Observing and Predicting Knowledge Worker Stress in the Wild

Authors will be visible after double blind review

ABSTRACT

Knowledge workers face many challenges as they work: work is fragmented, disruptions are constant, tasks are complex, and work hours can be long. These challenges can impact knowledge workers' stress, focus and awakeness, and in turn their interaction with the digital environment, the quality of work performed and their productivity in general. We report on a field study with 14 knowledge workers over an eight-week period in which we investigated, using experience sampling, how the workers experience stress over time. During this field study, we also collected biometric data, which we then used to investigate if it is possible to predict such productivity-related indicators as stress, focus and awakeness, in the moment. We observed and report on various trends in knowledge worker stress levels over several weeks and show that classifiers can be built that are able to predict stress, focus and awakeness from biometric data above a baseline.

Author Keywords

Biometrics, Stress, Awakeness, Focus, Computer Interaction, Ubiquituous Computing, Empirical Study, User Centered Design

INTRODUCTION

'The most valuable asset of a 21st-century institution (whether business or non-business) will be its knowledge workers and their productivity' [23]. Knowledge workers constantly face challenges, such as a high work fragmentation, continuous disruptions and distractions, highly complex and demanding tasks, and long working hours [34, 52, 20]. These challenges, amongst others, can lead to stress in the workplace. Stress is an ever-growing concern as it can lead to fatigue, burnout and various other illnesses, ultimately resulting in work absences and marked productivity losses [42, 71, 29].

Given the importance of understanding stress and its relationship and effect on work, a number of studies have been conducted using a variety of methods. Some have studied stress using invasive techniques, such as measuring cortisol as a stress biomarker (e.g., [66]). Such invasive techniques are best suited for laboratory environments. Others have shown the benefit of using autonomic nervous system (ANS) activity by analyzing biometric signals of the human body, such as blood pressure (e.g., [47]). These less invasive approaches can be used in longer-term field studies, opening up the ability to ask the question of what stress looks like in the wild.

In this paper, we build on this ability to non-invasively monitor stress to investigate two questions: RQ1) how do knowledge workers at work experience stress over time, and RQ2) can we predict whether a knowledge worker is experiencing stress in the moment based on biometric data. A better understanding of how stress is experienced (RQ1) can help inform the design of workplaces to manage and alleviate stress. An ability to predict stress in the moment (RQ2) can enable the development of digital tools to avoid stress.

Our work builds upon previous work and extends it by performing an eight-week study in the workplace with 14 knowledge workers, collecting biometric data as well as experience samples on stress, and two related human aspects: focus and awakeness. We collected participant reports on focus because the challenges that contribute to stress levels can also make it hard for knowledge workers to stay focused and the level of focus can affect productivity [53]. We collected participant reports on awakeness because a lack of awakeness (sleepiness) can result in undesired consequences on work [16]. The participants in our study, who perform various job functions for a research and development group within a single large corporation, wore a single biometric armband sensor[1] with low invasiveness that captured heart-, respiration- and skin-related measures. This modality was chosen to ease longitudinal deployment in the field. Using machine learning, we created classifiers and analyzed their ability to predict stress, focus, and awakeness levels. Considering all of these three human aspects helps us to consider whether one aspect might be easier to detect than another. If one aspect is easier to detect, it might

[1] Biovotion Everion sensor [11]

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04. . . $15.00

DOI: https://doi.org/10.1145/3313831.XXXXXXX

serve as a proxy or an indicator of the presence (or absence) of other aspects.

In our analysis, we identify several trends in day-to-day stress levels that emerged over the course of our eight-week study. The results of our analysis also show that biometric signals collected from a single minimally invasive sensor can be used to predict stress and its related aspects accurately, with the abstract concept of focus (predictably) being the hardest to detect. Our results further show that knowledge workers' self-reported levels of stress, focus and awakeness and their physiological manifestation and prediction can vary substantially between individuals. The main contributions of our work are:

- A qualitative examination of how knowledge worker stress presents and fluctuates in the wild.
- The creation and analysis of measures for the automatic monitoring of knowledge workers' stress, focus and awakeneess in the workplace based on an eight-week field study with 14 office workers.
- A discussion on the impact of applying this research to improve the interaction of knowledge workers with their digital environment leading to an improvement in their productivity and well-being, besides a reflection on aspects that can be further improved in future studies.

RELATED WORK

The study and prediction approach on which we report in this paper is related to previous studies of productivity-related indicators and studies using biometerics to predict these indicators. We consider related work in each of these categories in turn.

Productivity-related Indicators

Our study focuses on the productivity-related indicators of stress, focus and awakeness.

Stress

Much previous work relates to identifying and mitigating stress in an office environment. Previous studies have measured stress by taking one of two possible approaches. The first approach is to measure plasma catecholamine and cortisol as stress biomarkers [66]. This approach is impractical for use over prolonged periods of time, as in our eight-week study. Further, this approach is imprecise because of the delay from the stress stimulation to the stress response, which may take from minutes to hours [12, 39].

The second approach, which we have chosen for our study, is to measure autonomic nervous system (ANS) activity by analyzing biometric signals of the human body, such as blood pressure, heartbeat, and temperature [47, 77, 76]. In particular, changes in heart rate variability are associated with cognitive and emotional stress [55, 22]. This second approach has been used successfully by several past studies [61, 32, 56].

Hovsepian et al. [43] have worked to obtain a standard for continuous stress assessment. They used sensors to conduct a seven-day lab study with 26 participants, as well as a field study with 20 participants. They found that their model showed significant improvement over simple heart rate variability measurements. However, this model requires the use of a suite of invasive sensors that would be impractical for a study the length of ours.

Hernandez et al. [41] investigated the use of pressure-sensitive keyboard and a capacitative mouse as non-intrusive means for measuring computer users stress levels. They found participants exhibited significantly increased typing pressure and mouse contact when in stressful conditions. McDuff et al. [55] experimented with a camera to measure photoplethysmographic signals indicative of cognitive stress. Vizer et al. [78] used keystroke and linguistic features to automatically measure stress levels in response to cognitive and physical stress conditions. All of these studies were performed in a controlled lab setting over a short duration and the results have yet to be replicated in the field.

Evans and Johnson [25] investigated the correlation between noise in the workplace and stress levels. They found that workers exposed to open-office noise showed aftereffects that indicate motivational deficits. The population for their experiment comprised 40 female clerical workers, who were randomly assigned to a control condition or to three-hour low-intensity noise room designed to simulate typical open-office noise levels.

Stress in the workplace and its effects on service providers has been analyzed in call centers [40] and in the context of the perceived imbalance between resources and demands [14]. These studies considered several factors such as personality traits, career-related goals and attitudes, and life outside of work, and these factors were correlated with stress levels and burnout.

Kocielnik et al. [49] developed a framework for unobtrusive and continuous measurement of stress in real life conditions. They equipped university staff members with a wristband sensor and combined this data with information from the participants calendars over the course of four weeks. They observed that the data they collected reflected well the participants perceptions of their own stress levels. However this work focused mostly on providing retrospective information to users so they can work to improve their own stress balance. They did not attempt to make observations about the big picture of participants stress profiles or make predictions in real time.

Our study stands out from these works by nature of its length and focus on knowledge worker stress in everyday office life, using unobtrusive measures. To the best of our knowledge there is no longitudinal study which attempts to explain and predict knowledge worker stress that comes close to the length of our own study. The eight week duration gives us authority to speak on the nature of day-to-day fluctuations in knowledge worker stress levels.

Focus

Focus refers to the allocation of limited cognitive processing resources [4]. Mark et al. [53] studied engagement in workplace activities by analyzing the digital activity of 32 information workers in situ for 5 days to understand how attentional states change with context. They found that boredom is highest in the early afternoon and focus peaks in the middle of the afternoon. They also found that doing work that requires focus correlates with stress, while rote work correlates with happiness.

