level of demands was also significantly and positively (r=0.57, p<0.01) correlated with stress. In contrast, valence (r=-0.42, p<0.01) and resources (r=-0.42, p<0.01) were negatively and significantly correlated with stress. We did not find a significant correlation between arousal and daily stress levels.

When considering the broader range of signals, we found that the average time spent on different applications (r=0.11, p<0.01) and the number of open applications (r=0.13, p<0.01) were the only two features that were positively and significantly correlated with stress. In addition, as part of the daily check-out survey, participants also reported the occurrence of ten pre-defined stressors.

High pace workday, was positively correlated with average time spent on open applications (r=0.29, p<0.001), total number of busy slots (r=0.16, p<0.001), total duration of busy slots (r=0.16, p<0.05), total number of meetings (r=0.15, p<0.05), key press count (r=0.24, p<0.001), keyboard events (r=0.27, 0<0.001), mouse events (r=0.19, p<0.001), and total number of open applications (r=0.29, p<0.001). High values of these features indicate that individuals were spending more time on their work machines, suggesting that these signals represent high pace workday stressors.

Too many meetings, was positively correlated with number of busy slots (r=0.26, p<0.001), busy slot duration (r=0.25, p<0.001), total number of meetings (r=0.25, p<0.001), and median AU12 (r=0.18, p<0.005). The results are intuitive in that the number of busy slots, total busy slot duration, and total meeting count were a proxy for how many meetings an individual had on a given workday. In addition, individuals were probably more likely to smile during meetings (as they are social) which may be shown on AU12.

Too many emails, was positively correlated with median AU12 (r=0.13, p<0.05) and negatively correlated with mean AU12 (r=-0.15, p<0.05) and standard deviation of AU14 (r=-0.16, p<0.05), These signals may be indicative of prolonged sustained attention when reading and responding to emails.

Overly packed day, was positively correlated with number of busy slots (r=0.18, p<0.001), busy slot duration (r=0.18, p<0.001), total number of meetings (r=0.17, p<0.001), and median AU12 (r=0.14, p<0.001). These features can be indicative of high application usage, a high number of meetings, and more smiling gestures (AU12) associated with social interactions, respectively.

Too many ongoing activities, was positively correlated with the average time spent in different applications (r=0.20, p<0.001) and the number of open applications (r=0.20, p<0.05). In addition, it was negatively correlated with mean AU06 (r=-0.03, p<0.01).

Sitting for too long, was positively correlated with the total # of mouse events (r=0.15, p<0.05). However, sitting for too long was also negatively correlated with mean AU01 (r=-0.17, p<0.05), mean AU02 (r=-0.11, p<0.001), mean AU06 (r=-0.19, p<0.001), the standard deviation of AU01 (r=-0.14, p<0.001), the standard deviation of AU02 (r=-0.11, p<0.001), and the standard deviation of AU06 (r=-0.13, p<0.05). AU01 is the facial action unit for the inner brow raiser, AU02 is the facial action unit of the outer brow raiser, and AU06 corresponds to the cheek raiser, which may be indicative of body fatigue.

Lack of breaks, was positively correlated with total # of keyboard events (r=0.19, p<0.05) and total # of keyboard events (r=0.16, p<0.05), which are proxies of prolonged interaction with the work machine. Hence, if participants are spending more time on their work computers, they are likely to get less time for taking breaks. There was also a negative correlation with AU06 (r=-0.10, p<0.05), which may similarly indicate fatigue due to the lack of breaks.

Missing exercise due to work/personal life, was positively correlated with mean AU04 (r=0.07, p<0.05), mean AU07 (r=0.18, p<0.05), and standard deviation of AU07 (r=0.22, p<0.01). AU04 is a proxy for brow lowering facial action, and AU07 is a proxy for eyelid tightening. AU07 is also a proxy for the smiling gesture, which might mean that participants were having frequent meetings that day.

