Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Observing and predicting knowledge worker stress, focus and awakeness in the wild

Mauricio Soto [*,1,a], Chris Satterfield [1,b], Thomas Fritz [c], Gail C. Murphy [b], David C. Shepherd [d], Nicholas Kraft [e]

[a] *Hitachi ABB Power Grids, Raleigh, North Carolina, United States*
[b] *The University of British Columbia, Vancouver, British Columbia*
[c] *University of Zurich, Zurich, Switzerland*
[d] *Virginia Commonwealth University, Richmond, Virginia, United States*
[e] *UserVoice, Raleigh, North Carolina, United States*

ARTICLE INFO

Keywords:
Biometrics
Psycho-Physiological Data
Stress
Awakeness
Focus
Computer Interaction
Ubiquitous Computing
Empirical Study
User Centered Design

ABSTRACT

Knowledge workers face many challenges in the workplace: work is fragmented, disruptions are constant, tasks are complex, and work hours can be long. These challenges can affect 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 and awakeness over time. During this field study, we also collected biometric data including heart- and skin-related measures, which we then used to investigate if it is possible to predict stress, focus and awakeness, in the moment. We observed and report on various trends in knowledge worker stress and awakeness levels over several weeks, finding that people tend to have certain "baseline" levels for these aspects. Moreover, we found that days with high levels of stress tend to cluster, similarly as the days with low awakeness. We further show that machine learning models can be built from the data of a single minimally invasive device to predict stress, focus, and awakeness. Overall, we found that our models were capable of large improvements in precision and recall in comparison to a random classifier for stress (25.9% increase over random for precision, 4.2% for recall) and awakeness (52.4% increase in precision, 40.8% in recall). The abstract concept of focus proved to be the hardest to predict (26.0% increase in precision, 27.8% decrease in recall).

1. Introduction

'The most valuable asset of a 21st-century institution (whether business or non-business) will be its knowledge workers and their productivity' Drucker (1999). Knowledge workers constantly face challenges, such as a high work fragmentation, continuous disruptions and distractions, highly complex and demanding tasks, and long working hours Czerwinski et al. (2004); González and Mark (2004); Mark et al. (2008). 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 Hockey (1997); for the Improvement of Living and Conditions (2010); Setz et al. (2010).

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. Minimally invasive studies have shown the benefit of sensing autonomic nervous system (ANS) activity by analyzing certain biometric (*aka.* psycho-physiological) signals of the human body, such as skin temperature Kataoka et al. (2000). These approaches can be used in long-term field studies, opening up the question of what stress looks like in the wild. In contrast, other studies have studied stress using more invasive techniques, such as measuring cortisol as a stress biomarker Piazza et al. (2010). These studies are best suited for laboratory environments.

* Corresponding author.
E-mail addresses: mauricio.soto-gonzalez1@hitachi-powergrids.com (M. Soto), cds00@cs.ubc.ca (C. Satterfield), fritz@ifi.uzh.ch (T. Fritz), murphy@cs.ubc.ca (G.C. Murphy), shepherdd@vcu.edu (D.C. Shepherd), nkraft@gmail.com (N. Kraft).

[1] These authors contributed equally.

https://doi.org/10.1016/j.ijhcs.2020.102560
Received 1 April 2020; Received in revised form 12 October 2020; Accepted 17 October 2020
Available online 2 November 2020

Our work builds on this ability to non-invasively monitor stress by performing an eight-week study in the workplace with 14 knowledge workers examining stress as well as two related human aspects: focus and awakeness (i.e., wakefulness). We collected data on these two additional aspects because stress and the challenges that contribute to stress levels can also affect sleep. A lack of sleep can affect the ability of knowledge workers to stay awake and focused at work which, in turn, can result in undesired consequences on work and productivity Cohen et al. (1997); Connor et al. (2002); Mark et al. (2014).

In this paper, we investigate two research questions. Our first question examines how knowledge workers experience stress and awakeness in their workplace over a long period of time. Our second question explores the possibility of using biometric data to predict whether a knowledge worker is experiencing stress, is focused, and is awake at a given moment in time. A better understanding of how knowledge workers experience stress, focus and awakeness can help inform the design of workplaces to manage and alleviate stress, and to encourage focus and awakeness. An ability to predict stress, focus and awakeness in the moment can enable the development of digital tools to support knowledge workers in their work by better managing disruptions, automatically adjusting lighting to reduce sleepiness, or by avoiding stress, amongst other possibilities.

To the best of our knowledge, this is the longest in-situ study performed in a real-world environment of knowledge workers analyzing an extensive collection of biometric signals with experience samples examining real-time prediction of stress, focus and awakeness in the workplace. Previous studies are predominantly controlled lab studies, which used only a subset of the relevant biometric signals we captured, much more limited in duration, or not focused on a real-world environment Goyal and Fussell (2017); Healey and Picard (2005); Nakagawa et al. (2014); Parnin (2011); Radevski et al. (2015); Sano and Picard (2013); Wijsman et al. (2011); Züger and Fritz (2015).

The participants in our study perform various job functions for a research and development group within a single large corporation. For this study, they wore a novel, state-of-the-art biometric armband comprised of precise sensors[2] which captured heart-, respiration- and skin-related measures with low invasiveness. This modality was chosen to ease longitudinal deployment in the field. We then used machine learning to create classifiers and analyze their ability to predict stress, focus and awakeness levels based on the biometric signs gathered by the sensors. 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 predict, it might 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 show that biometric signals collected from a single minimally invasive device 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 determine that knowledge workers' self-reported levels of stress, focus and awakeness, and their physiological manifestation and prediction can vary substantially between individuals. The three main contributions of our work are:

- A qualitative examination of how knowledge worker stress and awakeness behaves and fluctuates in the wild based on an eight-week field study with 14 office workers.
- The creation and analysis of measures for the automatic monitoring of knowledge workers' stress, focus and awakeness in the workplace based on a machine learning model trained on biometric data and experience samples.
- 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.

2. Related work

The study and prediction approach used in this paper is related to previous studies of stress, focus, and awakeness and studies using biometrics to predict them. We consider related work in each of these categories in turn.

2.1. 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 Piazza et al. (2010). 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 Chandola et al. (2010); Hellhammer et al. (2009).

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 skin temperature van Eekelen et al. (2004); Kataoka et al. (2000); Valentini and Parati (2010). In particular, changes in heart rate variability are associated with cognitive and emotional stress Dishman et al. (2000); McDuff et al. (2016). This second approach has been used successfully by several past studies Gal and Vuksanovic (2007); Montano et al. (2009); of the European Society of Cardiology the North American Society of Pacing Electrophysiology (1996).

Zaman et al. Zaman et al. (2019) performed a study in which they measure stress and productivity in short sessions with 63 participants using biometric sensors (thermal facial camera, wrist EDA, chest breathing sensor, and facial camera). Different from this study, our corpus is gathered from real-world office workers. Also, our study spans for eight weeks allowing us to analyze the behavior or our participants through the weeks and providing a more extensive in-depth analysis of our participants over time.

Hovsepian et al. Hovsepian et al. (2015) 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. Hernandez et al. (2014) investigated the use of a pressure-sensitive keyboard and a capacitive 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. McDuff et al. (2016) experimented with a camera to measure photoplethysmographic signals indicative of cognitive stress. Vizer et al. Vizer et al. (2009) 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.

