Designing Opportune Stress Intervention Delivery Timing using Multi-modal Data

Akane Sano
Massachusetts Institute of Technology
Cambridge, MA, USA
akanes@media.mit.edu

Paul Johns
Microsoft Research
Redmond, WA, USA
Paul.Johns@microsoft.com

Mary Czerwinski
Microsoft Research
Redmond, WA, USA
marycz@microsoft.com

***Abstract*—This paper describes a micro-stress intervention system for information office workers in the workplace, their responses to the interventions and machine learning models to predict the most opportune timing for providing the interventions. We studied 30 office workers for 10 days and examined their work patterns by monitoring their computer and application usage, sleep, activity, heart rate and its variability, as well as the history of micro-stress interventions provided through our desktop software. We analyzed temporal patterns of stress intervention acceptance/rejection and the relationships between their subjective and objective responses to the interventions and perceived work engagement, challenge and stress levels. We then developed machine learning models to predict better stress intervention delivery timing based on this multi-modal data. We found that features from computer and application usage, activity, heart rate variability and stress intervention history showed up to 80.0% accuracy in predicting good or bad intervention timing using a multi-kernel support vector machine algorithm. These findings could help practitioners design the most effective, just-in-time, closed-loop, stress interventions. To our knowledge, this is one of the first papers to review opportune stress interventions' delivery timing research, which could have a big influence in designing stress intervention technologies.**

1. Introduction

Can we design a system to provide stress management advice with minimal interruption to office workers in the workplace? With IT technology advances, work productivity has increased; on the other hand, high negative stress in the workplace has been a serious problem. The continuous connection to our computing devices and the high demands for responding to emails/instant messengers promptly has added to perceived workplace stress.

When we design stress interventions, there are at least 3 factors we need to consider. First, one needs to consider what types of interventions are provided (content). Secondly, we need to consider how the intervention is delivered (i.e., modality). Finally, we need to consider when a system provides the intervention for maximal impact and effectiveness (e.g., timing). Even when a designer has appropriated the most effective interventions, presenting them at the most appropriate times could have a significantly positive benefit regarding effectiveness.

Just-In-Time Adaptive Interventions (JITAIs) have been investigated to support people while they manage their individual, daily stress experience [1]. It has been mentioned that, "A JITAI can be used to (a) remind people to engage in stress management techniques as they experience stress, (b) help people better identify and address emotionally laden situations as they occur, in their natural environment; and (c) support long-term learning of stress-management".

Providing stress interventions while users are experiencing high stress situations might not be a good strategy. For instance, users might be working on important tasks for which they do not want to be interrupted. Interruptions on IT workers' performance and emotion has been investigated in several HCI studies. Mark et al. investigated interruption at work and found that interruptions change work patterns and could add stress and frustration [2]. Czerwinski et al. studied responses to instant messenger (IM) notifications and found that the delays associated with an IM disruption depend on the computer tasks one is engaged in [3]. They found that a good time for notifications was early in the task, before the user became deeply engaged in the task goal; and that evaluation, planning, and execution phases in tasks were the most disruptive.

These findings show that designing the right timing of notifications is very important to provide less damaging interruptions; however, when it comes to stress interventions, the interruption that occurs while users were in the early phase of tasks could also make users more aware of the benefits of taking a break. Providing stress interventions with inappropriate timings could annoy users and increase their stress levels. However, long-term, continuous computer usage without such advice could lead to adverse health outcomes and reduce attention and performance. As described in the previous paragraph, JITAI could play a role in reminding people of how to manage their stress. Stress intervention with personalized and optimal timing could increase users' productivity and health.

In this paper, we approached this problem by answering the following research questions.

- RQ1 a) How do people respond to stress interventions over the course of the day?

978-1-5386-0563-9/17/$31.00

- RQ1 b) Does it relate to their emotion, stress, work engagement, and challenges?
- RQ1 c) Are stress interventions provided at opportune timing more effective in reducing stress?
- RQ2 a) With multi-modal data from information workers, how accurately can we predict better stress intervention delivery timing? Which signals can best predict that delivery timing?

To answer these questions, we (1) developed a system that provides micro-stress interventions to office workers in the office environment, (2) collected multi-modal, physiological and behavioral data from 30 office workers through a 10 business day study, (3) analyzed the data to understand their stress profiles and responses to micro-stress interventions presented on their computers and then (4) designed models to predict the optimal time to provide stress interventions. We found that the combination of computer and application usage, intervention history, users' activity and their heart rate variability features showed the best accuracy, at 80.0%.

2. Related Work

Here we review previous studies about stress detection, just-in-time stress intervention and users' availability detection and describe how our study is novel compared to previous related work.