Interruptions in the office are a common barrier keeping workers from sustaining focus on their work related activities, particularly when the interruptions occur at inopportune moments. Such interruptions may include emails, alerts, or interactions with co-workers[34, 15, 45]. Interruptions in inopportune times can have negative effects that range from higher error rate and lower overall performance to an increase in stress and frustration [5, 19, 52]. External interruptions may cause workers to enter a "chain of distraction" [45]. This chain is composed by stages of preparation, diversion, resumption and recovery that result in time away from an ongoing task.

Other studied constructs that relate to focus in the workplace include cognitive absorption, cognitive engagement, flow, and mindfulness. Cognitive absorption describes periods of time in which a person experiences total immersion in an activity. This state is also accompanied by a sense of deep enjoyment, a feeling of control, curiosity, and not realizing the passing of time. It has been associated with ease of use and perceived usefulness of information technology [2]. Cognitive engagement is described [79] as a period of strong focus in an activity without the feeling of a sense of control of the situation. Flow [18], and mindfulness [80, 21] are psychological states that describe periods of prolonged attention and total immersion in an activity. Flow occurs when a person is focused on an activity that requires high challenge and high use of the person's skills, whereas mindfulness is characterized by being aware of fine detail, affording the capacity to discover and manage unexpected events.

Awakeness

Sleepiness (lack of awakeness) and its associated risk of serious injury to passengers has been studied in the context of automobile accidents [16, 60]. These studies show a strong association between the level of acute driver sleepiness and the risk of injury crash. Connor et al. [16] conducted a population-based case study using the Stanford sleepiness scale, which is similar to a seven-point Likert scale and describes seven different levels of sleepiness from "Could not stay awake, sleep onset was imminent" (1) to "Felt active, wide awake" (7). Nordbakke and Sagberg [60] show that drivers are well aware of various factors influencing the risk of falling asleep while driving. Drivers also have good knowledge of the most effective measures to prevent falling asleep at the wheel. However, most of drivers continue driving even when recognizing sleepiness signals, due to the desire to arrive at a reasonable time, the length of the drive, or pre-planed commitments.

Biometrics

Our study investigates the prediction of multiple productivity-related indicators (stress, focus, and awakeness) using two different types of measurements, biometric signals and computer interaction, over eight weeks in a real-life office setting. Existing work [70, 38, 81, 85, 35, 64, 59, 67] analyzes a broad array of biometric signals and correlates them with individual's cognitive states and processes. For example, Zuger et al. [86] used biometrics to sense interruptibility in the office [86]. Biometric signals have also been studied in the context of technology users. For example, Parnin [64] analyzes electromyography to measure sub-vocal utterances, and how these might be correlated with the programmer's perceived difficulty of programming tasks. Similarly, biometrics have been used to measure code difficulty by using biometric sensors [31] and using Near Infrared Spectroscopy to measure developer's cerebral blood flow [59].

Eye tracking technology [7, 17, 68] and brain activity [44, 72] have been used in previous studies to analyze different tasks in an office environment. Eye tracking has been used to analyze memory load and processing load by inspecting task-evoked pupillary response and pupil size [6]. Similar studies have shown high correlation between pupil size and mental workload of subtasks [6] and cognitive load [48]. Brain activity has been associated with different mental states [8] by analyzing specific frequency bands (alpha, beta, gamma, delta, and theta) using electroencephalography (EEG). The increase or decrease of some of these frequencies is correlated with attentional demand and working memory load [73, 74]. In contrast to studies that use eye tracking or EEG, we focused on a less invasive technology that can be applied in a real world scenario.

FIELD STUDY

We conducted an eight-week field study with 14 participants using experience sampling and biometric sensors to investigate how knowledge workers experience stress over time and the feasibility of predicting stress, focus, and awakeness based on biometric signals.

Participants

We recruited 14 professionals via personal contacts from a large power and automation company. All participants work primarily in an office environment, though half of the participants spend at least 10% (and up to 50%) of their time in a laboratory environment. Office workers are a population that generalizes to a variety of contexts, and including part-time laboratory workers guarantees that our participants have varying work patterns that include different levels of computer usage, as well as different levels of activity in both individual and collaborative tasks.

Of the 14 participants, 11 are male and 3 are female. The average participant age is 40, with 5 in the age range 25-34, 7 in the age range 35-44, and 2 in the age range 55+. The participants have an average number of years of professional experience of 12, with 2 having less than 5 years, 10 having 5-15 years, and 2 having more than 25 years. All participants work for a research organization within the company, but their job functions span line management, laboratory science, scientific research, technology evaluation, and software development.

Procedure

We performed an in-situ study over the course of eight weeks. For this, we first informed participants about the study purpose and procedure, handed out biometric sensors and introduced and explained the self-reporting to the participants. For 10 of the participants, we further installed a computer activity

Figure 1. We used Biovotion's Everion to collect biometric measurements from the participants

tracker on their company-issued laptops (note: four participants from the original sample declined for privacy reasons). After the initial setup, participants were asked to fill out 3 surveys each day for the following eight weeks (see Section 3.3.3). At the end of the eight weeks, we collected the biometric sensors and performed a short follow-up interview on the study and the participants' experiences.

In the second month of the study, we offered participants two one-hour Tai Chi classes per week (each in the middle of the day) in addition to the daily experience sampling and asked them to attend one class a week. We chose to offer Tai Chi class for two reasons: first, to incentivise participants for their continued participation and for keeping them engaged; and second, for their said benefits of reducing stress based on our consultation with researchers in psychiatry, psychology and more specifically on mindfulness practices.

Data Collection

In this section we describe the two datasets that we collected from each study participant.

Biometric Sensors

Figure 1 illustrates Biovotion's Everion, which we used to track the biometric signals of the study participants. The Everion is worn on the upper arm and provides continuous monitoring of certain biometric measurements.[2] Previous studies [86, 70, 38, 81, 85, 35] have used similar devices [62, 24, 27] to capture psycho-physiological and biometric measurements for shorter periods of time or capturing a smaller number of measurements.

Table 1 lists the biometrics measurements that we collected using the Everion. Each measurement is collected once per second, and each recorded observation has an associated timestamp and quality rating. Data collected by the Everion are uploaded to a server, from which we downloaded the data for use in our study.

Biometric Measurement	Units of Measure
Physical Activity	[30, 3]
Intensity of motion	(No unit)
Energy Expenditure	Calories per second (cal/s)
Step counter	Steps
Heart	[37, 38, 57, 36]
Heart rate	Beats per Minute (bpm)
Blood pulse wave	(No Unit)
Heart rate variability (RMSSD)	Milliseconds (ms)
Blood oxygenation	Percent (%)
Blood perfusion	(No unit)
Skin	[38, 36]
Galvanic skin response	kOhm
Skin temperature	Degrees Celsius (°C)
Respiration	[57, 38, 36, 54]
Respiratory rate	Breaths per Minute (bpm)

Table 1. Biometric measurements captured by the Everion, organized by category and with references to previous works using similar data

Computer Interaction Data

To gain a better understanding of our participants day-to-day work activities, we installed an open source computer interaction monitor (reference omitted for doubleblin.). The monitor ran in the background on participants' computers and tracked the active windows, as well as the keyboard and mouse activity. Four of our participants opted out of this part of the study for privacy reasons (S6, S10, S12, and S13). Therefore, we only included the computer interaction data in our analysis for our more coarse-grained explanatory model but not in our analysis on the predictive model for stress in the moment.

Surveys

Following guidelines from previous studies [50, 63, 51] and following the preferrences of extensive user piloting, we sent via text message a survey request to each participant two times per work day. Pilot participants preferred these over other means, in part, due to them being accessible and noticeable anywhere in the office.