Kocielnik et al. Kocielnik et al. (2013) 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

[2] Biovotion Everion Biovotion (2019)

participants stress profiles or make predictions in real time. Also using a wristband, Hernandez et al. Hernandez et al. (2011) studied the stress level during calls in a call center over a seven day period. They collected skin conductance measures and examined the interpersonal variability of reporting stress.

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 that 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.

Rather than trying to measure or predict stress, several studies have induced stress and examined its effect on work performance, motivation and others. For instance, Sarsenbayeva et al. Sarsenbayeva et al. (2019) induced stress and studied how it affects mobile interaction, showing that it can reduce completion time and accuracy during target acquisition tasks. Evans and Johnson Evans and Johnson (2000) 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 but found no difference in cortisol levels. 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 also been analyzed in call centers Hernandez et al. (2011) and in the context of the perceived imbalance between resources and demands Cherniss (1980). These studies considered several factors such as personality traits, career-related goals and attitudes, as well as life outside of work, and examined their correlation with stress levels and burnout.

Finally, there is also some work that already used proprietary stress measures for other types of classification. For example, Mirjafari et al. Mirjafari et al. (2019) used a proprietary stress measure by Garmin together with other data collected through the sensing of mobile devices and built machine learning models to differentiate between low and high performing workers.

2.2. *Focus*

Focus refers to the allocation of limited cognitive processing resources Anderson (2004). Mark et al. Mark et al. (2014) studied attentional states, including focus, for workplace activities by analyzing the digital activity of 32 information workers in situ for 5 days. 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 barriers 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-workersChong and Siino (2006); González and Mark (2004); Iqbal and Horvitz (2007). 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 Bailey et al. (2001); Czerwinski et al. (2000); Mark et al. (2008). External interruptions may cause workers to enter a “chain of distraction” Iqbal and Horvitz (2007). This chain is composed by stages of preparation, diversion, resumption and recovery that result in time away from an ongoing task. Since interruptions can have a large impact on the focus and productivity of office workers, several studies have examined the prediction of interruptibility—the availability for interruptions—using a variety of features, including computer interaction and biometrics Bailey and Iqbal (2008); Chen and Vertegaal (2004); Fogarty et al. (2005a); Iqbal and Bailey (2008); Züger and Fritz (2015); Zuger et al. (2018). Most of these studies were again conducted for small and controlled tasks over shorter periods of time.

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 Agarwal and Karahanna (2000). Cognitive engagement is described Webster and Ho (1997) as a period of strong focus in an activity without the feeling of a sense of control of the situation. Flow Csikszentmihalyi (1990), and mindfulness Dane (2011); Weick and Sutcliffe (2006) 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.

2.3. *Awakeness*

Sleepiness (lack of awakeness) and its associated risk of serious injury to passengers has been studied in the context of automobile accidents Connor et al. (2002); Nordbakke and Sagberg (2007). These studies show a strong association between the level of acute driver sleepiness and the risk of injury crash. Connor et al. Connor et al. (2002) 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 Nordbakke and Sagberg (2007) 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.

2.4. *Biometrics*

Our study investigates the prediction of stress, focus, and awakeness using two different types of measurements, and biometric signals over eight weeks in a real-life office setting. Existing work Goyal and Fussell (2017); Healey and Picard (2005); Nakagawa et al. (2014); Parnin (2011); Radevski et al. (2015); Sano and Picard (2013); Wijsman et al. (2011); Züger and Fritz (2015)—some of it already mentioned above—analyzes a broad array of biometric signals and correlates them with individual’s cognitive states and processes. For example, Zuger et al. Zuger et al. (2018) used biometrics to sense interruptibility in the office Zuger et al. (2018). Biometric signals have also been studied in the context of technology users. For example, Parnin Parnin (2011) 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 Fritz et al. (2014) and using Near Infrared Spectroscopy to measure developer’s cerebral blood flow Nakagawa et al. (2014).

Eye tracking technology Bednarik and Tukiainen (2006); Crosby and Stelovsky (1990); Rodeghero et al. (2014) and brain activity Ikutani and Uwano (2014); Siegmund et al. (2014) have been used in previous studies to examine 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 Beatty (1982). Similar studies have shown high correlation between pupil size and mental workload of subtasks Beatty (1982) and cognitive load Klingner (2010). Brain activity has been associated with different mental states Berger (1929) 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 Smith and Gevins (2005); Sterman et al. (1993). 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. Similarly, we considered adding additional metrics such as the Depression, Anxiety and Stress Scale (DASS) and the Perceived Stress Scale (PSS) Ferdous et al. (2015) but opted for the ones presented in the paper after several discussions with experts in the area, and after piloting surveys to maximize participant compliance.

3. 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 and awakeness over time and the feasibility of predicting stress, focus, and awakeness using biometric signals.

3.1. Participants

We recruited a group of 14 professionals via personal contacts from a large power and automation company. The group is diverse in terms of age, work experience, gender, and work responsibilities. 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, including part-time laboratory workers, and 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.

From 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 average number of years of professional experience of the participants is 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.

3.2. Procedure

We performed a study over the course of eight weeks. This study includes the collection of the participant's biometric data and self-reporting surveys during their work hours. We also collected computer interaction data, but given privacy concerns of some participants chose not to use this data (see Section 3.3.2). We informed participants about the study purpose and procedure, handed out biometric sensors, introduced and explained the self-reporting to the participants, handed out consent forms and ensured informed consent from each participant for the study.

After the initial setup, participants were asked to fill out three surveys each day for the following eight weeks (see Section 3.3.3) and wear the biometric sensor during their work hours. 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.

Keeping participants engaged in an eight-week study can be difficult. In the second month of the study, largely to incentivize participants to continue, we offered participants two one-hour Tai Chi classes per week (each in the middle of the day) and asked them to attend one class a week. The choice of Tai Chi provided a link to mindfulness, a topic of growing interest in communities. Consultations we held with researchers in psychiatry and psychology suggested Tai Chi as a good link to mindfulness and an intervention that might alleviate stress. By recording attendance of participants at the Tai Chi sessions, we are able to analyze the impact of the sessions on the results. As a brief summary, the impact of the sessions was minimal; we present an analysis of the effects in Section 6.

3.3. Data collection

We collected three datasets from each participant as described below:

3.3.1. 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.[3] Previous studies Goyal and Fussell (2017); Healey and Picard (2005); Sano and Picard (2013); Wijsman et al. (2011); Züger and Fritz (2015); Zuger et al. (2018) have used similar devices Electro (2019); Fitbit (2019); Okada et al. (2011) to capture psycho-physiological and biometric measurements for shorter periods of time or capturing a smaller number of measurements. A comparative study has shown that the Everion can be used as a valid proxy for HRV metrics for knowledge workers Barrios et al. (2019).

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 is uploaded to a server, from which we downloaded the data for use in our study. We chose these biometric measurements based on previous research which indicates their potential (see references in Table 1) and the availability and feasibility of collecting these measurements with low invasiveness over a long duration.

3.3.2. Computer interaction data

To gain a better understanding of our participants day-to-day work activities, we asked participants to install an open source computer interaction monitor Meyer et al. (2017). 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).

3.3.3. Surveys

Following guidelines from previous studies Lalle et al. (2016); Luo et al. (2018); Panwar and Collins (2018) and following the preferences of extensive user piloting, we sent via text message a survey request to each participant two times per workday. Pilot participants preferred the usage of text message, in part, due to them being accessible and noticeable anywhere in the office.