Stress has been intensively studied in ubiquitous computing and HCI research fields. One of the active research topics is about detecting stress conditions using passively collected data from a mouse [4], a keyboard [5], [6], a smartphone and wearable sensors [7]. Hovsepian et al. developed a stress model using ECG and respiration data from a laboratory stress induced study and validated the model using the data from both in a laboratory in a week-long field study [8]. Sarker also build a model to predict stress episodes from time series data including physiology, activity, GPS and previous stress history data [9].

There have been studies about providing just-in-time stress interventions. Sharmin et al. visualized physiological sensor data on stress to inform the design of JITAIs [1]. Their findings revealed the importance of contextual and temporal visualizations for making decisions on the timing of interventions. Other studies have developed models to predict stress levels or provide interventions with appropriate timing, automatically using physiological, contextual and user self-report data. Jaimes et al. used Hidden Markov Models to predict heart rate variability, a proxy of stress up to 3 minutes in advance, and developed a just-in-time stress intervention system using reinforcement learning [10].

Ubiquitous computing research has been done to automatically identify/predict users' availability and design notifications which are less interruptible or more effective for users on mobile phones and computers. Fogarty et al. examined predicting when IT workers can be interrupted using sensors and found microphone, time of day, phone usage and mouse and key-board usage can estimate it with 76.3% accuracy [11]. Pejovic and Musolesi developed "InterruptMe", a mobile phone interruption system using user activity, location, time of day, emotions and engagement to increase user satisfaction and reduce response time [12]. Sarker et al. modeled user availability by analyzing physiological, behavioral and self-report data in a week-long field study [13]. They used the delay in responding to a system prompt to objectively measure user availability. They found the 30 most discriminating features to train a machine learning model for predicting availability, including location, affect, activity type, stress, time, and day of the week. Their model showed an accuracy of 74.7% in 10-fold cross-validation and 77.9% with leave-one-subject-out training techniques.

These related prior studies revealed the importance of leveraging multi-modal data to detect stress, design just-in-time interventions and users' availability; however, providing interventions when high stress is observed might disrupt workers. That is why in our paper, we designed models for detecting "good timing of stress intervention delivery" rather than detecting stressful moments. We included user's availability (whether they had time or cognitive capacity to receive stress interventions) and physiological stress and cognitive load levels in our models.

Furthermore, when we consider designing practical models, we need to reduce the users' active input, which can be a burden for users and must rely more on passive data to be usable. Therefore, in our paper, we minimized users' input and leverage multi-modal passive data as much as possible.

In addition, there are several reasons why we designed this study specialized to office workers on their computers, not on their mobile phones. 1) many IT workers tend to keep staring at their computers without taking a break and providing interventions to them could reduce adverse health outcomes and promote their attention and performance, 2) they might dismiss just-in-time interventions provided on the phone as they use their computers as a primary machine, 3) we are able to design a in-situ semi-controlled study where the participants mostly sat or stood then we were able to collect relatively clean data, 4) we can also get rich contextual information from their computer and application usage.

There are computer applications available to tell users to take a break; however, they are simply timer-based or require blacklists from the user and dont take into account users' activity or contextual information.

3. Methods

3.1. Measurements

Thirty participants (male: 12, female: 18, age: ¡ 25 years old: N=3, 25-29: 8, 30-49: 14, 50-59: 5) joined the Health Aware study, as it was called at the time. All the participants were employees at a technology company in the United

States. They were compensated with $250 giftcard on the study completion.

The study consisted of two parts. The first phase was a five-day emotional state and work profile, activity and physiology monitoring period. The second phase, also lasting five days, continued with activity and physiology monitoring and micro-stress interventions were provided on their computers at work.

3.1.1. First Week. At the beginning of the first week, participants filled out pre-study surveys about gender, marital status, race, age, height, weight, and job titles. They also completed the Perceived Stress Scale (PSS) [14], a well-validated screener for stress at the beginning of the study.

During the first week were participants' emotional state, work engagement and challenges surveyed on their computers a few times per hour and whenever participants unlocked their computer screens. The participants sometimes dismissed it and the sampling frequency was lower. Their emotional states were measured with the conventional 2 axes of arousal and valence [15] (-200: less arousal/negative valence) to 200 (more arousal/positive valence) and engagement and challenge were measured with 6 Likert scales (0: Not at all - 5: Extremely). All participants wore an activity tracker (Fitbit, USA) and a heart rate monitor (Zephyr, USA) to monitor their activity, sleep and heart rate and heart rate variability (HRV).In addition, their computer activity (keyboard, mouse, touch panel usage) and application usage (email, website and all applications) were monitored. The sensor and computer activity data were sent to our server and used for further offline analysis.