Loss of sleep due to longer working hours or deadlines, was positively correlated with mean AU04 (r=0.18, p<0.001), mean AU07 (r=0.15, p<0.001), median AU04 (r=0.15, p<0.05), median AU07 (r=0.15, p<0.05), standard deviation of AU04 (r=0.20, p<0.001), and standard deviation of AU07 (r=0.15, p<0.05). Loss of sleep was negatively correlated with the standard deviation of face height (r=-0.25, p<0.001) and the standard deviation of face width (r=-0.25, p<0.001). These may be indicative of closed eyes due to tiredness.

Unable to separate work and life demands, was positively correlated with mean AU04 (r=0.15, p<0.001) and standard deviation of AU04 (r=0.18, p<0.001). It was negatively correlated with the standard deviation of AU02 (r=-0.08, p<0.05), the standard deviation of face width (r=-0.15, p<0.05), and the standard deviation of face height (r=-0.15, p<0.05). These findings may indicate that our participants were more stationary in front of their cameras when experiencing this stressor.

B. RQ2: Can we accurately recognize self-reported stress from passively sensed data?

1) Can we detect instantaneous and daily stress levels of remote information workers?: To assess the possibility of recognizing self-reported stress, we conducted a 10-fold cross-validation approach in which the entire dataset was split into 10 separate folds. These folds were then iteratively used to generate 10 models. While the data from the training and

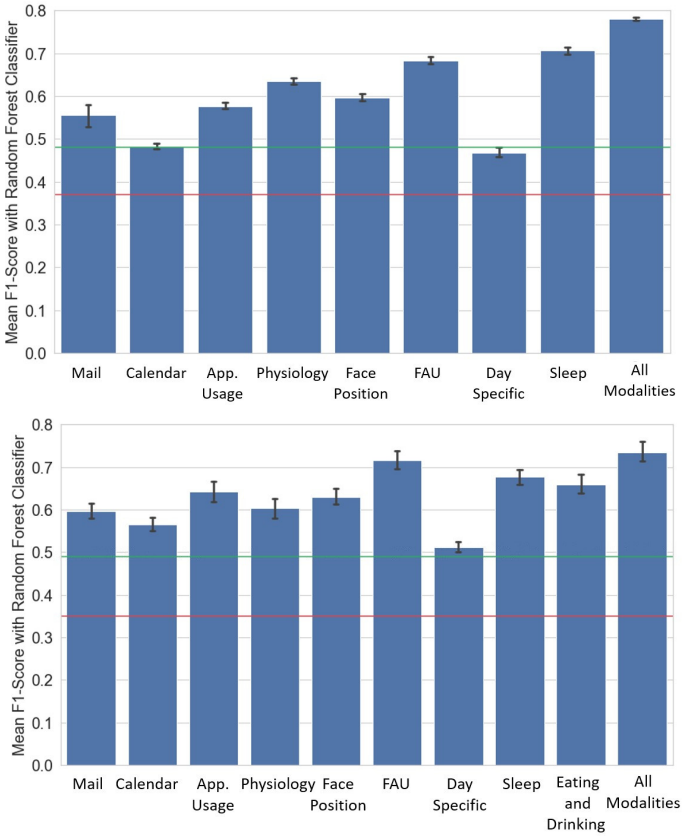

Fig. 1. Average F1-scores and their standard errors when considering different modalities in for instantaneous (top) and daily stress (bottom) predictions. Red and green horizontal lines represent the Majority and Random baseline classifiers, respectively. FAU: Facial action units. App. usage: application usage.