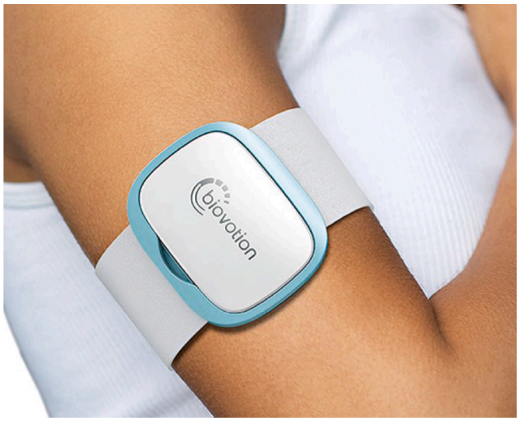


Fig. 1. We used Biovotion's Everion to collect biometric measurements from the participants.

[3] https://biovotion.zendesk.com/hc/en-us/articles/213613165

Table 1
Biometric measurements captured by the Everion, organized by category and with references to previous works using similar data. *(RMSSD denotes the root mean square of successive heartbeat interval differences)*.

Biometric Measurement	Units of Measure
Physical Activity	Aldana et al. (1996); Fox (1999)
Intensity of motion	(No unit)
Energy Expenditure	Calories per second (cal/s)
Step counter	Steps
Heart	Haag et al. (2004); Haapalainen et al. (2010); Healey and Picard (2005); Mulder (1992)
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	Haag et al. (2004); Healey and Picard (2005)
Galvanic skin response	kOhm
Skin temperature	Degrees Celsius (°C)
Respiration	Haag et al. (2004); Healey and Picard (2005); Masaoka and Homma (1997); Mulder (1992)
Respiratory rate	Breaths per Minute (bpm)

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 Adams et al. (2018) since this would discourage participation in stressed time frames and would bias the corpus. The same survey was sent each time consisting of the following three questions:

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 wording of the questions is based on a previous survey of individuals in an organizational context Gloor et al. (2010). The use of awakeness (rather than sleepiness) in Question 1 is inspired by previous work Wilhelm and Schoebi (2007) and to some extent also captures the “arousal” aspect of the affective space Russell (1980).

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 workday?

Following guidelines from similar previous studies Fogarty et al. (2005b); Tanaka and Fujita (2011), 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.

We did not ask our participants about their focus levels in the end of the day survey, as focus is more of an in-the-moment aspect than stress and awakeness.

4. Observed trends over time in stress and awakeness levels

To gain insights into how knowledge workers experience stress and awakeness 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. As we noted in the last section, we did not ask participants about focus in the end of day surveys as focus is an aspect relevant at a particular moment in time rather than an aspect for an extended period of work. 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.

4.1. Stress Levels

Overall, we identified three prominent characteristics in the stress levels.

4.1.1. 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. Seven participants (54%) reported feeling average stress levels most frequently (rating 3 on our scale), while five (38%) reported feeling little stress (rating 2) and one (8%) reported feeling no stress at all (rating 1). Figure 2 illustrates these points using data from two participants, showing the 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). Gaps in the chart for S1 represent days for which the participant did not report their stress level.

4.1.2. 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 one 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.

4.1.3. 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 one 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 participant’s 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.

4.2. Awakeness

We applied the same analyses described above to the self-reported awakeness levels of our participants. Compared with the reported

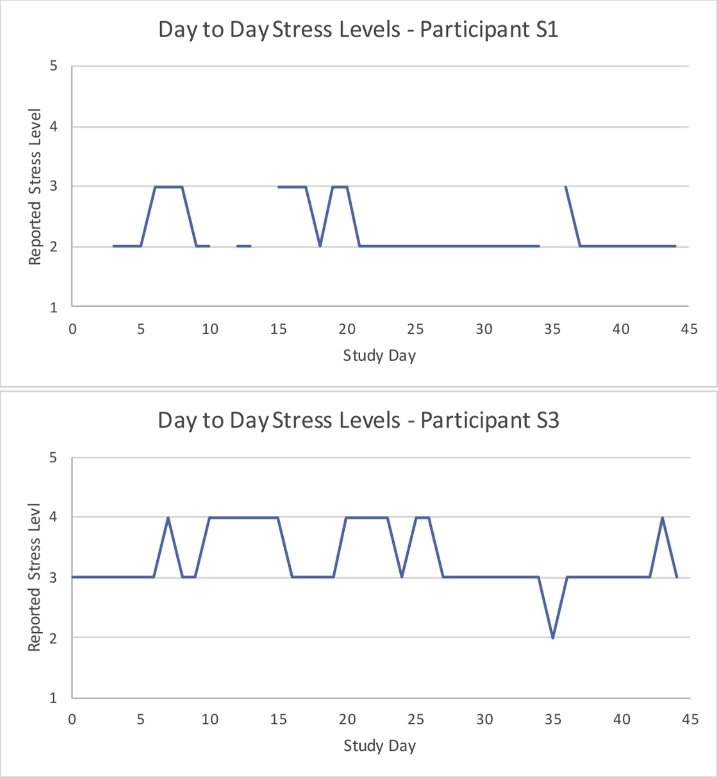


Fig. 2. The day-to-day stress levels reported by two participants (S1 and S3) are shown. The y-axis represents values on the 5-point Likert scale we asked participants to respond with ranging from 1/Not at all stressed to 5/Extremely stressed. The x-axis represents the day of the study on which the value was recorded (from 0-45). Some gaps are present in the chart for S1 as they did not report their stress level on those days.

stress levels, we observed the same trend of participants reporting one baseline far more commonly than any other. We did not find there to be a significant correlation between reported stress and awakeness levels. Overall, participant’s awakeness levels fluctuated significantly less than their stress levels (73.1% of reports were at the baseline level, compared to 65.1% for stress, $p < 0.05$). Participants were unlikely to experience days with heightened (above baseline) awakeness. Such days made up only 6.0% of the total observed workdays across all participant. Most of these days came from one participant, S2, who reported 15 heightened awakeness days compared to the next highest, S13 with four. Large deviations (>1 point deviation) from each participants baseline awakeness levels were extremely uncommon, accounting for only nine (2.1%) of our total observations. Similarly to what we observed with high stress days, low awakeness days frequently came in clusters of two or three days in a row. Given that the immediately preceding day was a low awakeness day; a given day was 228.2% more likely to be a low awakeness day than when considering any day at random ($p < 0.0001$).

Table 2
Fixed effects estimates (F.E.E.), 95% confidence intervals (C.I.) and associated p-values for the explanatory variables (day of week and proximity to month end) that we examined in our linear mixed model analysis.

	Stress			Awakeness		
Variable	**F.E.E.**	**p-value**	**C.I. (95%)**	**F.E.E.**	**p-value**	**C.I. (95%)**
Monday	-0.033	0.745	(− 0.234, 0.167)	-0.058	0.497	(− 0.227, 0.110)
Tuesday	0.028	0.773	(− 0.164, 0.221)	0.036	0.662	(− 0.126, 0.198)
Wednesday	0.163	0.100	(− 0.031, 0.357)	-0.018	0.830	(− 0.181, 0.145)
Thursday	0.022	0.821	(− 0.169, 0.213)	0.076	0.927	(− 0.085, 0.238)
Friday	0.082	0.473	(− 0.142, 0.306)	-0.178	0.025	(− 0.334, − 0.022)
Month End	0.044	0.687	(− 0.171, 0.259)	-0.082	0.374	(− 0.263, 0.099)

4.3. Explaining fluctuations

In an attempt to explain some of the fluctuations in stress and awakeness that our participants were experiencing, we created a linear mixed model with the self-reported daily stress and awakeness levels as dependent variables 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. Ultimately, the analysis showed that none of the variables that we examined had a significant explanatory power with respect to our participants perceived stress levels. For awakeness, we found that there was a small (fixed effects estimate: -0.178) yet significant decrease in awakeness levels on Fridays in particular. Table 2 shows some of the detailed results of these analyses. These results show the difficulty of explaining a person’s stress and awakeness via simple measures, and point to the need for additional instrumentation and data collection if we are to successfully understand and make predictions about these human aspects in the workplace.