We sent the first request at a random time between 9am and 11am and sent the second request at a random time between 1pm and 3pm. We randomized the request times to avoid either establishing or observing a standard behavioral pattern. That is, we did not want the participants to plan for the arrival of the survey request at a set time, and we did not want the survey request to overlap with a set daily behavior (e.g., coffee break every day at 2:30pm). Similarly, we avoided using tools which allow for too much freedom in response time [1] since this would discurage participation in stressed timeframes and would bias the corpus. The same survey was sent each time:

1. How awake are you right now?

2. How stressed do you feel right now?

3. How focused on work are you right now?

We used the phrase "right now" to capture each aspect in the moment (so as to permit later prediction of each aspect based on biometric data). The wordings of the questions are based on a previous survey of individuals in an organizational context [33]. The use of awakeness (rather than sleepiness) in Question 1 is inspired by previous work [82] and to

[2] https://biovotion.zendesk.com/hc/en-us/categories/201633909-Everion-Device

some extent also captures the “arousal” aspect of the affective space [69].

One last survey was sent at the end of the day, at 4:25pm which asked the four different questions detailed below:

1. How awake have you been today?
2. How stressed did you feel today?
3. How productive have you been today?
4. How do you feel about your work day?

Following guidelines from similar previous studies [28, 75], we asked the participants to respond to each question using a 5-point Likert scale ranging from 1 (not at all awake/stressed/focused) to 5 (extremely awake/stressed/focused). Each participant response, as stored by Survey Gizmo, comprised the date, the time at which the response was initiated, the time at which the survey was submitted, the unique identifier for the participant, and the responses submitted by the participant.

OBSERVED TRENDS OVER TIME IN STRESS LEVELS

To gain insights into how knowledge workers experience stress over an extended period of work, we examined the end of day survey responses collected from each participant to see if any identifiable trends emerged. We used data from 13 of the 14 participants - we excluded one participant from this analysis as they experienced atypical stress levels in the latter half of the study due to factors outside of our control.

Baseline Stress Levels

Common amongst all participants was a trend to select one stress rating far more frequently than any other. We will refer to this value as the participant’s baseline stress level. All but one participant reported their perceived stress level for the day as their baseline stress level more than 50% of the time. In total, the baseline values made up 65% of the reported values collected from participants. Interestingly, while participants sometimes saw periods of sustained increases in stress, lasting as many as 6 consecutive workdays in the most extreme case, participants would always return to their baseline stress level at some point.

The baseline stress level varied significantly between participants. 7 participants (54%) reported feeling average stress levels most frequently (rating 3 on our scale), while 5 (38%) reported feeling little stress (rating 2) and 1 (8%) reported feeling no stress at all (rating 1).

Figure 2 illustrates these points, showing both participants tendency to report and return to baseline stress levels, as well as a distinct difference in baseline stress level (rating 2 for S1 vs rating 3 for S3).

Stressful Days Tend to Cluster

Accounting for the variance between participants perceived stress baselines, we consider a stressful day to be one that represents a deviation of 1 or more stress levels above the participant’s baseline. Of the 93 stressful days we observed in total, we found that 39 (41%) of these days occurred in groupings of two or more consecutive stressful workdays. The most common size of these groups was two workdays, while the largest group we observed was six workdays. The day after a stressful day is much more likely to be a stressful day as compared to any other day with a 0.55 average increase over baseline, compared to 0.02 average increase over baseline.

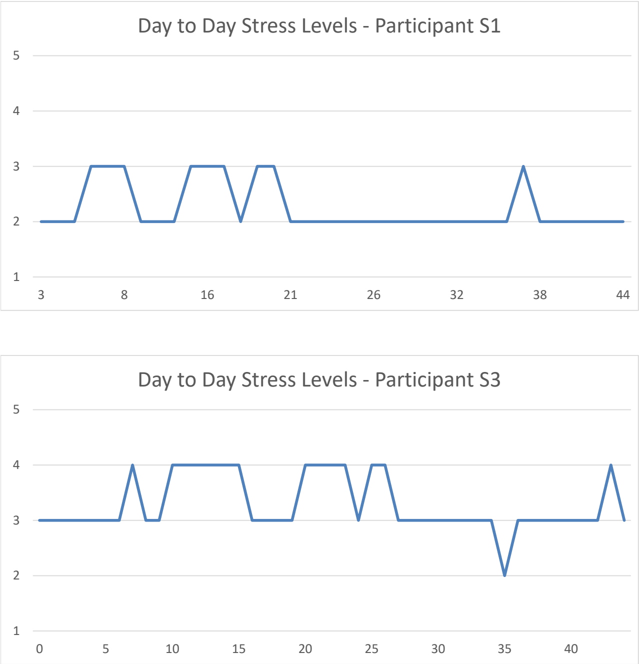


Figure 2. The day-to-day stress levels reported by two participants (S1 and S3) are shown. The y-axis represents values on the 5-point Likert scale we asked participants to respond with ranging from 1/Not at all stressed to 5/Extremely stressed. The x-axis represents the day of the study on which the value was recorded (from 0-45). The x-axes are slightly different for the 2 participants as they did not begin and end the study on exactly the same days.

Extreme Changes in Stress Levels are Rare

After accounting for each participant’s perceived stress baseline, we examined the frequency of deviations from the baseline. We found that participants were far more likely to report a stress level that was within 1 point of their baseline, than to report a stress level 2 or more points away. These extreme deviations represented only 15% of all reported values which differed from the participants baseline. As well, the majority (78%) of these deviations came from just two participants. This suggests that some people may be less resilient to the stress of the workplace than others. For most participants, extremely stressful days were few and far between.

Explaining Stress Fluctuations

In an attempt to explain some of the stress that our participants were experiencing, we created a linear mixed model with the self-reported daily stress levels as dependent variable and the participants as random effects. We experimented with day of the week and proximity to beginning or end of month as possible explanatory variables. We also calculated daily features from the gathered computer interaction data where available—namely total time spent actively on the computer, percentage of time idle, and amount of time spent on non work-related web browsing per day. Ultimately, the analysis showed that none of the variables that we examine had a significant explanatory power with respect to our participants perceived stress levels. This indicates that aspects such as the time spent on a computer or the non work-related web browsing do not have a significant impact on the stress level. Furthermore, this points to the need for additional instrumentation and data collection, if we are to succesfully understand and predict stress in the workplace.

PREDICTING STRESS IN THE MOMENT

To investigate whether stress, focus and awakeness can be predicted in the moment based on biometric measures. we investigated classifiers trained for each individual and across all participants. We report on the effectiveness of these classifiers and the features that are important in prediction of stress, focus and awakeness.

Data Preparation

In this section we describe how we preprocessed the collected data for use in training and testing machine learning models.

Data Linking

We linked the collected biometric data and survey responses for each participant. Linking the data is necessary to construct training and test datasets for use in creating and evaluation machine learning models.

Our linking approach is as follows. From the start time of each survey response, we look back one hour for available biometric data. For each minute in that hour-long time window, we check for biometric data to associate with the survey response. For example, if a participant started a survey response at 11:05am, we look for biometric data in the time frame 10:05am to 11:05am. If biometric data is available in the hour-long time window, we consider the survey response to have associated biometric data. Otherwise, we exclude the survey response from our dataset.

Reasons for a survey response to lack associated biometric data include:

- The participant not wearing the Everion in the hour before before beginning the survey
- The Everion not recording data in the hour before the participant began the survey (e.g., due to low battery)
- Biometric data not being uploaded successfully to the server

Figure 3 illustrates the number of survey responses with associated biometric data for each study participant. The total number of responses per participant is affected by their response rate and by the number days out-of-office (e.g., vacations, holidays, etc.). Participant S2 and S12 have particularly low numbers of usable survey responses. In each of these cases, the issue related to biometric data not being uploaded successfully to the server.