3.1.2. Second Week. During the second week, the participants continued with activity, sleep, heart rate and computer activity monitoring (no emotion, engagement and challenge monitoring). Our HealthAware application was also installed on their desktop computers to provide interventions (suggestions and recommendations) for stress. The details about the application are described in the next section.

3.2. Micro-stress Intervention Design

The HealthAware application was run in the background of participants' computers throughout the second week of the study.

3.2.1. Stress Intervention. We designed micro-stress interventions which can be done in a short time (¡ 3 minutes). The interventions were grouped into 4 categories (positive psychology, cognitive behavioral, meta-cognitive and somatic practices) and subdivided each category into "individual" and "social" (activities that people perform alone and with others) interventions (5 (interventions per each subcategory) x 2 (individual or social) x 4 (categories) = 40 interventions in total) [16] [17]. See some examples of the interventions in the papers. On each day, the first stress intervention window popped up between 30 and 60 minutes after the first computer activity of the day. After

Figure 1. Example of stress intervention (Individual somatic practice).

that, the stress interventions, chosen randomly from those that had not yet been provided to the participant, were popped up at a randomly chosen time between 90 and 150 minutes after the previous intervention (some studies have shown that human alertness and performance includes ultradian rhythm). Here we show one example of the stress interventions that suggests a quick stretch categorized as an individual somatic intervention (Figure 1). Participants were first asked if this was good timing for receiving a stress intervention using a 7-point Likert scale (1: Extremely bad - 7: Extremely good). If they said "good timing", they were then asked to rate their stress level (1: Not stressed at all - 7: Extremely stressed). They were then provided with a stress intervention and asked to evaluate how much they liked the advice, whether they found it effective or not, and then self-rated their stress level again. We also designed another option of stress interventions; whenever participants wanted to receive interventions, they could also request one by clicking the "stress advisor" button in the HealthAware application. The responses to the interventions were sent to our server and used for further analysis.

4. Opportune Intervention Delivery Timing Prediction Model

In this section, we describe machine learning models to predict good or bad timing for providing stress interventions. Note that we did not use participants' self-reported measurements, such as affect, engagement, arousal, and valence because we aimed to fully automate designing this model mostly with passive data. We included both behavioral and physiological features because we wanted to incorporate participants' behavioral availability (whether they had the time or cognitive capacity to receive stress interventions) and participants' physiological stress and cognitive load levels to our models.

4.1. Feature Extraction

Based on the measurements described in the section 3.1, we extracted 6 different modalities of participants' physiological and behavioral features from data on week2. Table 1 shows a list of features. We used 1, 2, 5 and 10 minutes prior to the intervention pops-ups as windows with which to compute the features.

TABLE 1. A LIST OF FEATURES FOR TIMING PREDICTION

Modality	# of features	Features
Time	2	Time and minutes
Heart rate	6	HeartRate (mean, median, variance), RMSSD (mean, median, variance)
Application Usage	8	Email and Calendar usage [seconds] (1, 2, 5, 10 mins before)
Computer Usage	12	Mouse + keyboard + touch panel usage, mouse usage, keyboard usage [seconds] (1, 2, 5, 10 mins before)
Activity	5	Last night sleep start time and duration, # of awakenings, Steps yesterday, Steps today
Stress Intervention History	5	# of interventions prior on the day, # of good timing interventions prior on the day, # of voluntary interventions prior on the day, time elapsed since last intervention, time elapsed since last good timing intervention [minutes]

4.1.1. Heart Rate and Heart Rate Variability. We computed mean, median and variance of heart rate and HRV to capture participants' autonomic nervous activity. For HRV, we used the square root of the mean squared differences of successive NN intervals (RMSSD), which is a well validated measure of time domain HRV features [18] and one of the most robust features historically used to measure autonomous nervous activity.

4.1.2. Computer and Application Usage. We counted mouse, keyboard and touch activities for each of the time windows used in the model. We also used the number of seconds that the primary email and calendar applications in the company were in the foreground window, ending when the user either changed windows or the computer had no keyboard or mouse activity for a period of five minutes. We used only the data for when the computer was actively used. We chose email and calendar applications since they are 2 of the top 3 software applications used by IT workers [19].

4.1.3. Activity. We also included activity features about sleep and steps, because sleep and physical activity are also well validated as to be related to stress levels [20] [21].

4.1.4. Stress Intervention History. From stress interventions, we extracted # of interventions and time latencies from prior interventions provided on the day. To minimize the use of users' active input in the model, we decided to use features (# of voluntary interventions, # of good timing interventions prior on this day and time latencies from prior good timing interventions on the day) which are least burdensome to users and still practical to develop actual models.