5. Predicting stress, focus and awakeness 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 predicting stress, focus and awakeness.

5.1. Data Preparation

In a machine learning context, data preparation and utilization is an essential part of the proposed solution. To prepare the collected data for use in training and testing our proposed machine learning models, we performed the following steps.

5.1.1. Data Cleaning

All data recorded by the Everion is associated with a quality score ranging from 0-1 that is calculated using proprietary methods. In accordance with the recommendations of Biovotion, to prevent our results from being effected by erroneous data we set a quality threshold of 0.5 and discarded any data gathered which had a quality rating below this threshold.

5.1.2. 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 evaluating machine learning models.

To link the data, we look back one hour from the start time of each survey response for available biometric data. For example, if a participant started a survey response at 11:05am, we look for biometric data from between 10:05am to 11:05am. If no biometric data was recorded in the hour time window, we exclude the survey response from the dataset. Otherwise, we consider the survey response to have associated biometric data.

There are several reasons for a survey response to lack associated biometric data:

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

5.1.3. Feature Extraction

We extracted features from the biometric data to provide as input to machine learning models. Previous studies Bernstein and Zurfluh (2005); Züger and Fritz (2015) 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 Zuger et al. (2018), 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. If there is a large portion ($\geq$50%) of data missing (either because of a recording issue or because of low quality data) from the time window considered, then the time window is marked as missing. In this case, features are imputed based on the mean of other samples of the same feature for that participant. This is an effective and commonly used technique, which is preferable to the alternative of deletion as it preserves our already small sample size Hawthorne et al. (2005). 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, 25th percentile, 75th percentile, 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. To attempt to account for inter-participant differences, we normalized all features on a per-participant basis.

5.1.4. Response Transformations

Table 3 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. Given this distribution and based on our earlier observation that the participants tended to adhere to a baseline reporting level for stress and awakeness, we elected to simplify the problem from five classes to two. This transformation enables us to more easily represent patterns in the data, such as when a participant fluctuates from a normal to high stress level. To perform this transformation, we began by calculating the median response value for each participant and each question. We classified each response below the median as 0 ('negative') and each response above the median as 1 ('positive'). The

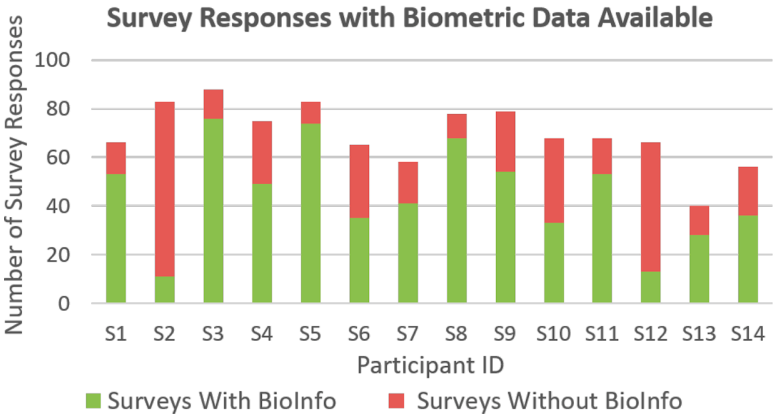
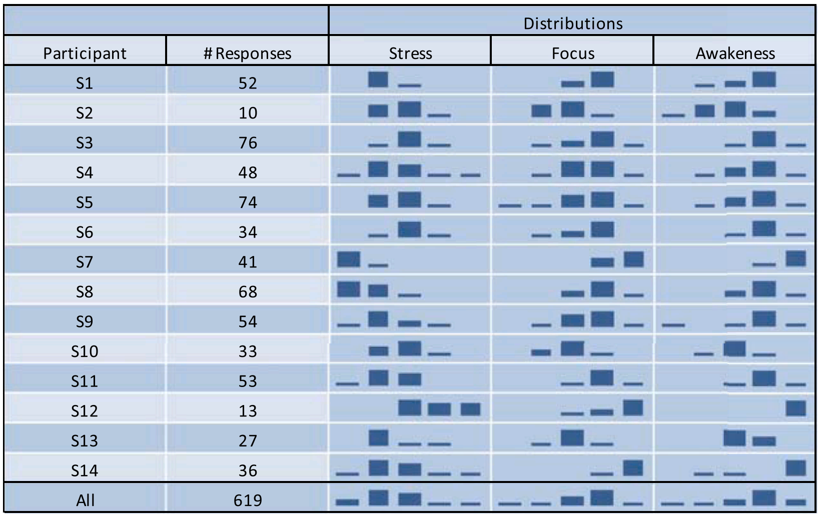


Fig. 3. The figure shows the biometric data available per each participant. Green sections represent survey responses with biometric data available, red sections responses with no biometric data available.

Table 3
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 five 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.



		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			

distribution for the stress question skewed left, so we included the median values in the 'positive' class (i.e. 'stressed'), while the distributions for focus and awakeness skewed right, so we included those median values in the 'negative' class (i.e. 'not focused', 'not awake').

With this method, we transformed the survey data into a two-point scale, representing negative or positive responses for each of the three human aspects of interests (e.g., not stressed or stressed).

5.1.5. 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 mitigate this effect, we applied random oversampling to our training sets, which artificially rebalances the dataset by creating randomly replicated data in the minority class. This technique has commonly been used in previous studies on unbalanced datasets Chawla et al. (2004); Yap et al. (2014).

5.2. Selecting a Classifier Algorithm

Many different algorithms can be used to build a classifier. To select an algorithm, we compared multiple classifiers using the popular machine learning library scikit-learn Pedregosa et al. (2011), evaluating each one by using leave-one-out cross validation. Our analysis showed that random forest outperforms all other classifiers, including Naïve Bayes, decision trees, support vector machine, and a multilayer perceptron neural network. For the remainder of this paper, we refer to a random forest classifier.

5.3. Individual Classifiers

Since peoples' experience of stress, focus, and awakeness (as well as

Table 4
Results of predictions using the individual models. The distribution columns show the proportion of the minority class out of the total number of responses for each of the three variables. The baseline rows represents the averaged results of our baseline classifiers. The general row shows the averaged results of our models trained on all participants.