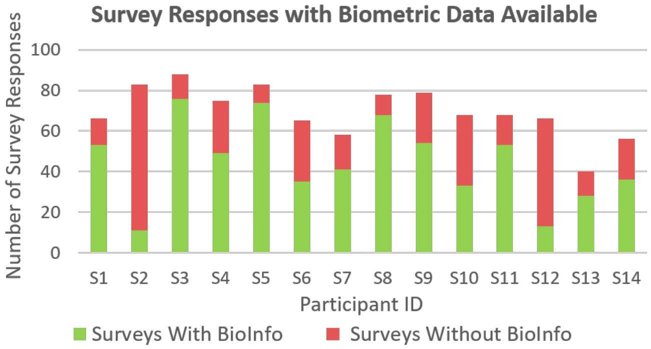


Figure 3. The figure shows the biometric data available per each participant. Green sections represent survey responses with biometric data available, red sections responses with no biometric data available

Feature Extraction

We extracted features from the biometric data to provide as input to machine learning models. Previous studies [9, 85] identify time windows as an important factor that impacts the prediction accuracy of a classifier. We considered many time windows from the literature on biometric analysis [86], ranging from 10 seconds to 3 hours. Specifically, we considered the following time windows: *10sec, 20sec, 30sec, 45sec, 1min, 2min, 3min, 5min, 7.5min, 10min, 20min, 30min, 45min, 1hour, 2hour, 3hour*.

From the start time of each survey response, we look back the amount of time that corresponds to each time window, and we create features for all of the biometric data available in that time window. For example, if a participant started a survey response at 11:05am, for the 30min time window, we create features using all of the available biometric data from 10:35am to 11:05am. For each time window, we calculate 10 statistical measurements from the biometric data to create 10 distinct features. Specifically, the 10 statistical measurements are: mean, standard deviation, variance, median, percentile25, percentile75, interquartile range, maximum, minimum, and range. Thus, for each survey response, we generate a large number of corresponding features based on three factors: biometric measurement, time window, and statistical measurement. In addition to these biometric features, we also considered the time of day in which the questions were asked. These features are created to predict the responses described by the ground truth.

Response Transformations

Table 2 illustrates the distribution of responses from each participant for each of the three survey questions (listed in Section 3.3.3). The figure shows that there is a notable imbalance in the distribution of the self-reported responses provided by the participants. Most participants did not use all five points of the five-point Likert scale in their responses, and the distributions tend to skew toward one side or the other, depending on the question. Thus, we binarized the survey data into a two-point scale to give the machine learning models the best possible chance to make useful predictions. The two points in the binary scale represent negative or positive responses for each of the three human aspects of interests (e.g., not stressed or stressed).

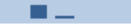
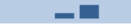
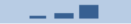
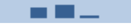
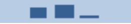
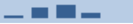
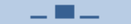
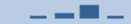
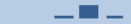
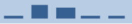
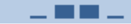
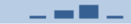
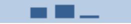
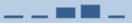
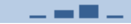
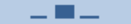
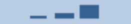
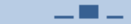
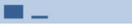
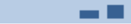
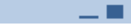
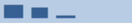
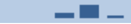
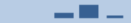
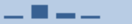
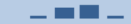
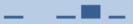
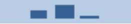
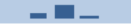
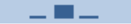
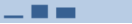
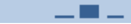
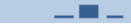
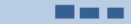
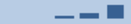
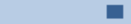
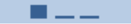
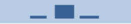
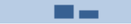
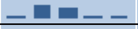
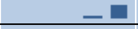
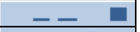
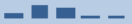
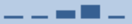
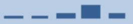
		Distributions		
Participant	# Responses	Stress	Focus	Awakeness
S1	52			
S2	10			
S3	76			
S4	48			
S5	74			
S6	34			
S7	41			
S8	68			
S9	54			
S10	33			
S11	53			
S12	13			
S13	27			
S14	36			
All	619			

Table 2. The distribution of the responses of each participant to the three questions asked during the day are shown. Each bar in the histograms represent one of the 5 values on the 5-point Likert scale we asked participants to respond with, where the far left side of the histograms are 1/Not at all, and the far right sides are 5/Extremely

We binarized the survey responses as follows. For each participant, we calculated the median response value for each question. We classified each response below the median as 0 ('negative') and each response above the median as 1 ('positive'). The distribution for the stress question skewed left, so we included the median values in the 'positive' class, while the distributions for focus and awakeness skewed right, so we included those median values in the 'negative' class.

Oversampling

Even after binarizing the responses as described in the previous section, we found the distribution of responses was still quite imbalanced for many of our participants. This can be seen in the distribution columns in Table 4. To combat this, we applied random oversampling to our training sets, which artificially rebalances the dataset by creating randomly replicated data in the minority class. This has been a commonly used technique in previous studies on unbalanced datasets [13, 83].

Variable	K	# Estimators	Minimum Samples Split
Stress	300	100	4
Focus	200	50	4
Awakeness	800	100	4

Table 3. The hyperparameters selected by grid search analysis to tune our random forest models. K refers to the number of features selected.

Selecting a Classifier Algorithm

There are many different algorithms that can be used to build a classifier. To select an algorithm, we compared multiple classifiers using the popular machine learning library scikit-learn [65] and performed a grid search analysis to determine the optimal hyperparameters for each classifier. Our analysis showed that random forest outperforms all other classifiers, including Naïve Bayes, decision trees, support vector machine, and a multilayer perceptron neural network. The optimal values for random forest and the three output measures are listed in Table 3. For the remainder of this paper, we refer to random forest classifiers trained with these hyperparameters.

Individual Classifiers

Since peoples' experience of stress, focus, and awakeness (as well as their physiological manifestations) can vary substantially (e.g., [40]), we first trained individual classifiers for each participant (rather than a general one for all participants). The results of our analysis are reported in Table 4. For our analysis, we report values of accuracy, one of the most commonly used metric to compare performance, as well as precision and recall of the classes of interest: 'stressed', 'not focused', and 'not awake'. Since the imbalance in the data can lead to high accuracy values if a classifier always just predicts the most likely/frequent class while ignoring the class of higher importance and interest, precision and recall of the class of interest are also important to consider [83, 10, 40]. For some users (i.e., S1. as seen In Table 2), the imbalance in their data was so extreme that even after adjusting by oversampling it was impossible to create a reasonable classifier. These scenarios are difficult to predict as any classifier will not have enough variance in its training data for the 'stressed' situation to adequately distinguish it from the non-stressed case.

Overall, we were able to use extracted physiological features to predict all three aspects with reasonable accuracy, precision, and recall, as well as to improve upon the baseline—a stratified random classifier that randomly chooses one of the two classes with a bias towards the larger class. While the individually trained classifiers improved on average across all participants upon the baseline in all cases excepting recall of 'not focused', the improvement was substantially higher for awakeness (85% improvement in precision, 62% in recall, and 7% in accuracy) than for stress or focus. Also, the performance of the individually trained classifiers varied greatly across participants. While some participants showed a large improvement, for others the baseline performed much better than the individually trained classifier. For instance, for predicting 'stressed', the individual classifiers improved upon the baseline for S4, S6, S8, S11, S12, and S14 with a maximum improvement of 88.4% in precision and 184.6% in recall for S12, while they did worse for S1, S3, S5, S7, S9, S10, and S13, and in the worst cases did not correctly predict a single instance of 'stressed'. Typically users that have the lowest precision and recall values are those where the data is the most unbalanced.