4.2. Models

We used the participants' answers to a 7-point Likert scale prompt "Is this a good timing?" (1:Very Bad - 7:Very Good timing) and the data the participants missed answering the question to split out data into good timing and bad timing. When they missed responding to the interventions, they were engaged in their tasks and did not want to be interrupted. We defined good timing as answers 5-7 to the good timing prompt and bad timing as missed and 1-3. We obtained 362 sets of data. where 58% of all events were categorized as occurring at a bad timing.

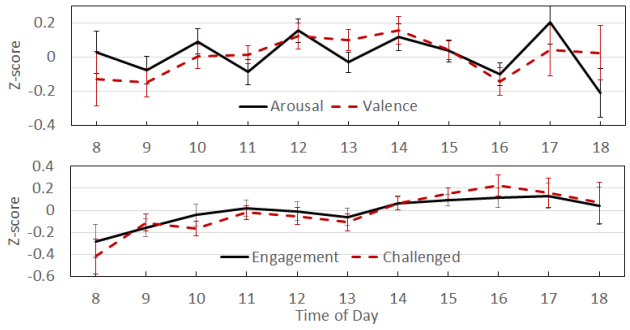

Figure 2. Participants' self-reported emotion and work profiles vs Time in Week 1 (Error bars: +/- 1 Standard Error of the mean).

We compared the following algorithms (1) Neural Network (a two-layer feedforward neural network with 10 hidden layers) (2) Random Forest and (3) Support Vector Machine (radial base function (RBF) Kernel) for each modality of data and all features (the combination of all modalities) which were used in previous stress/availability detection studies. We also applied a multi-kernel learning (MKL) algorithm [22] to all feature data, saw which features showed higher weights on the kernel and integrated different features by simultaneously learning an optimal linear combination of RBF kernels. We used 10-cross validation to repeat training a model with 90% of the data and testing it with the remaining 10% of the data. Within training with 90% of the data, we also used 10-cross validation to optimize parameters. In addition, for algorithms except Random Forest, we used random sampling to balance the numbers of samples for 2 classes. We compared the performance (accuracy and F1 score) of the models with the 6 different modalities of features and all features.

5. Results

5.1. Participants' Emotion and Work Profile

During week 1, we surveyed participants' valence, arousal, engagement and perceived challenge at work, subjectively. We examined the difference of the self-reported measures at different times of the day using Linear Mixed-Effect models that use random and fixed effects to account for the repeated measures within participants. (Figure 2).

As we observed individual differences in these self-reported measures, we normalized the data within partici-

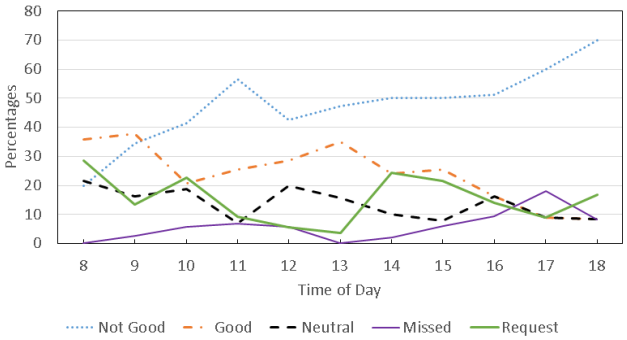

Figure 3. Stress intervention profiles vs Time.

pants (using a z-score transformation). Our results showed that challenge and engagement was lowest at the beginning of the day and highest around 3-4pm ($p < 0.05$) for challenge and 2-3pm for engagement ($p < 0.05$). Valence increased toward the afternoon and dropped at 4pm. Arousal was not statistically different in different times of days but engagement increased in the morning and in the afternoon toward the end of the day except 12-1pm and 6pm.

Overall, averaged PSS, perceived stress scores, at the start of the study among our population was at 13.6, which was around the average compared to the population norm for the same age group (age:18-64, PSS mean: 11.9-14.2) [14].

5.2. Participants' Response to Stress Interventions

Next, we describe stress intervention profiles. First, we examined the responses to stress interventions. Overall, 36% of the stress interventions were rated as good timing (Extremely/very/slightly good timing), 8% were missed and the rest 58% were rated as bad timing. We captured response time from the moment the experience sampling window popped up to the moment participants reacted to it. The response time was not statistically significantly associated with their self-reported timing and stress ratings or PSS scores (a well validated stress measurement tool) at the beginning of the study. We plotted time-series stress intervention profiles (the percentages of good timing, bad timing, neutral, missed and requests) for our users (Figure 3).