	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	0.750	0.111	0.166	9/52	0.609	0.193	0.143	14/52	0.628	0.396	0.396	16/52
S2	0.866	0.000	0.000	1/10	0.133	0.063	0.083	4/10	0.367	0.000	0.000	4/10
S3	0.842	0.500	0.222	12/76	0.644	0.476	0.385	26/76	0.965	0.000	0.000	2/76
S4	0.506	0.460	0.440	22/48	0.861	0.000	0.000	4/48	0.701	0.617	0.537	18/48
S5	0.919	0.200	0.067	5/74	0.802	0.196	0.127	11/74	0.658	0.517	0.372	26/74
S6	0.824	0.400	0.400	5/34	0.451	0.143	0.112	12/34	0.902	0.400	0.222	3/34
S7	0.976	0.000	0.000	1/41	0.642	0.571	0.334	16/41	0.934	0.000	0.000	2/41
S8	0.603	0.566	0.548	31/68	0.745	0.531	0.315	18/68	0.755	0.355	0.432	15/68
S9	0.629	0.324	0.229	16/54	0.451	0.348	0.334	23/54	0.741	0.250	0.083	12/54
S10	0.970	0.000	0.000	1/33	0.768	0.611	0.407	9/33	0.848	0.000	0.000	3/33
S11	0.667	0.614	0.530	22/53	0.711	0.056	0.034	10/53	0.792	0.000	0.000	8/53
S12	0.744	0.600	0.500	4/13	0.769	1.000	0.250	4/13	1.000	-	-	0/13
S13	0.926	0.000	0.000	2/27	0.963	0.000	0.000	1/27	1.000	-	-	0/27
S14	0.527	0.526	0.555	18/36	0.834	0.200	0.166	4/36	0.963	0.625	0.833	2/36
Overall	0.768	0.322	0.261	149/619	0.670	0.313	0.192	156/619	0.804	0.263	0.240	111/619
Baseline - Majority	0.735	0.000	0.000		0.743	0.000	0.000		0.833	0.000	0.000	
Baseline - Random	0.716	0.256	0.251		0.650	0.249	0.266		0.755	0.173	0.170	
Improvement Over Random (%)	7.249	25.850	4.188		3.146	25.990	-27.754		6.468	52.368	40.828	
General	0.543	0.246	0.436		0.521	0.268	0.519		0.554	0.189	0.450	
General Baseline - Majority	0.759	0.000	0.000		0.748	0.000	0.000		0.821	0.000	0.000	
General Baseline - Random	0.635	0.255	0.268		0.585	0.198	0.212		0.687	0.151	0.162	

their physiological manifestations) can vary substantially (e.g., Hernandez et al. (2011)), we first trained and evaluated individual classifiers for each participant (as opposed to a general one for all participants) using leave-one-out cross validation. 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 Bhattacharyya et al. (2011); Hernandez et al. (2011); Yap et al. (2014).

Besides our results, for the purposes of comparison, we also present two commonly used baselines in this table - a majority classifier, which always predicts the larger of the two classes considered, and a stratified random classifier which randomly chooses between the two classes, but with a proportional bias towards the larger class.

For some users (i.e., S1. as seen in Table 3), the imbalance in their data was so extreme that even after adjusting by oversampling, we were not able 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. We present a comparison between the averaged results of our individual classifiers and those of the baseline stratified random classifier in the “Improvement Over Random” row of Table 4. This is calculated as the difference between the overall average and the baseline results, divided by the baseline results (for example, $\frac{Acc_{Overall} - Acc_{Random}}{Acc_{Random}}$). We do not compare our results with the majority classifier directly as this classifier achieved precision and recall scores of zero when predicting stress, lack of focus, and lack of awakeness, making a meaningful comparison unfeasible. The improvement percentages demonstrate that the predictions made by our classifiers are much better than random after correcting for the imbalance in our dataset.

While the individually trained classifiers improved on average across all participants upon the baseline in all cases except in recall of ‘not focused’, the improvement was substantially higher for awakeness (52.4% improvement in precision, 40.8% in recall, and 6.5% in accuracy) than for stress or focus. In addition, 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 152.0% in precision and 111.1% 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 imbalanced.

5.4. 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 similar results with respect to accuracy, precision, and recall for the individual models. Ultimately, we elected not to utilize any feature selection in order to simplify our approach, as we found there to be minimal differences in performance between the techniques, and the random forest algorithm is capable of (and robust for) handling datasets with many features.

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. Heart rate variability proved to be an important measure for all of the aspects of interest, ranking as the most important feature category for both stress and awakeness, and the second most important one for focus. This is not surprising, as heart rate variability and skin temperature have been shown in several previous studies to be possible indicators for stress levels Dishman et al. (2000); Kataoka et al. (2000); McDuff et al. (2016). We also found blood pulse wave to be an important indicator for both focus and awakeness, but less important for stress, while respiration rate was important in stress and focus but not awakeness. Besides these mentioned feature categories, there was great variation in which measures were important to which of the three aspects. This shows that there is a clear benefit to having multiple biometric streams available for predicting stress, focus and awakeness.

5.5. 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 (leave-one-participant-out cross-validation), repeating this process for all 14 participants.

The bottom three rows of Table 4 present the averaged performance results for this approach in terms of accuracy, precision, and recall, as well as the results of baseline stratified random and majority classifiers

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

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%

following the same leave-one-group-out cross-validation procedure. 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.

6. 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.

6.1. 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 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 Züger et al. (2017). 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 Karjalainen (2007)). We believe that our results present a good step to more in situ usage of biometric sensing. Future studies can build on the evidence, and for example, reduce the effort required for the experience sampling from continuously rating stress levels to using biometrics as a ground truth with occasional validation of the predicted stress levels, or identify how to balance needs across a group of office workers and how to handle conflicting levels between different group members.

6.2. The Effect of Tai Chi on Stress

As described in our study procedure, due to our focus on stress, we offered participants an opportunity to be part of Tai Chi classes. The two primary reasons for this intervention were to keep participants motivated to continue the self-reports over the long study period, and to offer a technique that might help reduce stress. After consulting with other researchers, we decided to offer Tai Chi in the last month of the study, leaving the first month of the study unaltered. While participants were not required to take a Tai Chi class, they all took one per week except for two participants that opted out of the last two weeks of classes. Four participants explicitly stated that they liked the Tai Chi sessions or that they were "excellent". These classes were performed once per week with a duration of one hour.

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 workdays 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.

6.3. 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 Hernandez et al. (2011); Hovsepian et al. (2015).

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.

6.4. 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 3. 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. Exler et al. (2016). 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. Rebalancing the dataset using such techniques is a common and effective practice Branco et al. (2016). Alternative techniques such as SMOTE Chawla et al. (2002) exist and can perform better than those we employed, however they are impractical considering the limited amount of data we have to work with.

Oversampling the dataset as we did may also lead to an increased risk of model overfitting. However, we believe the benefits of rebalancing the dataset outweigh this possible downside. In light of the 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.

6.5. 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.

7. Threats to Validity

There are numerous threats to validity to our study.

7.1. 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 departments with different age ranges, genders, work experience, and working in different positions therefore providing evidence that our approach's performance is comprised of and can generalize to a wide range of knowledge workers.

Secondly, our results may not generalize 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 worldwide such as controlled temperature and lighting.

7.2. Internal Validity

This study tries to find correlations between biometric features and the human aspects of 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 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 sufficient to draw valid conclusions. To address this threat, we collected a data for an eight-week period, which is 400% longer than the longest previous studies Muller and Fritz (2016); Zuger et al. (2018).

Due to the small amount of data available to us for the purposes of this study, there is a risk that our models may be overfitting to some degree. Such overfitting indicates that the results presented in our work may be less than optimal, however any overfitting applies strictly to the training data and given that there is no overlap between the data used for training and testing, we do not believe this invalidates our results. We leave further optimization of our approach to future researchers.