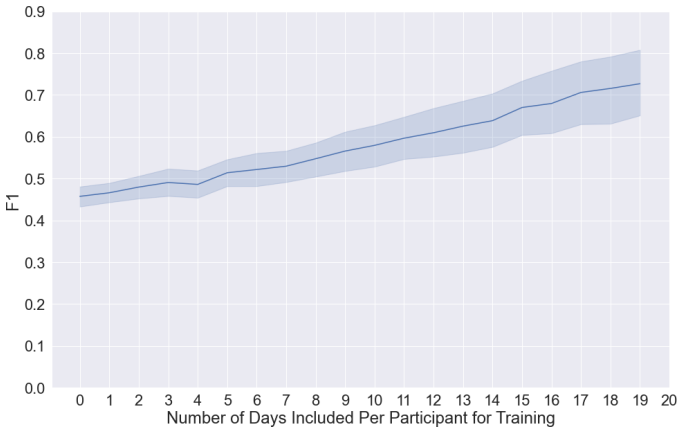

Fig. 2. Average performance across participants when varying the number of training days for instantaneous stress prediction.

testing sets were always different, it is important to note that data from the same user may be in both training and testing sets simultaneously (a.k.a., person-dependent models). Random Forest and Gradient Boosting outperformed other classifiers in terms of average F1-score, which was consistent when considering daily stress. In particular, Random Forest reached a performance of 78% when predicting instantaneous stress (rightmost bar in top Figure 1) and 73% when predicting daily stress (rightmost bar in bottom Figure 1). For simplicity, we report Random Forest results in the rest of the analysis.

2) What is the predictive value of each sensing modality for instantaneous and daily stress prediction?: Figure 1 (top) shows the average performance across separate modalities when predicting instantaneous stress. As can be seen, sleep-based features yielded the best F1-score performance (72%), which was followed by those associated with facial action units (68%), camera-based physiological sensing (63%), mail (58%), application usage (57%), calendar (48%) and day specific features (46%). The Majority and Random baseline classifiers yielded a performance of 37% and 48%, respectively.

Similarly, Figure 1 (bottom) shows the average performance across separate modalities when predicting daily stress. In this case, features associated with facial action units yielded the best F1-score performance (71.52%), which was followed by those associated with sleep (67.73%), eating and drinking (65.95%), application usage (64.20%), camera-based physiological sensing (60.39%), mail (59.63%), calendar (55.63%), and day specific features (51.23%). In this case, the Majority and Random baseline classifiers yield performance of 35% and 49%, respectively.

3) How much data do we need to personalize stress detection models?: One difficult yet important challenge in the stress detection literature is the ability to generalize to unseen participants [38] (a.k.a., person-independent models). In our case, conducting a leave-one-out protocol to predict instantaneous stress yield an F1-score performance of 46% (leftmost score of Figure 2), which is significantly lower than the results reported in the previous section.

Hence, we wanted to investigate how much data from a particular participant would be needed to ensure good performance. To do so, we iteratively added an increasing number of days from a particular participant as part of the training set and used the generated models to make predictions in the following days. Figure 2 shows the average F1-score performance and their standard deviation for a different number of days. As can be seen, the average performance monotonically increases with the amount of data considered as part of the training data. For instance, with 15 days of training from each individual (3 weeks), the model achieved a performance of 68% F1-score on the remaining week.

C. RQ3: What are key end-user considerations when deploying stress sensing systems?

At the end of the study, participants provided their preferences in relation to having a stress sensing system at work. We summarize some of the main findings in this section.

1) What sensing modalities are preferred by the participants?: Participants were more comfortable sharing their keyboard and mouse activity, followed by computer usage, wearable device data, microphone, smartphone, and webcam. Keyboard and mouse activity were significantly higher than any other (Wilcoxon signed-rank test, $p<0.05$), and computer usage was significantly higher than the microphone, smartphone, and webcam modalities ($p<0.01$). However, it is important to note that all the ratings were above the average, indicating that

our studied population could potentially accept all of them. Such insights provide a direct implication for keeping user preferences in mind while developing automated and multimodal systems that can be deployed to detect employees' stress levels continuously.