Feature Selection and Importance

There are a large variety of features that can be (and have been) calculated in previous research for each of the basic measurements listed in Table 1, such as the mean, standard deviation, maximum, and interquartile range. In addition, each of these metrics can be combined with the various time windows captured of a basic measurement, resulting in a large feature space. To reduce the feature space, we experimented with multiple feature selection methods, including selecting the top k highest correlated features by various metrics such as mutual information, Pearson's correlation coefficient, ANOVA's F-value, as well as wrapper methods such as recursive feature elimination, optimizing mean decrease accuracy by iteratively permuting features, and only selecting features that exceed a certain Gini importance threshold. We found that all methods produced

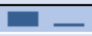
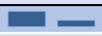
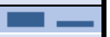
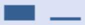
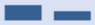
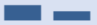
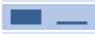
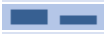
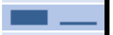
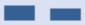
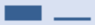
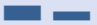
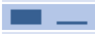
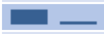
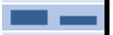
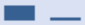
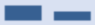
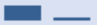
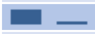
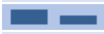
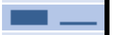
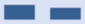
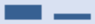
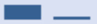
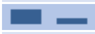
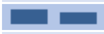
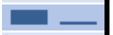
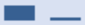
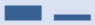
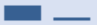
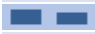
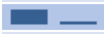
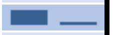
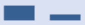
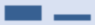
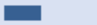
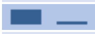
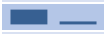
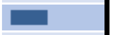
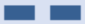
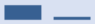
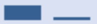
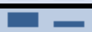
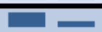
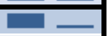
	Stress				Focus				Awakeness			
Participant	Accuracy	Precision of class 'stressed'	Recall of class 'stressed'	Distribution	Accuracy	Precision of class 'not focused'	Recall of class 'not focused'	Distribution	Accuracy	Precision of class 'not awake'	Recall of class 'not awake'	Distribution
S1	0.694	0.000	0.000		0.533	0.143	0.150		0.706	0.520	0.525	
S2	0.900	0.000	0.000		0.070	0.088	0.150		0.370	0.000	0.000	
S3	0.775	0.033	0.017		0.646	0.484	0.400		0.961	0.000	0.000	
S4	0.519	0.477	0.514		0.825	0.000	0.000		0.677	0.566	0.600	
S5	0.915	0.000	0.000		0.793	0.196	0.127		0.649	0.500	0.381	
S6	0.824	0.400	0.400		0.497	0.216	0.158		0.912	0.500	0.333	
S7	0.976	0.000	0.000		0.607	0.495	0.444		0.932	0.000	0.000	
S8	0.643	0.606	0.619		0.694	0.376	0.233		0.775	0.490	0.380	
S9	0.548	0.183	0.156		0.472	0.372	0.348		0.765	0.383	0.083	
S10	0.970	0.000	0.000		0.761	0.573	0.489		0.842	0.000	0.000	
S11	0.668	0.603	0.586		0.719	0.025	0.010		0.757	0.000	0.000	
S12	0.900	0.785	0.925		0.792	0.917	0.375		1.000	-	-	
S13	0.915	0.000	0.000		0.959	0.000	0.000		1.000	-	-	
S14	0.583	0.582	0.599		0.833	0.263	0.250		0.988	0.870	1.000	
Overall	0.773	0.262	0.273		0.657	0.296	0.224		0.809	0.319	0.275	
Baseline	0.716	0.256	0.251		0.650	0.249	0.266		0.755	0.173	0.170	
Improvement (%)	8.049	2.306	8.719		1.156	19.120	-15.837		7.206	84.646	61.747	
General	0.530	0.246	0.460		0.554	0.274	0.466		0.522	0.130	0.294	

Table 4. Results of predictions using the individual models. The distribution columns show a bar chart of the response distribution (negative/positive) for each of the three variables. The baseline row represents the averaged results of our baesline classifier. The general row shows the averaged results of our models trained on all participants.

similar results with respect to accuracy, precision, and recall for the individual models. Ultimately, we chose to use the top k features with the highest ANOVA F-value, as it is relatively simple and efficient to calculate. The values of k used were selected by grid search analysis, and are shown in Table 3.

Overall, the features that were selected as the important ones for the individual models based on the random forest algorithm varied greatly across participants. Yet, some feature categories were considered to be important more frequently than others. Table 5 shows the averaged Gini importance for the feature categories used for predicting stress. Particularly important for stress were the feature categories heart rate variability (18.3%) and skin temperature (15.7%), both of which have been shown in several previous studies to be indicators for stress [22, 55, 47]. For predicting focus, the feature categories for blood pulse wave (14.2%) and heart rate variability (13%) showed to be the most important categories, while for awakeness the most important ones were heart rate variability (13.6%) and blood pulse wave (13.1%).

Feature Category	Stress	Focus	Awakeness
Heart Rate Variability	18.3%	13%	13.6%
Blood Pulse Wave	10%	14.2%	13.1%
Heart Rate	8.7%	12.6%	10.3%
Skin Temperature	15.7%	9.8%	10%
Galvanic Skin Response	6.6%	8.2%	5.1%
Respiration Rate	14.8%	12.7%	10%
Oxygen Saturation	5.6%	4.3%	2%
Energy Expenditure	6%	7.7%	4.8%
Activity	4.6%	7.8%	7.6%
Steps	1.7%	0.8%	0.9%
Time of Day	0%	0.1%	0.5%

Table 5. The averaged Gini importance of each feature category, per response variable.

Individual vs. General Model

Individual models are trained specifically for each individual and thus require a data collection period before they are capable of making accurate predictions. On the other hand, the idea of general models is to be able to train them on already collected data and then to be able to apply them even to new and unseen individuals, thus overcoming the cold-start problem. Given the large individual differences in biometrics, training a general model to achieve an adequate accuracy for new individuals is not necessarily possible.

To examine the performance of a general model for our participants, we trained three general models, one for focus, one for awakeness and one for stress. We roughly followed the same procedure as for the individual models. Due to the larger amount of data available in the general case, we used the more common random undersampling, which randomly selects elements in the majority class to exclude from the dataset, instead of random oversampling to balance the distribution of the dataset. The models were trained on the datasets of 13 of the 14 participants, and then evaluated on the dataset of the last, repeating this process for all 14 participants.

The bottom row of Table 4 presents the averaged performance results for this approach in terms of accuracy, precision, and recall. Although the averaged precision and recall are comparable or better than those of the averaged individual results, this was at the cost of a large decrease in overall accuracy. Upon closer investigation into the performance of the general model when testing on each participant, we found that individually trained models for each participant performed much better than a general model trained over all participants. Using stress as an example, for participant S12, for whom we saw the greatest increase compared to the baseline in individual models, the general model was unable to predict a single instance of 'stressed' correctly. This is consistent with our expectations because biometric features are highly specific to individuals.

DISCUSSION

Humans experience stress, focus and awakeness in different ways. In this paper, we have attempted to study these mental states in the workplace using both participant self-reports and biometric data. We discuss implications from our study for the workplace, including ways in which the information might inform digitally-controlled or digitally-informed parts of the workplace. We also discuss challenges imposed by the data and possible future paths of research.

Implications for Workplaces

Being able to accurately recognize periods of high-stress in knowledge workers could enable more respectful workplaces. For example, an ability to sense and predict stress in the moment could help companies to prevent or de-escalate potentially dangerous situations in the workplace, e.g. confrontation between co-workers. Building an understanding of stress and focus over time and in the moment could also help create workplaces that are more conducive to enabling knowledge workers to be more productive. For example, this information could be used to feed an awareness dashboard of a team's stress level, and avoid digital interruptions in high-stress or high-focus periods similar to previous studies [84]. Building an understanding of awakeness and focus could also enable the creation of workplaces that are conducive to workers producing higher-quality work. For example, if awakeness or focus decreases, they might be enhanced by adapting lighting in the workplace or scheduling breaks to prevent focus loss.