7.3. 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 a 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 the studied human aspects of stress, focus and awakeness. In an effort to maintain this research applicable to real-world environments, we picked the already existing Everion 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.

A future more thorough statistical analysis of the relationships between the aspects of interest, such as stress, and the physiological data may further provide deeper insights into the data and how it might be used in prediction.

8. 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 these human aspects over time and to examine the ability of biometrics to predict these aspects. The longitudinal and in-situ placement of the study support and extend previous work. Based on daily collected survey responses, we observed that although participants sometimes saw periods of sustained stress or sleepiness, they would always return to their baseline reporting level 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 biometric data with which we were able to create a model that is able to predict user stress, sleepiness, and lack of focus with small improvements in accuracy (from 3.1% to 7.2%, depending on the aspect in question), and moderate improvements in precision (25.9% - 52.4%) in comparison to a stratified random classifier. While the precision and recall scores we report are still low overall, this is a difficult problem to solve and our improvements indicate the potential for future researchers to build upon.

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.

CRediT authorship contribution statement

Mauricio Soto: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Chris Satterfield:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Thomas Fritz:** Conceptualization, Methodology, Investigation, Resources, Writing - review & editing. **Gail C. Murphy:** Conceptualization, Methodology, Writing - review & editing. **David C. Shepherd:** Conceptualization, Methodology, Investigation, Resources, Writing - review & editing. **Nicholas Kraft:** Conceptualization, Methodology, Investigation, Resources, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adams, A.T., Murnane, E.L., Adams, P., Elfenbein, M., Chang, P.F., Sannon, S., Gay, G., Choudhury, T., 2018. Keppi: A tangible user interface for self-reporting pain. CHI Conference on Human Factors in Computing Systems.

Agarwal, R., Karahanna, E., 2000. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. MIS Quarterly, 24 (4), pp. 665–694.

Aldana, S.G., Sutton, L.D., Jacobson, B.H., Quirk, M.G., 1996. Relationships between leisure time physical activity and perceived stress. Perceptual and Motor Skills 82 (1), 315–321. https://doi.org/10.2466/pms.1996.82.1.315.

Anderson, J.R., 2004. Cognitive Psychology and Its Implications.

Bailey, B.P., Iqbal, S.T., 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. ACM Transactions on Computer-Human Interaction (TOCHI) 14 (4), 21.

Bailey, B.P., Konstan, J.A., Carlis, J.V., 2001. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. Proceedings of INTERACT, 1, pp. 593–601.

Barrios, L., Oldrati, P., Santini, S., Lutterotti, A., 2019. Evaluating the accuracy of heart rate sensors based on photoplethysmography for in-the-wild analysis. Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare. Association for Computing Machinery, New York, NY, USA, pp. 251–261. https://doi.org/10.1145/3329189.3329215.

Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin, 91(2).

Bednarik, R., Tukiainen, M., 2006. An eye-tracking methodology for characterizing program comprehension processes. Proc. of ETRA.

Berger, H., 1929. Uber das elektrenkephalogramm des menschen. European Archives of Psychiatry and Clinical Neuroscience, 87, pp. 527–570.

Bernstein, P.V.A., Zurfluh, A., 2005. Interruptability prediction using motion detection. Workshop on Managing Context Information in Mobile and Pervasive Environments.

Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C., 2011. Data mining for credit card fraud: A comparative study. Decision Support Systems 50 (3), 602–613.

Biovotion, 2019. Everion. https://www.biovotion.com/everion/. [Online; accessed 9-July-2019].

Branco, P., Torgo, L., Ribeiro, R.P., 2016. A survey of predictive modeling on imbalanced domains. ACM Computing Surveys (CSUR) 49 (2), 1–50.

Chandola, T., Heraclides, A., Kumari, M., 2010. Psychophysiological biomarkers of workplace stressors. Neurosci Biobehav Rev., 35, pp. 51–57.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Chawla, N.V., Japkowicz, N., Kotcz, A., 2004. Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explor. Newsl. 6 (1), 1–6. https://doi.org/10.1145/1007730.1007733.

Chen, D., Vertegaal, R., 2004. Using mental load for managing interruptions in physiologically attentive user interfaces. CHI'04 extended abstracts on Human factors in computing systems. ACM, pp. 1513–1516.

Cherniss, C., 1980. Staff Burnout - Job Stress in the Human Services. Sage Publications, Inc.

Chong, J., Siino, R., 2006. Interruptions on software teams: a comparison of paired and solo programmers. Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work. ACM, pp. 29–38.

Cohen, S., Kessler, R.C., Gordon, L.U., 1997. Measuring stress: A guide for health and social scientists. Oxford University Press on Demand.

Connor, J., Ameratunga, S., Norton, R., Robinson, E., Dunn, R., Bailey, J., Civil, I., Jackson, R., 2002. Driver sleepiness and risk of serious injury to car occupants: population based case control study. BMJ.

Crosby, M., Stelovsky, J., 1990. How do we read algorithms? a case study. Computer 23 (1).

Csikszentmihalyi, M., 1990. Flow: The Psychology of Optimal Experience.

Czerwinski, M., Cutrell, E., Horvitz, E., 2000. Instant messaging: Effects of relevance and timing. People and computers XIV: Proceedings of HCI, 2. British Computer Society, pp. 71–76.

Czerwinski, M., Horvitz, E., Wilhite, S., 2004. A diary study of task switching and interruptions. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 175–182.

Dane, E., 2011. Paying attention to mindfulness and its effect on task performance in the workplace. Journal of Management, 37 (4), pp. 997–1018.

Dishman, R.K., Nakamura, Y., Garcia, M.E., Thompson, R.W., Dunn, A.L., Blair, S.N., 2000. Heart rate variability, trait anxiety, and perceived stress among physically fit men and women. International Journal of Psychophysiology 37 (2), 121–133.

Drucker, P.F., 1999. Knowledge-worker productivity: The biggest challenge. California management review 41 (2), 79–94.

van Eekelen, A.P.J., Houtveen, J.H., Kerkhof, G.A., 2004. Circadian variation in base rate measures of cardiac autonomic activity. European Journal of Applied Physiology, 93, pp. 39–46.

Electro, P., 2019. Equine H7 heart rate sensor belt set. https://www.polar.com/en/products/equine/accessories/equine_H7_heart_rate_sensor_belt_set. [Online; accessed 9-July-2019].

Evans, G.W., Johnson, D., 2000. Stress and open-office noise. Journal of Applied Psychology, 25(5), pp. 779–783.

Exler, A., Schankin, A., Klebsattel, C., Beigl, M., 2016. A wearable system for mood assessment considering smartphone features and data from mobile ecgs. Pervasive and Ubiquitous Computing, pp. 1153–1161.

Ferdous, R., Osmani, V., Mayora, O., 2015. Smartphone app usage as a predictor of perceived stress levels at workplace. International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth).

Fitbit, I., 2019. Fitbit Charge 2. https://www.fitbit.com/de/charge2. [Online; accessed 9-July-2019].

Fogarty, J., Ko, A.J., Aung, H.H., Golden, E., Tang, K.P., Hudson, S.E., 2005. Examining task engagement in sensor-based statistical models of human interruptibility. Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM, pp. 331–340.

Fogarty, J., Ko, A.J., Aung, H.H., Golden, E., Tang, K.P., Hudson, S.E., 2005. Examining task engagement in sensor-based statistical models of human interruptibility. SIGCHI Conference on Human Factors in Computing Systems, pp. 331–340.