In the early morning, the percentage of good timing opportunities for stress interventions (orange line in Figure 3) was higher than for that of suboptimal timing (blue line in Figure 3); however, the perceived percentage of suboptimal intervention timings gradually increased toward the end of the day and the percentages of good timing opportunities decreased once and slightly increased toward 1pm and decreased again. Interestingly, participants requested stress interventions more frequently in the early morning and the early afternoon (2pm) timeframes (green line in Figure 3).

If we compare z-scored averaged self-reported measures on week 1 vs stress intervention profiles on week2, we found engagement and challenge are highly related with the percentages of "Not Good" timing interventions (r=0.83, 0.79) and "missed" interventions (r = 0.69, 0.65) and inversely related with the percentage of "Good" timing interventions (r= -0.80, -0.74). The percentage of "Requested" interventions was neither related with engagement nor with challenge.

These findings confirm that intervention timing can be an important factor in the design of better intervention delivery and that information about time of day and engagement and challenge are associated with good timing of stress intervention deliver.

Among our micro-stress interventions, somatic individual (e.g. breathing exercise) and social (e.g. taking a walk with someone) interventions were rated as the most preferred (mean: 4.1 and 4.2 out of 7, respectively) and effective (mean: 3.8 and 3.7 out of 7, respectively) of all; however the effectiveness was much lower than likeness. Somatic social intervention also showed the largest percentage of self-reported stress reduction, the difference between pre- and post- self-reported stress levels, followed by somatic individual intervention. On the other hand, the least likable and least effective interventions were social positive psychology. Likeness and effectiveness for somatic individual and social interventions were statistically higher than for social positive psychology. Effectiveness was statistically higher at 10 am and 1pm than 5pm (lowest) ($p < 0.05$ respectively, Linear Mixed-Effects Models).

We also analyzed if participants' self-reported likeness and effectiveness were higher when the interventions were provided at good timing. In our study, we collected self-reported likeness and effectiveness only when participants rated good or neutral timing and they requested stress interventions voluntarily. Therefore, we compared their likeness and effectiveness among the following 3 cases: (1) good timing (2) neutral timing (3) voluntary request (Linear Mixed-Effects Models). Effectiveness was higher when interventions were provided at good timing (mean: 3.4) than at neutral timing and voluntarily (mean: 3.2, 3.3, respectively). The average likeness was highest when the interventions were provided voluntarily (3.8), followed by when rated good timing (3.7) and rated neutral timing (3.6); however, we could not find statistical significant differences. These results suggest providing stress interventions at opportune timing could become more effective.

Furthermore, participants with higher PSS scores (stress scores measured in the pre-study survey) showed higher averaged likeness scores for stress interventions (correlation analysis, r = 0.40, statistically significant $p < 0.05$). In other words, participants were experienced higher levels of stress did like getting the stress interventions.

5.3. Predicting Stress Intervention Timing

Table 2 shows a summary of our model's performance in predicting good intervention timing with the different modalities of model features (accuracy and F1 score). Random Forest showed the best results, followed by SVM. In comparing different modalities, the combination of all features showed the best results. Different algorithms showed slightly different orders of performance in different features.

TABLE 2. A SUMMARY OF PERFORMANCE IN PREDICTING GOOD INTERVENTION TIMING

	Neural Network		SVM (RBF kernel)		Random Forest	
Modality	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
Time	50.9	0.49	60.1	0.60	61.7	0.51
Stress intervention history	62.4	0.61	60.6	0.60	65.2	0.55
Computer usage	56.0	0.53	58.3	0.55	58.7	0.48
Activity	52.9	0.53	69.3	0.69	72.0	0.70
Heart rate	56.7	0.59	56.3	0.53	64.3	0.50
Application usage	55.9	0.53	60.1	0.60	64.2	0.55
All	66.7	0.63	64.3	0.64	71.7	0.53

In random forest and SVM, activity features showed the best performance, followed by stress intervention history. In neural networks, stress intervention history features were the best and heart rate features were the second.

5.4. Finding the Better Contributing Features

With MKL SVM, we obtained 80.0% accuracy with F1 score 0.80. Figure 4 showed the top features with highest average kernel weights for SVM. Computer usage, calendar and email usage, intervention history, activity and heart rate variability features highly contributed to the model.

We also considered including participants' voluntarily chosen stress intervention requests from the app or including participants' neutral rating for the stress intervention timing as good timing indicators and computed the performance of the good intervention timing prediction models based on these labeled data. The results showed that accuracy for all features using multi kernel SVM was 71.4 % (F1 score: 0.71) and 59.8% (F1 score: 0.58). The performance was better in the model using labels excluding voluntary stress intervention requests or neutral rating as good timing (80%, 0.8).