Currently, the cost of biometric sensors and necessary infrastructure, such as automated light and sound systems for adjusting the environment, makes our approach most appropriate for high-value workspaces, such as control rooms, command centers, or dispatch offices. However, as standard office settings become more personalizable (e.g., via adjustable desks, lighting, and sound showers) and sensor costs decrease, our approach could be applied to any office environment, and thus could impact a large percentage of modern workers. As in modern cars, temperature and lighting could be regulated on a per-person basis, which would allow the environment to react to the person's current state and to maximize each person's preferences and productivity (e.g., preferences of men and women in temperature [46]). Further research would be needed to identify how to balance needs across a group of office workers and how to handle conflicting levels between different group members.

The Effect of Tai Chi on Stress

As described in our study procedure, we offered and asked participants to attend a one-hour Tai Chi class per week for the last four weeks of the study. We recorded each participant's attendance, including which day they attended the offered classes. The two primary reasons for this intervention was to keep participants motivated to continue the self-reports over the long study period, and to offer a technique that might help reduce stress.

While this was not the focus of our study, we performed a secondary analysis to examine whether the Tai Chi classes had any effect on the participants' stress levels in the work days directly following the Tai Chi session. For this, we build a linear mixed model with the self-reported daily stress level as dependent variables and the participants as random effects. We found that Tai Chi attendance contributed a small amount to decreased stress in the work week immediately following the Tai Chi session (slope of -0.188, p <0.05). However, we did not find a connection between attendance and stress on the day of the session suggesting that Tai Chi might have long-term effects, but is not conducive at relieving stress close in time to the intervention. Overall, the Tai Chi thus had a small impact on the collected data of the second month of the study, which poses a threat to the validity to our observed trends for this period of time. However, since knowledge workers might attend these kinds of classes on their own doing, we believe that this is negligible, and the analysis rather provides a weak indication that this kind of stress intervention might in fact help to reduce stress in the wild.

Ground Truth and Self-Reporting

Studying mental states, such as stress, awakeness and focus, requires collecting a valid ground truth from each participant. We spent considerable time when designing the study determining the exact questions to ask of participants, consulting experts in the area, and basing the questions and wording on previous research and studies. Despite the care taken, it is possible that the gathered data lacks reliability and validity. Some have questioned the reliability and validity of self-reports as we used in our study due to subjective biases, lack of care in reporting, and the highly individual nature of reporting aspects such as stress [40, 43]. In addition, in contexts such as the workplace, as in our study, participants might be afraid to genuinely report levels of aspects, such as sleepiness. Hence, there is a chance that the self-reports we gathered do not adequately reflect the ground truth of the underlying variables under investigation. It could even be the case that certain biometrics might represent a more accurate ground truth of the studied phenomena than the self-reports. This suggests that a more confirmative study rather than an inquiry study could be a better approach, and we will explore such routes in future work.

Imbalanced Data

Study participants provided highly imbalanced data in their survey responses, with most participants only taking advantage of a subset of the Likert-scale values and the data points mostly being clustered around the middle of the scale, as can be seen in Table 2. While some of the imbalance is expected due to certain classes, such as 'not stressed', being more common in the workplace, this imbalance also provides challenges in the training and assessment of a machine learning classifier, as also found by others, e.g. [26]. We addressed this for the training by oversampling in case of few data samples for the individual models and undersampling in case of a general model where more data was available. Especially in light of this imbalance in the data, the results we achieved with our models are encouraging. For the assessment of the classifiers' performance we addressed the imbalance by not just presenting accuracy, but also by focusing on prediction and recall to examine the classifier's performance in predicting the infrequent (yet more important) cases, such as when a user is struggling to stay awake and an intervention or warning might be needed most.

Predicting Stress with Computer Interaction Data

Knowledge workers, a focus of our work and study, often spend a large amount of time each day interacting with information on their computer at work. An interesting direction for future study is to consider whether this computer interaction data, which can be gathered non-invasively as work occurs, could serve to sense and predict stress, focus and awakeness. Features that could be investigated include keystrokes per minute, mouse clicks per minute and changes in the active window title.

THREATS TO VALIDITY

There are numerous threats to validity to our study.

External Validity: The results of our study may not generalize to a broader population of office workers. To mitigate this risk, we collected participants from a wide variety of different departments with different age ranges, genders, work experience, and working in different positions.

Another threat is that our results may notgeneralize to a different office environment. We conducted this study in a typical office environment, similar to many among technology workers across the world. These office environments control for a series of variables to make them standard world wide such as controlled temperature and lighting.

Internal Validity: This study tries to find correlations between biometric features and different productivity-related indicators (stress, focus, and awakeness). Nonetheless, biometric signals are influenced by far more variables than the ones this study comprehends. Therefore, trying to draw a strong causality between the biometric features and the productivity-related aspects would be inaccurate. To mitigate this risk, we collected the data in a rote environment and in a regular manner to minimize the number of external causes that may affect each participant's biometric signals.

It is possible that the amount of data collected is not sufficienet to draw valid conclusions. To address this threat, we collected a data for an eight week period, which is 400% longer than previous studies [86, 58].

Construct Validity: A threat to the study is that there are other factors that might either influence the human aspects of interest or that were considered but are unrelated biometric signals. To mitigate this risk, we used an state-of-the-art device that captures a large number of highly accurate biometric signals. We collected the most commonly analyzed biometrics that historically have shown correlation with productivity aspects from each user. In an effort to maintain this research applicable to real-world environments, we picked this already existing device, even when, as a trade-off, we could not capture more descriptive and more intrusive signals such as SDNN, SCL, SCR, eye tracking, or brain activity. We also perform a grid analysis in our machine learning model to pick the number of features that is more predictive of the productivity-related indicators, therefore removing the biometric features that may be unrelated to the indicators.

CONCLUSIONS

Stress, awakeness and focus at work are highly relevant aspects when it comes to productivity and well-being at the workplace. In this paper, we presented the results of a study with 14 professional knowledge workers in their workplace over an eight week period to better understand how workers experience stress over time and to examine the ability of biometrics to predict these productivity-related aspects. The longitudinal and in situ placement of the study support and extend previous work. Based on twice collected daily survey responses, we observed that although participants sometimes saw periods of sustained stress, they would always return to a baseline stress level for them at some point. We also observed that stress levels seldom spiked, but when they did rise, the rise in stress tended to last more than a day. In addition to the survey responses, we continually collected biometrics data with which we were able to create a model that was able to predict awakeness, stress, and focus.

These results open up new opportunities to help increase knowledge workers' productivity and well-being, ranging from instantaneously taking action to prevent potentially risky situations and prevent accidents due to a lack of focus or awakeness, all the way to recommending interventions to reduce stress if it becomes more chronic.

REFERENCES

[1] Alexander T. Adams, Elizabeth L. Murnane, Phil Adams, Michael Elfenbein, Pamara F. Chang, Shruti Sannon, Geri Gay, and Tanzeem Choudhury. 2018. Keppi: A Tangible User Interface for Self-Reporting Pain. In *CHI Conference on Human Factors in Computing Systems*.

[2] Ritu Agarwal and Elena Karahanna. 2000. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. In *MIS Quarterly*, Vol. 24 (4). 665–694.

[3] Steven G. Aldana, Leanne D. Sutton, Bert H. Jacobson, and Michael G. Quirk. 1996. Relationships between Leisure Time Physical Activity and Perceived Stress. *Perceptual and Motor Skills* 82, 1 (1996), 315–321. DOI: http://dx.doi.org/10.2466/pms.1996.82.1.315

[4] John R. Anderson. 2004. *Cognitive Psychology and Its Implications*. 519 pages.

[5] Brian P Bailey, Joseph A Konstan, and John V Carlis. 2001. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *Proceedings of INTERACT*, Vol. 1. 593–601.

[6] Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. In *Psychological Bulletin*, Vol. 91(2). Issue 276.

[7] Roman Bednarik and Markku Tukiainen. 2006. An Eye-tracking Methodology for Characterizing Program Comprehension Processes. In *Proc. of ETRA*.