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

Fritz, T., Begel, A., Müller, S.C., Yigit-Elliott, S., Züger, M., 2014. Using psycho-physiological measures to assess task difficulty in software development. Proceedings of the 36th International Conference on Software Engineering. ACM, pp. 402–413.

Gal, V., Vuksanovic, V., 2007. Heart rate variability in mental stress aloud. Medical Engineering and Physics, 29, pp. 344–349.

Gloor, P., Oster, D., Raz, O., Pentland, A., Schoder, D., 2010. The virtual mirror: Reflecting on the social and psychological self to increase organizational creativity. International Studies of Management & Organization 40 (2), 74–94.

González, V.M., Mark, G., 2004. Constant, constant, multi-tasking craziness: managing multiple working spheres. Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, pp. 113–120.

Goyal, N., Fussell, S.R., 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), 52.

Haag, A., Goronzy, S., Schaich, P., Williams, J., 2004. Emotion recognition using bio-sensors: First steps towards an automatic system. Tutorial and research workshop on affective dialogue systems. Springer, pp. 36–48.

Haapalainen, E., Kim, S., Forlizzi, J.F., Dey, A.K., 2010. Psycho-physiological measures for assessing cognitive load. Proceedings of the 12th ACM international conference on Ubiquitous computing. ACM, pp. 301–310.

Hawthorne, G., Hawthorne, G., Elliott, P., 2005. Imputing cross-sectional missing data: comparison of common techniques. Australian & New Zealand Journal of Psychiatry 39 (7), 583–590.

Healey, J.A., Picard, R.W., 2005. Detecting stress during real-world driving tasks using physiological sensors. Transactions on intelligent transportation systems 6 (2), 156–166.

Hellhammer, D.H., Wust, S., Kudielka, B.M., 2009. Salivary cortisol as a biomarker in stress research. Psychoneuroendocrinology, 34, pp. 163–171.

Hernandez, J., Morris, R.R., Picard, R.W., 2011. Call center stress recognition with person-specific models. International Conference on Affective Computing and Intelligent Interaction, pp. 125–134.

Hernandez, J., Paredes, P., Roseway, A., Czerwinski, M., 2014. Under pressure: Sensing stress of computer users. CHI Conference on Human Factors in Computing Systems.

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

Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., Kumar, S., 2015. cstress: Towards a gold standard for continuous stress assessment in the mobile environment. ACM international conference on Ubiquitous computing, pp. 493–504.

Ikutani, Y., Uwano, H., 2014. Brain activity measurement during program comprehension with NIRS. Proc. of SNPD.

Iqbal, S.T., Bailey, B.P., 2008. Effects of intelligent notification management on users and their tasks. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 93–102.

Iqbal, S.T., Horvitz, E., 2007. Disruption and recovery of computing tasks: Field study, analysis and directions. SIGCHI Conference on Human Factors in Computing Systems, pp. 677–686.

Karjalainen, S., 2007. Gender differences in thermal comfort and use of thermostats in everyday thermal environments. Building and Environment, 42, pp. 1594–1603.

Kataoka, H., Kano, H., Yoshida, H., Saijo, A., Osumi, M.Y..M., 2000. Development of a skin temperature measuring system for non-contact stress evaluation. IEEE Engineering in Medicine and Biology Society, 20.

Klingner, J., 2010. Fixation-aligned pupillary response averaging. Symposium on Eye-Tracking Research and Applications, pp. 275–282.

Kocielnik, R., Sidorova, N., Maggi, F.M., Ouwerkerk, M., Westerink, J.H.D.M., 2013. Smart technologies for long-term stress monitoring at work. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems 53–58. https://doi.org/10.1109/cbms.2013.6627764.

Lalle, S., Conati, C., Carenini, G., 2016. Predicting confusion in information visualization from eye tracking and interaction data. International Joint Conference on Artificial Intelligence.

for the Improvement of Living, E.F., Conditions, W., 2010. Work-related stress. Technical Report.https://www.eurofound.europa.eu/printpdf/publications/report/2010/work-related-stress

Luo, Y., Lee, B., Wohn, D.Y., Rebar, A.L., Conroy, D.E., Choe, E.K., 2018. Time for break: Understanding information workers' sedentary behavior through a break prompting system. CHI Conference on Human Factors in Computing Systems.

Mark, G., Gudith, D., Klocke, U., 2008. The cost of interrupted work: more speed and stress. Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM, pp. 107–110.

Mark, G., Iqbal, S.T., Czerwinski, M., Johns, P., 2014. Bored mondays and focused afternoons: the rhythm of attention and online activity in the workplace. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 3025–3034.

Masaoka, Y., Homma, I., 1997. Anxiety and respiratory patterns: their relationship during mental stress and physical load. International Journal of Psychophysiology 27 (2), 153–159.

McDuff, D.J., Hernandez, J., Gontarek, S., Picard, R.W., 2016. Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera.

Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, pp. 4000–4004.

Meyer, A.N., Murphy, G.C., Zimmermann, T., Fritz, T., 2017. Design recommendations for self-monitoring in the workplace: Studies in software development. Proceedings ACM Human Computer Interaction, p. 24. https://doi.org/10.1145/3134714.

Mirjafari, S., Masaba, K., Grover, T., Wang, W., Audia, P.G., Campbell, A.T., Chawla, N. V., Swain, V.D., Choudhury, M.D., Dey, A.K., D'Mello, S.K., Gao, G., Gregg, J.M., Jagannath, K., Jiang, K., Lin, S., Liu, Q., Mark, G.J., Martinez, G.J., Mattingly, S.M., Moskal, E., Mulukutla, R., Nepal, S., Nies, K.A., Reddy, M.D., Robles-Granda, P.D., Saha, K., Sirigiri, A., Striegel, A.D., 2019. Differentiating higher and lower job performers in the workplace using mobile sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, pp. 1–24.

Montano, N., Porta, A., Cogliati, C., Costantino, G., Tobaldini, E., Casali, K.R., Iellamo, F., 2009. Heart rate variability explored in the frequency domain: A tool to investigate the link between heart and behavior. Neuroscience and Biobehavioral Reviews, 33, pp. 71–80.

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

Muller, S., Fritz, T., 2016. Using (bio)metrics to predict code quality online. In Proceedings of the ICSE.

Nakagawa, T., Kamei, Y., Uwano, H., Monden, A., Matsumoto, K., German, D.M., 2014. Quantifying programmers' mental workload during program comprehension based on cerebral blood flow measurement: A controlled experiment. Companion Proc. of ICSE.

Nordbakke, S., Sagberg, F., 2007. Sleepy at the wheel: Knowledge, symptoms and behaviour among car drivers. Transportation Research Part F: Traffic Psychology and Behaviour, 10, pp. 1–10.

Okada, Y., Yoto, T.Y., aki Suzuki, T., Sakuragawa, S., Sakakibara, H., Shimoi, K., Sugiura, T., 2011. Wearable ecg recorder with acceleration sensors for monitoring daily stress*: Office work simulation study. Journal of Medical and Biological Engineering, 34, pp. 420–426.

of the European Society of Cardiology the North American Society of Pacing Electrophysiology, T.F., 1996. Heart rate variability, standards of measurement, physiological interpretation, and clinical use. European Heart Journal, 93, pp. 1043–1065.