2) What kind of storage (e.g., local and cloud) do participants prefer for storing their stress-related data?: To investigate if participants have any preference for where the stress-related features and inferred stress scores should be stored, we asked participants to identify their comfort levels if their data were stored locally on their computers and/or in the cloud. For both options, participants could identify their preferences using a 5-point Likert scale, ranging from very comfortable (option 1) to very uncomfortable (option 5). The median response for local storage was "somewhat comfortable" (option 4), and the median response for cloud storage was "neither comfortable nor uncomfortable" (option 3), which were close to being significantly different (Wilcoxon signed-rank test, p: 0.0576). This indicates that participants felt more comfortable with the local option, probably due to privacy concerns and the ability to control the information more easily.

V. Discussion

Our work is focused on understanding workplace stress through passive sensing technologies. Based on our findings, we discuss potential implications for future workplace stress sensing systems and research.

1) Identify stressors and their digital phenotypes to mitigate stress.: As part of our analysis, we found several signals correlated with frequent workplace stressors. For example, a highly paced workday is strongly correlated with a greater amount of computer activity (e.g., application usage, keyboard, and mouse usage) and leads to higher stress. Our choice of stressors to investigate was informed by empirical data from our target population. However, these stressors might not be as pertinent to other target populations. Because stress is a response to a variety of contextual factors that a worker is situated in [16], thus, such response varies by individuals, it is important to identify appropriate stressors and their digital traits to inform the design of targeted interventions to mitigate stress. Therefore, future research on understanding workplace stressors should consider context- and population-specific stressors.

2) Understand workplace demands and available resources to design person-specific interventions.: We found that high workplace demand was positively correlated with stress and negatively correlated with the number of available resources for both instantaneous and daily stress. Workplace demands could be physical, psychological, social, or organizational aspects of the job that require sustained physical and/or psychological (cognitive and emotional) effort to cope. Available resources could range from physical, psychological, social, or organizational aspects of life. Note that having high demands at work may not be a negative experience or lead to higher stress. Understanding what individuals perceive as high demands that are negatively affecting them and what specific resources they need to meet those demands or to cope with the stress generated by them is necessary to design targeted interventions that address specific resource needs.

3) Take personalization into account when building stress models.: We found that our one-size-fits-all stress models did not generalize to completely unobserved participants (Section IV-B3). On the other hand, we found that personalized models (i.e., models built using within-person data) improved the stress prediction within our dataset, which aligns with prior research findings [38]. These findings suggest that a one-size-fits-all model may not work for accurate stress detection in the workplace, possibly because stress manifests differently for different people. Building personalized models can be challenging to bootstrap because of a low volume of data generated by a single person (compared to a population of hundreds of workers). In our analysis, we found that by using data for 3 weeks or 15 working days, our stress prediction model was able to detect stress levels with an F1-score of 68%. We also find that adding one additional week of data (i.e., four weeks of data collection) achieves an even higher F1-score (74%). However, this improved model performance comes at the cost of potentially disrupting the users' work with frequent Experience Sampling questions required for training data collection. This implies that systems that train and leverage personalized stress sensing models should consider the upfront cold-start costs and temporal performance variability.

4) Privacy and Ethics: User privacy is a major concern with any application that captures personal information. Privacy in work-related stress sensing is even more sensitive since, in a toxic work environment, work stress-related concerns can be stigmatized [10], [37]. Hence, protecting users' privacy of stress sensing systems and their tracked stress-related data is critical and must be well regulated within respective organizations. Note that for inferring stress, we used high-level activity data from each participant (e.g., the total number of emails in a given window, the total number of minutes in meetings) instead of low-level, sensitive data (e.g., email text). Such high-level activity data pose relatively few privacy challenges, especially in contrast to asking someone about specific kinds of email content that they found stressful. We also used webcams that constantly collected data from participants. Although the features we extracted from webcams can be decoupled from individuals' identity (e.g., facial action units, face position), physiological data may be considered sensitive and constant monitoring can lead to discomfort, as we have observed from our participants. Irrespective of the granularity and anonymity of such data, system designers, data privacy officers, and organizations should establish strong regulations to protect the workers' privacy and data while respecting their preferences for comfort level, storage, and retention (Section IV-C2).