[8] Hans Berger. 1929. Uber das Elektrenkephalogramm des Menschen. In *European Archives of Psychiatry and Clinical Neuroscience*, Vol. 87. 527–570.

[9] Peter Vorburger Abraham Bernstein and Alen Zurfluh. 2005. Interruptability Prediction Using Motion Detection. In *Workshop on Managing Context Information in Mobile and Pervasive Environments*.

[10] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J. Christopher Westland. 2011. Data mining for credit card fraud: A comparative study. *Decision Support Systems* 50, 3 (2011), 602 – 613.

[11] Biovotion. 2018. Everion. `https://www.biovotion.com/everion/`. (2018). [Online; accessed 9-July-2019].

[12] Tarani Chandola, Alexandros Heraclides, and Meena Kumari. 2010. Psychophysiological biomarkers of workplace stressors. In *Neurosci Biobehav Rev.*, Vol. 35. 51–57.

[13] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. 2004. Editorial: Special Issue on Learning from Imbalanced Data Sets. *SIGKDD Explor. Newsl.* 6, 1 (June 2004), 1–6. DOI: http://dx.doi.org/10.1145/1007730.1007733

[14] Cary Cherniss. 1980. *Staff Burnout - Job Stress in the Human Services*. Sage Publications, Inc.

[15] Jan Chong and Rosanne Siino. 2006. Interruptions on software teams: a comparison of paired and solo programmers. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM, 29–38.

[16] Jennie Connor, Shanthi Ameratunga, Robyn Norton, Elizabeth Robinson, Roger Dunn, John Bailey, Ian Civil, and Rod Jackson. 2002. Driver sleepiness and risk of serious injury to car occupants: population based case control study. In *BMJ*.

[17] M.E. Crosby and J. Stelovsky. 1990. How do we read algorithms? A case study. *Computer* 23, 1 (1990).

[18] Mihaly Csikszentmihalyi. 1990. *Flow: The Psychology of Optimal Experience*.

[19] Mary Czerwinski, Edward Cutrell, and Eric Horvitz. 2000. Instant messaging: Effects of relevance and timing. In *People and computers XIV: Proceedings of HCI*, Vol. 2. British Computer Society, 71–76.

[20] Mary Czerwinski, Eric Horvitz, and Susan Wilhite. 2004. A Diary Study of Task Switching and Interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, 175–182.

[21] Erik Dane. 2011. Paying attention to mindfulness and its effect on task performance in the workplace. In *Journal of Management*, Vol. 37 (4). 997–1018.

[22] Rod K. Dishman, Yoshio Nakamura, Melissa E. Garcia, Ray W. Thompson, Andrea L. Dunn, and Steven N. Blair. 2000. Heart rate variability, trait anxiety, and perceived stress among physically fit men and women. *International Journal of Psychophysiology* 37, 2 (2000), 121 – 133.

[23] Peter F Drucker. 1999. Knowledge-worker productivity: The biggest challenge. *California management review* 41, 2 (1999), 79–94.

[24] Polar Electro. 2017. Equine H7 heart rate sensor belt set. `https://www.polar.com/en/products/equine/accessories/equine_H7_heart_rate_sensor_belt_set`. (2017). [Online; accessed 9-July-2019].

[25] Gary W. Evans and Dana Johnson. 2000. Stress and open-office noise. In *Journal of Applied Psychology*, Vol. 25(5). 779 – 783.

[26] Anja Exler, Andrea Schankin, Christoph Klebsattel, and Michael Beigl. 2016. A wearable system for mood assessment considering smartphone features and data from mobile ECGs. In *Pervasive and Ubiquitous Computing*. 1153–1161.

[27] Inc. Fitbit. 2017. Fitbit Charge 2. `https://www.fitbit.com/de/charge2`. (2017). [Online; accessed 9-July-2019].

[28] James Fogarty, Andrew J. Ko, Htet Htet Aung, Elspeth Golden, Karen P Tang, and Scott E. Hudson. 2005. Examining Task Engagement in Sensor-Based Statistical Models of Human Interruptibility. In *SIGCHI Conference on Human Factors in Computing Systems*. 331–340.

[29] European Foundation for the Improvement of Living and Working Conditions. 2010. *Work-related stress*. Technical Report. `https://www.eurofound.europa.eu/printpdf/publications/report/2010/work-related-stress`.

[30] Kenneth R. Fox. 1999. The influence of physical activity on mental well-being. *Public Health Nutrition* 2, 3 (1999), 411–418.

[31] Thomas Fritz, Andrew Begel, Sebastian C Müller, Serap Yigit-Elliott, and Manuela Züger. 2014. Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 402–413.

[32] Vera Gal and Vesna Vuksanovic. 2007. Heart rate variability in mental stress aloud. In *Medical Engineering and Physics*, Vol. 29. 344–349.

[33] P.A. Gloor, D. Oster, O. Raz, A. Pentland, and D. Schoder. 2010. The Virtual Mirror: Reflecting on the Social and Psychological Self to Increase Organizational Creativity. *International Studies of Management & Organization* 40, 2 (2010), 74–94.

[34] Victor M González and Gloria Mark. 2004. Constant, constant, multi-tasking craziness: managing multiple working spheres. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 113–120.

[35] Nitesh Goyal and Susan R Fussell. 2017. Intelligent Interruption Management using Electro Dermal Activity based Physiological Sensor for Collaborative Sensemaking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 52.

[36] Andreas Haag, Silke Goronzy, Peter Schaich, and Jason Williams. 2004. Emotion recognition using bio-sensors: First steps towards an automatic system. In *Tutorial and research workshop on affective dialogue systems*. Springer, 36–48.

[37] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 301–310.

[38] Jennifer A Healey and Rosalind W Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *Transactions on intelligent transportation systems* 6, 2 (2005), 156–166.

[39] Dirk H. Hellhammer, Stefan Wust, and Brigitte Martina Kudielka. 2009. Salivary cortisol as a biomarker in stress research. In *Psychoneuroendocrinology*, Vol. 34. 163–171.

[40] Javier Hernandez, Rob R. Morris, and Rosalind W. Picard. 2011. Call Center Stress Recognition with Person-Specific Models. In *International Conference on Affective Computing and Intelligent Interaction*. 125–.

[41] Javier Hernandez, Pablo Paredes, Asta Roseway, and Mary Czerwinski. 2014. Under pressure: Sensing Stress of Computer Users. In *CHI Conference on Human Factors in Computing Systems*.

[42] G. Robert J. Hockey. 1997. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology* 45, 1 (1997), 73 – 93.

[43] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. In *ACM international conference on Ubiquitous computing*. 493–504.

[44] Y. Ikutani and H. Uwano. 2014. Brain activity measurement during program comprehension with NIRS. In *Proc. of SNPD*.

[45] Shamsi T. Iqbal and Eric Horvitz. 2007. Disruption and Recovery of Computing Tasks: Field Study, Analysis and Directions. In *SIGCHI Conference on Human Factors in Computing Systems*. 677–686.

[46] Sami Karjalainen. 2007. Gender differences in thermal comfort and use of thermostats in everyday thermal environments. In *Building and Environment*, Vol. 42. 1594–1603. Issue 4.

[47] H. Kataoka, H. Kano, H. Yoshida, A. Saijo, and M. Yasuda ; M. Osumi. 2000. Development of a skin temperature measuring system for non-contact stress evaluation. In *IEEE Engineering in Medicine and Biology Society*, Vol. 20.

[48] Jeff Klingner. 2010. Fixation-aligned pupillary response averaging. In *Symposium on Eye-Tracking Research and Applications*. 275–282.