6. Discussion

6.1. Stress and Intervention Profiles

Our stress intervention profiles showed that the highest intervention acceptance rate (orange line in Figure 4) occurred at the beginning of the day, with a decreasing trend throughout the day except a peak around 1pm, which can be considered similar to previous findings where interruptions in the early stages of tasks are the least disruptive. This trend is inversely related with the trend of engagement and challenge. We observed slight increases of the intervention acceptance rate around noon when people tended to take breaks for lunch. We also found increases of requested interventions from 2pm to 3pm (green line). It is also known that there is a post-lunch dip in circadian rhythm that lowers performance and increases sleepiness. That could explain why people were amenable to interventions around or after noon [23]. At the end of the day, we assumed our participants wanted to finish their work before they left for

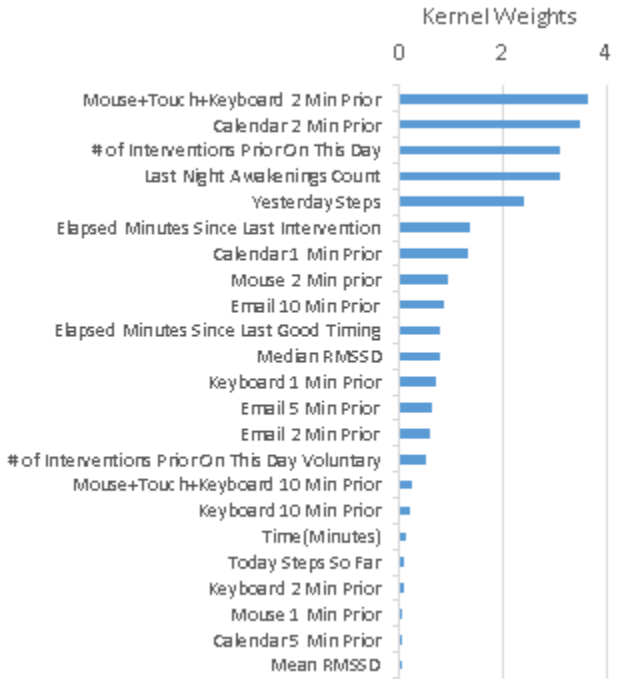

Figure 4. Kernel Weights for good timing prediction.

home, rather than receiving the stress interventions (highest engagement and challenge at the end of the day (Figure 2). Therefore, if we designed better intervention delivery carefully around 2-3pm, we would most likely be able to increase acceptance rates. We also confirmed that self-reported engagement profile was consistent with the focus profile of IT workers from a previous study [24].

Some previous papers used response time to interruptions as an objective user availability measure [12]; however, in our analysis, we did not find the relationship between response time and self-reported user availability. This might be because, in our in-situ study, we did not set up our users' natural work environment to ask participants to respond to the notifications we sent as quickly as possible; however, participants with higher perceived stress levels at the start of the study self-reported wanting to receive the interventions more frequently. This implies that people with higher stress could be more open to receive interventions.

As described in [25], "Poor engagement and burden are likely to hinder intervention effectiveness. Burden is an indication that intervention requirements exceed the momentary personal resources of the participant." In our study, participants found our interventions more effective when they were provided at good timing than at neutral timing. With more data, the results could be statistically significant. In addition, participants preferred somatic interventions consistently with previous finding [16].

6.2. Good Timing Prediction for Stress Interventions

Our results showed that computer and application usage, intervention history, and activity are the top features in predicting better intervention delivery timing. We also found HRV (median and mean RMSSD) could help explain predicting opportune timing to receive interventions.

We discuss similarities and discrepancies between our findings and previous findings. Computer and application usage 1 or 2 minutes prior affected the model. Consistent with previous findings, mouse and keyboard usage is one of the strongest predictor to user availability or stress [26] [27]. As previous study showed, our data also showed that email use decreases heart rate variability that can be a sign of high cognitive load or high stress [28]. Activity features such as yesterday steps and last night awakening count could affect users' stress baseline of the day. "Steps today" could tell us how long they have been sitting. If they walk around for attending meetings or meeting their colleagues, they might rather prefer focusing on their work without trying stress interventions.

Our results also indicated heart rate variability features contribute to predicting opportune timing. Hovsepian et al. found heart rate features (e.g. 80th Percentage of RR Intervals and mean of RR Intervals from ECG) contribute to recognize stress [8]; however our model is not designed to recognize stress, but rather recognize good timing to receive stress interventions, where people's stress level is high but their cognitive load is low and they are available to be interrupted. Heart rate variability could help explain automatic stress responses as well as cognitive load [29] [30]. Another study found heart rate variability feature (median absolute deviation, a measure of ECG variability) was the one of the best predictors for cognitive load [31]. Jaimes et al. used heart rate variability as an objective physiological stress marker to optimize stress interventions using a Hidden Markov Model and a reinforcement learning model [32] [10].