VI. Conclusion

Leveraging a multimodal data logging platform, this work evaluated the potential use of different signals in a 4-week study that included 46 information workers performing their regular work in situ. The findings of our study help advance our understanding of workplace stress and its potential digital manifestations. In addition, we demonstrate that these signals can be automatically detected and used for predictions of stress with machine learning.

REFERENCES

[1] Andrew A Bennett, Emily D Campion, Kathleen R Keeler, and Sheila K Keener. Videoconference fatigue? exploring changes in fatigue after videoconference meetings during covid-19. *Journal of Applied Psychology*, 106(3):330, 2021.

[2] Douglas W Billings, Royer F Cook, April Hendrickson, and David C Dove. A web-based approach to managing stress and mood disorders in the workforce. *Journal of occupational and environmental medicine/American College of Occupational and Environmental Medicine*, 50(8):960, 2008.

[3] J Blascovich and W B Mendes. Challenge and threat appraisals: The role of affective cues. In Joseph P Forgas, editor, *Feeling and thinking: The role of affect in social cognition. Studies in emotion and social interaction, second series*, pages 59–82. Cambridge University Press, 2000.

[4] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[5] Yekta Said Can, Bert Arnrich, and Cem Ersoy. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics*, 92:103139, 2019.

[6] Sheldon Cohen, Tom Kamarck, Robin Mermelstein, et al. Perceived stress scale. *Measuring stress: A guide for health and social scientists*, 10(2):1–2, 1994.

[7] Thomas W Colligan and Eileen M Higgins. Workplace stress: Etiology and consequences. *Journal of workplace behavioral health*, 21(2):89–97, 2006.

[8] Terry L Conway, Ross R Vickers Jr, Harold W Ward, and Richard H Rahe. Occupational stress and variation in cigarette, coffee, and alcohol consumption. *Journal of health and social behavior*, pages 155–165, 1981.

[9] Evangelia Demerouti, Arnold B Bakker, Friedhelm Nachreiner, and Wilmar B Schaufeli. The job demands-resources model of burnout. *Journal of Applied psychology*, 86(3):499, 2001.

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

[11] Rosenberg Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[12] Alberto Fung and Daniel McDuff. A scalable approach for facial action unit classifier training usingnoisy data for pre-training. *arXiv preprint arXiv:1911.05946*, 2019.

[13] Javier Hernandez, Rob R. Morris, and Rosalind W. Picard. Call center stress recognition with person-specific models. In *Affective Computing and Intelligent Interaction*, pages 125–134, 2011.

[14] Javier Hernandez, Pablo Paredes, Asta Roseway, and Mary Czerwinski. Under pressure: sensing stress of computer users. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 51–60, 2014.

[15] Esther Howe, Jina Suh, Mehrab Bin Morshed, Daniel McDuff, et al. Design of digital workplace stress-reduction intervention systems: Effects of intervention type and timing. In *CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022.

[16] Hirohito Ide, Guillaume Lopez, Masaki Shuzo, Shunji Mitsuyoshi, Jean-Jacques Delaunay, and Ichiro Yamada. Workplace stress estimation method based on multivariate analysis of physiological indices. In *HEALTHINF*, pages 53–60, 2012.

[17] Takayuki Kageyama, Noriko Nishikido, Toshio Kobayashi, et al. Self-reported sleep quality, job stress, and daytime autonomic activities assessed in terms of short-term heart rate variability among male white-collar workers. *Industrial health*, 36(3):263–272, 1998.

[18] Hannah K Knudsen, Lori J Ducharme, and Paul M Roman. Job stress and poor sleep quality: data from an american sample of full-time workers. *Social science & medicine*, 64(10):1997–2007, 2007.

[19] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *NeurIPS*, 2020.

[20] Adrian Loerbroks, Sung-Il Cho, Maureen F Dollard, Jianfang Zou, et al. Associations between work stress and suicidal ideation: Individual-participant data from six cross-sectional studies. *Journal of psychosomatic research*, 90:62–69, 2016.