[49] Rafal Kocielnik, Natalia Sidorova, Fabrizio Maria Maggi, Martin Ouwerkerk, and Joyce H. D. M. Westerink. 2013. Smart technologies for long-term stress monitoring at work. *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems* (2013), 53–58. DOI: `http://dx.doi.org/10.1109/cbms.2013.6627764`

[50] Sebastien Lalle, Cristina Conati, and Giuseppe Carenini. 2016. Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data. In *International Joint Conference on Artificial Intelligence*.

[51] Yuhan Luo, Bongshin Lee, Donghee Yvette Wohn, Amanda L. Rebar, David E. Conroy, and Eun Kyoung Choe. 2018. Time for Break: Understanding Information Workers' Sedentary Behavior Through a Break Prompting System. In *CHI Conference on Human Factors in Computing Systems*.

[52] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 107–110.

[53] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, and Paul Johns. 2014. 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*. ACM, 3025–3034.

[54] Yuri Masaoka and Ikuo Homma. 1997. Anxiety and respiratory patterns: their relationship during mental stress and physical load. *International Journal of Psychophysiology* 27, 2 (1997), 153–159.

[55] Daniel J. McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W. Picard. 2016. COGCAM: Contact-free Measurement of Cognitive Stress During Computer Tasks with a Digital Camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 4000–4004.

[56] Nicola Montano, Alberto Porta, Chiara Cogliati, Giorgio Costantino, Eleonora Tobaldini, Karina Rabello Casali, and Ferdinando Iellamo. 2009. Heart rate variability explored in the frequency domain: A tool to investigate the link between heart and behavior. In *Neuroscience and Biobehavioral Reviews*, Vol. 33. 71–80.

[57] LJM Mulder. 1992. Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological psychology* 34, 2 (1992), 205–236.

[58] Sebastian Muller and Thomas Fritz. 2016. Using (Bio)Metrics to Predict Code Quality Online. In *In Proceedings of the ICSE.*

[59] Takao Nakagawa, Yasutaka Kamei, Hidetake Uwano, Akito Monden, Kenichi Matsumoto, and Daniel M. German. 2014. Quantifying Programmers' Mental Workload During Program Comprehension Based on Cerebral Blood Flow Measurement: A Controlled Experiment. In *Companion Proc. of ICSE.*

[60] Susanne Nordbakke and Fridulv Sagberg. 2007. Sleepy at the wheel: Knowledge, symptoms and behaviour among car drivers. In *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 10. 1–10.

[61] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. 1996. Heart Rate Variability, Standards of Measurement, Physiological Interpretation, and Clinical Use. In *European Heart Journal*, Vol. 93. 1043–1065.

[62] Yoshio Okada, Tsuyoshi Yi Yoto, Taka aki Suzuki, Satoshi Sakuragawa, Hiroyuki Sakakibara, Kayoko Shimoi, and Toshifumi Sugiura. 2011. Wearable ECG Recorder with Acceleration Sensors for Monitoring Daily Stress*: Office Work Simulation Study. In *Journal of Medical and Biological Engineering*, Vol. 34. 420–426.

[63] Prateek Panwar and Christopher M. Collins. 2018. Detecting Negative Emotion for Mixed Initiative Visual Analytics. In *CHI Conference on Human Factors in Computing Systems.*

[64] Chris Parnin. 2011. Subvocalization - Toward Hearing the Inner Thoughts of Developers. In *Proceedings of International Conference on Program Comprehension.*

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Clondel, P. Prettenhofer, R.Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research 12*. 2825–2830.

[66] Jennifer R. Piazza, David M. Almeida, Natalia O. Dmitrieva, and Laura C. Klein. 2010. Frontiers in the Use of Biomarkers of Health in Research on Stress and Aging. In *35th Annual International Conference of the IEEE EMBS*, Vol. 65B (5). 513–525.

[67] Stevche Radevski, Hideaki Hata, and Kenichi Matsumoto. 2015. Real-Time Monitoring of Neural State in Assessing and Improving Software Developers' Productivity. *Proceedings of Connected Health: Applications, Systems and Engineering Technologies* (2015).

[68] Paige Rodeghero, Collin McMillan, Paul W. McBurney, Nigel Bosch, and Sidney D'Mello. 2014. Improving Automated Source Code Summarization via an Eye-tracking Study of Programmers. In *International Convernce on Software Engineering.*

[69] J.A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[70] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 671–676.

[71] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Troster, and U. Ehlert. 2010. Discriminating Stress From Cognitive Load Using a Wearable EDA Device. *IEEE Transactions on Information Technology in Biomedicine* 14, 2 (2010), 410–417.

[72] Janet Siegmund, Christian Kästner, Sven Apel, Chris Parnin, Anja Bethmann, Thomas Leich, Gunter Saake, and André Brechmann. 2014. Understanding Source Code with Functional Magnetic Resonance Imaging. In *Proceedings of International Conference on Software Engineering.*

[73] Michael E. Smith and Alan Gevins. 2005. Neurophysiologic monitoring of mental workload and fatigue during operation of a flight simulator. In *The International Society for Optical Engineering*. 116–126.

[74] M. B. Sterman, C. A. Mann, and D. A. Kaiser. 1993. Quantitative EEG patterns of differential in-flight workload. In *6th Annual Workshop on Space Operations Applications and Research*, Vol. 2.

[75] Takahiro Tanaka and Kinya Fujita. 2011. Study of user interruptibility estimation based on focused application switching. In *Conference on Computer Supported Cooperative Work*. 721–724.

[76] Mariaconsuelo Valentini and Gianfranco Parati. 2010. Variables Influencing Heart Rate. In *Progress in Cardiovascular Diseases*, Vol. 52. 11–19. Issue 1.

[77] Alexander P. J. van Eekelen, Jan H. Houtveen, and Gerard A. Kerkhof. 2004. Circadian variation in base rate measures of cardiac autonomic activity. In *European Journal of Applied Physiology*, Vol. 93. 39–46.

[78] Lisa M. Vizer, Lina Zhou, and Andrew Sears. 2009. Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies* 67, 10 (Oct. 2009), 870–886. DOI: **http://dx.doi.org/10.1016/j.ijhcs.2009.07.005**

[79] Jane Webster and Hayes Ho. 1997. Audience engagement in multimedia presentations. In *Data Base for the Advancement in Information Systems*, Vol. 28(2). 63–77.

[80] Karl E. Weick and Kathleen M. Sutcliffe. 2006. Mindfulness and the quality of organizational attention. In *Organization Science*, Vol. 17. 514–524.

[81] Jacqueline Wijsman, Bernard Grundlehner, Hao Liu, Hermie Hermens, and Julien Penders. 2011. Towards mental stress detection using wearable physiological sensors. In *Engineering in Medicine and Biology Society*. IEEE, 1798–1801.

[82] P. Wilhelm and D. Schoebi. 2007. Assessing modd in daily life: Structural validity to change, and reliability of a short-scale to measure three basic dimensions of mood. *European Journal of Psychological Assessment* 23, 4 (2007), 258–267.

[83] Bee Wah Yap, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, and Nik Nik Abdullah. 2014. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, Tutut Herawan, Mustafa Mat Deris, and Jemal Abawajy (Eds.). Springer Singapore, 13–22.

[84] Manuela Züger, Christopher Corley, André N Meyer, Boyang Li, Thomas Fritz, David Shepherd, Vinay Augustine, Patrick Francis, Nicholas Kraft, and Will Snipes. 2017. Reducing Interruptions at Work: A Large-Scale Field Study of FlowLight. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 61–72.

[85] Manuela Züger and Thomas Fritz. 2015. Interruptibility of software developers and its prediction using psycho-physiological sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2981–2990.

[86] Manuela Zuger, Sebastian Muller, Andre Meyer, and Thomas Fritz. 2018. Sensing Interruptibility in the Office: A Field Study on the Use of Biometric and Computer Interaction Sensor. In *Conference on Human Factors in Computing Systems*.