In addition, we also considered building models using only passive data without using any of users' active input (users' responses to stress interventions (e.g., good or bad timing ratings)). The model with intervention history including only # of interventions prior on this day, # of voluntary interventions, and time elapsed since last intervention reduced the performance to predict opportune timing (random forest: accuracy 60.6%, F1 score 0.45, with all features, MKL SVM: 61.1%, F1 score 0.59). Therefore, we showed that some users' input which is less burdensome than their self-reported mood helps improve the model performance.

6.3. Limitations and Future Work

Regarding limitations and future work in this paper, the length of our study (10 days) and the number of the participants (N=30) are limited. The results from the 5-day intervention period could be biased by the novelty effect. Second, We provided interventions with random timing; however the responses to the interventions could include delayed effect of earlier interventions. Third, We built general models for predicting the optimal timing for delivering stress interventions. However, if we collect longer-term data and consider inter-individual differences in stress related behaviors and physiological responses, we could improve the performance of our models. We are also able to build a model to predict self-reported effectiveness of interventions. In the study, we measured heart rate using a chest-band sensor which can be a burden for users. In designing more practical system as a next step, a wrist band heart rate sensor or non-contact physiological sensing with a camera on the computer would be less burdensome but it can be more sensitive to noise. In addition, other modalities of measurement could also be used for this system to improve performance, such as the user's calendar information including deadlines and meetings [33].

7. Conclusions

In this paper, we developed a micro-stress intervention system for IT office workers, studied 30 IT office workers for 10 days, and examined their work patterns, emotion and stress profiles and responses to stress interventions by monitoring their computer and application usage, sleep, activity patterns and heart rate, as well as micro-stress intervention history provided through our desktop software application.

We examined our participants' likeness/effectiveness to the interventions and temporal patterns of stress intervention acceptance/rejection. We also analyzed the relationships among their self-reported responses to the interventions as well as response time as a objective measure and perceived stress levels.

IT workers' reported challenge levels were lowest in the morning and elevated with a peak at 4pm at work; their voluntary stress intervention requests were highest at 2pm. Our participants preferred individual and social somatic interventions and social somatic intervention showed the largest self-reported stress reduction after they received them. People with higher perceived stress levels also showed higher self-reported stress intervention acceptance rates. They also reported high effectiveness in stress interventions when they were delivered at opportune timing.

We also developed machine learning models to predict good/bad timing to provide stress interventions using multimodal data. We found that a combination of the features (mouse and keyboard usage, application usage, stress intervention history, activity, and heart rate variability) was best and provided 80.0% accuracy (F1 score : 0.8) using a multi-kernel support vector machine algorithm in predicting good or bad timing of the interventions for stress.

We believe that our results provide the first, basic and practical information on stress intervention timings in the workplace and that these findings could help designers of just-in-time, closed-loop, stress intervention systems.

References

[1] M. Sharmin, A. Raij, D. Epstien, I. Nahum-Shani, J. G. Beck, S. Vhaduri, K. Preston, and S. Kumar, "Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM Press, 2015, pp. 505–516.

[2] G. Mark, D. Gudith, and U. Klocke, "The cost of interrupted work," in *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems*. ACM Press, 2008, p. 107.

[3] E. Czerwinski, M., Cutrell, E. and Horvitz, "Instant messaging: Effects of relevance and timing," in *Proc. of HCI'00*. British Computer Society, 2000, pp. 71–76.

[4] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, "Activity-Aware Mental Stress Detection Using Physiological Sensors," 2012, pp. 211–230.

[5] J. Hernandez, P. Paredes, A. Roseway, and M. Czerwinski, "Under pressure: Sensing Stress of Computer Users," *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, pp. 51–60, 2014.

[6] L. M. Vizer, L. Zhou, and A. Sears, "Automated stress detection using keystroke and linguistic features: An exploratory study," *International Journal of Human Computer Studies*, vol. 67, no. 10, pp. 870–886, 2009.

[7] A. Sano and R. W. Picard, "Stress Recognition Using Wearable Sensors and Mobile Phones," *Affective Computing and Intelligent Interaction, 2013 Humaine Association Conference on*, pp. 671–676, 2013.

[8] K. Hovsepian, M. Al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "cStress," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM Press, 2015, pp. 493–504.

[9] H. Sarker, I. Nahum-Shani, M. Al'Absi, S. Kumar, M. Tyburski, M. M. Rahman, K. Hovsepian, M. Sharmin, D. H. Epstein, K. L. Preston, C. D. Furr-Holden, and A. Milam, "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 Press, 2016, pp. 4489–4501.