[21] Franci Suni Lopez, Nelly Condori-Fernandez, and Alejandro Catala. Towards real-time automatic stress detection for office workplaces. In *Annual International Symposium on Information Management and Big Data*, pages 273–288. Springer, 2018.

[22] Peter F Lovibond and Sydney H Lovibond. The structure of negative emotional states: Comparison of the depression anxiety stress scales (dass) with the beck depression and anxiety inventories. *Behaviour research and therapy*, 33(3):335–343, 1995.

[23] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. Focused, aroused, but so distractible: Temporal perspectives on multitasking and communications. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 903–916. ACM, 2015.

[24] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, and Paul Johns. Bored mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3025–3034. ACM, 2014.

[25] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, Paul Johns, et al. Email duration, batching and self-interruption: Patterns of email use on productivity and stress. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 1717–1728, 2016.

[26] Gloria Mark, Stephen Voida, and Armand Cardello. " a pace not dictated by electrons" an empirical study of work without email. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 555–564, 2012.

[27] Daniel McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W. Picard. COGCAM: Contact-free Measurement of Cognitive Stress During Computer Tasks with a Digital Camera. In *Conference on Human Factors in Computing Systems*, pages 4000–4004, New York, New York, USA, 2016. ACM Press.

[28] Daniel McDuff, Eunice Jun, Kael Rowan, and Mary Czerwinski. Longitudinal observational evidence of the impact of emotion regulation strategies on affective expression. *IEEE Transactions on Affective Computing*, 2019.

[29] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. Affectaura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 849–858, 2012.

[30] Yoshio Mino, Akira Babazono, Toshihide Tsuda, and Nobufumi Yasuda. Can stress management at the workplace prevent depression? a randomized controlled trial. *Psychotherapy and psychosomatics*, 75(3):177–182, 2006.

[31] Mehrab Bin Morshed, Samruddhi Shreeram Kulkarni, Koustuv Saha, Richard Li, Leah G Roper, Lama Nachman, Hong Lu, et al. Food, mood, context: Examining college students' eating context and mental well-being. *ACM Transactions on Computing for Healthcare*, 2022.

[32] Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. Affectivespotlight: Facilitating the communication of affective responses from audience members during online presentations. *Conference on Human Factors in Computing Systems - Proceedings*, may 2021.

[33] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, et al. Poptherapy: Coping with stress through pop-culture. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, pages 109–117, 2014.

[34] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[35] Akane Sano, Andrew J. Phillips, Amy Z. Yu, Andrew W. McHill, Sara Taylor, et al. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *Wearable and Implantable Body Sensor Networks*, pages 1–6, jun 2015.

[36] Jessica Schroeder, Jina Suh, Chelsey Wilks, Mary Czerwinski, Sean A Munson, James Fogarty, and Tim Althoff. Data-driven implications for translating evidence-based psychotherapies into technology-delivered interventions. In *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 274–287, 2020.

[37] Natasha A Schvey, Rebecca M Puhl, and Kelly D Brownell. The stress of stigma: exploring the effect of weight stigma on cortisol reactivity. *Psychosomatic medicine*, 76(2):156–162, 2014.

[38] Elena Smets, Emmanuel Rios Velazquez, Giuseppina Schiavone, Imen Chakroun, Ellie D'Hondt, et al. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ digital medicine*, 1(1):1–10, 2018.

[39] Sabine Sonnentag, Alexander Pundt, and Laura Venz. Distal and proximal predictors of snacking at work: A daily-survey study. *Journal of Applied Psychology*, 102(2):151, 2017.

[40] Tanja GM Vrijkotte, Lorenz JP Van Doornen, and Eco JC De Geus. Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability. *Hypertension*, 35(4):880–886, 2000.

[41] Eoin Whelan, Daniel McDuff, Rob Gleasure, and Jan Vom Brocke. How emotion-sensing technology can reshape the workplace. *MIT Sloan Management Review*, 59(3):7–10, 2018.