Panwar, P., Collins, C.M., 2018. Detecting negative emotion for mixed initiative visual analytics. CHI Conference on Human Factors in Computing Systems.

Parnin, C., 2011. Subvocalization - toward hearing the inner thoughts of developers. Proceedings of International Conference on Program Comprehension.

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

Piazza, J.R., Almeida, D.M., Dmitrieva, N.O., Klein, L.C., 2010. Frontiers in the use of biomarkers of health in research on stress and aging. 35th Annual International Conference of the IEEE EMBS, 65B (5), pp. 513–525.

Radevski, S., Hata, H., Matsumoto, K., 2015. Real-time monitoring of neural state in assessing and improving software developers' productivity. Proceedings of Connected Health: Applications, Systems and Engineering Technologies.

Rodeghero, P., McMillan, C., McBurney, P.W., Bosch, N., D'Mello, S., 2014. Improving automated source code summarization via an eye-tracking study of programmers. International Converence on Software Engineering.

Russell, J., 1980. A circumplex model of affect. Journal of Personality and Social Psychology 39 (6), 1161–1178.

Sano, A., Picard, R.W., 2013. Stress recognition using wearable sensors and mobile phones. Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, pp. 671–676.

Sarsenbayeva, Z., van Berkel, N., Hettiachchi, D., Jiang, W., Dingler, T., Velloso, E., Kostakos, V., Gonçalves, J., 2019. Measuring the effects of stress on mobile interaction. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, pp. 1–18.

Setz, C., Arnrich, B., Schumm, J., Marca, R.L., Troster, G., Ehlert, U., 2010. Discriminating stress from cognitive load using a wearable eda device. IEEE Transactions on Information Technology in Biomedicine 14 (2), 410–417.

Siegmund, J., Kästner, C., Apel, S., Parnin, C., Bethmann, A., Leich, T., Saake, G., Brechmann, A., 2014. Understanding source code with functional magnetic resonance imaging. Proceedings of International Conference on Software Engineering.

Smith, M.E., Gevins, A., 2005. Neurophysiologic monitoring of mental workload and fatigue during operation of a flight simulator. The International Society for Optical Engineering, pp. 116–126.

Sterman, M.B., Mann, C.A., Kaiser, D.A., 1993. Quantitative eeg patterns of differential in-flight workload. 6th Annual Workshop on Space Operations Applications and Research, 2.

Tanaka, T., Fujita, K., 2011. Study of user interruptibility estimation based on focused application switching. Conference on Computer Supported Cooperative Work, pp. 721–724.

Valentini, M., Parati, G., 2010. Variables influencing heart rate. Progress in Cardiovascular Diseases, 52, pp. 11–19.

Vizer, L.M., Zhou, L., Sears, A., 2009. Automated stress detection using keystroke and linguistic features: An exploratory study. International Journal of Human-Computer Studies 67 (10), 870–886. https://doi.org/10.1016/j.ijhcs.2009.07.005.

Webster, J., Ho, H., 1997. Audience engagement in multimedia presentations. Data Base for the Advancement in Information Systems, 28(2), pp. 63–77.

Weick, K.E., Sutcliffe, K.M., 2006. Mindfulness and the quality of organizational attention. Organization Science, 17, pp. 514–524.

Wijsman, J., Grundlehner, B., Liu, H., Hermens, H., Penders, J., 2011. Towards mental stress detection using wearable physiological sensors. Engineering in Medicine and Biology Society. IEEE, pp. 1798–1801.

Wilhelm, P., Schoebi, D., 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), 258–267.

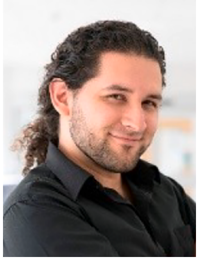
Yap, B.W., Rani, K.A., Rahman, H.A.A., Fong, S., Khairudin, Z., Abdullah, N.N., 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: Herawan, T., Deris, M.M., Abawajy, J. (Eds.), Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Springer Singapore, pp. 13–22.

Zaman, S., Wesley, A., Silva, D.R.D., Buddharaju, P., Akbar, F., Gao, G., Mark, G., Gutierrez-Osuna, R., Pavlidis, I., 2019. Stress and productivity patterns of interrupted, synergistic, and antagonistic office activities. Scientific data, pp. 1–18.

Züger, M., Corley, C., Meyer, A.N., Li, B., Fritz, T., Shepherd, D., Augustine, V., Francis, P., Kraft, N., Snipes, W., 2017. Reducing interruptions at work: A large-scale field study of flowlight. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, pp. 61–72.

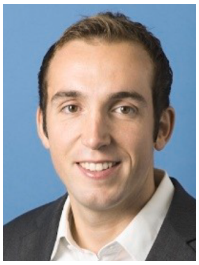
Züger, M., Fritz, T., 2015. Interruptibility of software developers and its prediction using psycho-physiological sensors. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, pp. 2981–2990.

Zuger, M., Muller, S., Meyer, A., Fritz, T., 2018. Sensing interruptibility in the office: A field study on the use of biometric and computer interaction sensor. Conference on Human Factors in Computing Systems.



Mauricio Soto Senior Machine Learning R&D Engineer at Hitachi ABB Power Grids. His Ph.D. work encompasses primarily problems relating automatic error repair, data mining, and machine learning. His thesis focuses on creating automatic program repair approaches to find higher quality patches for real bugs in existing software. Currently Mauricio focuses on applied machine learning using time series data. PC Member for ESEM '20 Industry Track, MSR '20 Challenge Track, EMSE '19 Research Track, and MSR '19 Challenge Track.

Chris Satterfield M.Sc. in Computer Science from the Software Practices Lab at the University of British Columbia. His research focuses on the automatic identification and summarization of the tasks that professionals perform throughout their day. He is also interested how biometric sensors may be used to provide insights into developer productivity.



Thomas Fritz Associate Professor in the Department of Informatics at the University of Zurich. In his research, he focuses on empirically studying software developers and on using personal and biometric data to improve software developers' productivity. By better understanding what software developers need, what they experience, and how they operate on a daily basis, we will be able to provide better and more tailored support to developers as well as improve their productivity and the quality of the software they produce. In particular, Thomas is interested in three areas: developer productivity, biometric sensing and information needs.

Gail Murphy Full Professor and the founder of the Software Practices Lab. She is also the Vice-President Research & Innovation for UBC and a co-founder and Director at Tasktop Technologies. Gail's research interests are in software engineering with a particular interest in improving the productivity of knowledge workers, including software developers. Her group develops tools to aid with the evolution of large software systems and performs empirical studies to better understand how developers work and how software is developed.

David Shepherd Associate Professor in the Department of Computer Science at Virginia Commonwealth University. His research has produced tools that have been used by thousands, innovations that have been featured in the popular press, and practical ideas that have won business plan competitions. Co-Editor-in-Chief of the Journal of Systems & Software. Program Co-Chair for FSE 2017 industrial track, ICSE 2017, and ICSME 2015. His current work focuses on enabling end-user programming for industrial machines and increasing diversity in computer science.

Nicholas Kraft Principal software engineer at UserVoice in Raleigh. His previous positions include software engineering researcher at ABB Corporate Research, part-time teaching assistant professor in the Department of Computer Science at NC State University, and associate professor (with tenure) in the Department of Computer Science at The University of Alabama. PC member for ICSE'20, ASE'19 Tool Demos, ICSME'19 Doc-Sym, ESEC/FSE'19 Industry