[10] L. G. Jaimes, M. Llofriu, and A. Raij, "PREVENTER, a Selection Mechanism for Just-in-Time Preventive Interventions," *IEEE Transactions on Affective Computing*, vol. 7, no. 3, pp. 243–257, jul 2016.

[11] J. Fogarty, S. E. Hudson, and J. Lai, "Examining the robustness of sensor-based statistical models of human interruptibility," *Proceedings of the 2004 conference on Human factors in computing systems*, vol. 6, no. 1, pp. 207–214, 2004.

[12] V. Pejovic and M. Musolesi, "InterruptMe," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM Press, 2014, pp. 897–908.

[13] H. Sarker, M. Sharmin, A. A. Ali, M. M. Rahman, R. Bari, S. M. Hossain, and S. Kumar, "Assessing the availability of users to engage in just-in-time intervention in the natural environment," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM Press, 2014, pp. 909–920.

[14] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress." *Journal of health and social behavior*, vol. 24, no. 4, pp. 385–96, dec 1983.

[15] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[16] P. Paredes, R. Gilad-Bachrach, M. Czerwinski, A. Roseway, K. Rowan, and J. Hernandez, "PopTherapy: Coping with Stress through Pop-Culture," in *Pervasive Health '14*, 2014.

[17] A. Sano, P. Johns, and M. Czerwinski, "HealthAware: An advice system for stress, sleep, diet and exercise," in *2015 International Conference on Affective Computing and Intelligent Interaction*. IEEE, sep 2015, pp. 546–552.

[18] "Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology." *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[19] G. Mark, S. Voida, and A. Cardello, ""A pace not dictated by electrons"," *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, p. 555, 2012.

[20] S. Green, "Health Psychology: A Textbook, 2nd edition by Jane Ogden. Open University Press, Buckingham, 2000, 396 pages, £17·99, ISBN 0 335 20596 8." *Journal of Advanced Nursing*, vol. 33, no. 5, pp. 696–696, jul 2008.

[21] R. Leproult, G. Copinschi, O. Buxton, and E. Van Cauter, "Sleep loss results in an elevation of cortisol levels the next evening." *Sleep*, vol. 20, no. 10, pp. 865–70, oct 1997.

[22] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. ACM Press, 2009, pp. 1–8.

[23] T. H. Monk, D. J. Buysse, C. F. Reynolds, and D. J. Kupfer, "Circadian determinants of the postlunch dip in performance." pp. 123–33, 1996.

[24] G. Mark, S. T. Iqbal, M. Czerwinski, and P. Johns, "Bored mondays and focused afternoons," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM Press, 2014, pp. 3025–3034.

[25] I. Nahum-shani, S. N. Smith, K. Witkiewitz, L. M. Collins, B. Spring, and S. A. Murphy, "Just-in-time adaptive interventions (JITAIs): An organizing framework for ongoing health behavior support." *The Methodology Center Technical Report*, vol. 073975, no. 14, pp. 1–37, 2014.

[26] D. Sun, P. Paredes, and J. Canny, "MouStress," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM Press, 2014, pp. 61–70.

[27] J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang, "Predicting human interruptibility with sensors," *ACM Transactions on Computer-Human Interaction*, vol. 12, no. 1, pp. 119–146, 2005.

[28] G. Mark, S. T. Iqbal, M. Czerwinski, P. Johns, and A. Sano, "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 - CHI '16*, vol. 21, no. 1, 2016, pp. 98–109.

[29] N. Kudo, H. Shinohara, and H. Kodama, "Heart rate variability biofeedback intervention for reduction of psychological stress during the early postpartum period." *Applied psychophysiology and biofeedback*, vol. 39, no. 3-4, pp. 203–11, dec 2014.

[30] A. Shearer, M. Hunt, M. Chowdhury, and L. Nicol, "Effects of a brief mindfulness meditation intervention on student stress and heart rate variability." *International Journal of Stress Management*, vol. 23, no. 2, pp. 232–254, 2016.

[31] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psychophysiological measures for assessing cognitive load," *Proceedings of the 12th ACM international conference on Ubiquitous computing*, p. 301, 2010.

[32] L. Jaimes, M. Llofriu, and A. Raij, "A Stress-Free Life: Just-in-Time Interventions for Stress via Real-Time Forecasting and Intervention Adaptation," in *Proceedings of the 9th International Conference on Body Area Networks*. ICST, 2014.

[33] J. Bakker, L. Holenderski, R. Kocielnik, M. Pechenizkiy, and N. Sidorova, "Stess@Work," in *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics 2012*. ACM Press, 2012, p. 673.