Sensing Interruptibility in the Office: A Field Study on the Use of Biometric and Computer Interaction Sensors

Züger, Manuela ; Müller, Sebastian ; Meyer, André ; Fritz, Thomas

Abstract: Knowledge workers experience many interruptions during their work day. Especially when they happen at inopportune moments, interruptions can incur high costs, cause time loss and frustration. Knowing a person's interruptibility allows optimizing the timing of interruptions and minimize disruption. Recent advances in technology provide the opportunity to collect a wide variety of data on knowledge workers to predict interruptibility. While prior work predominantly examined interruptibility based on a single data type and in short lab studies, we conducted a two-week field study with 13 professional software developers to investigate a variety of computer interaction, heart-, sleep-, and physical activity-related data. Our analysis shows that computer interaction data is more accurate in predicting interruptibility at the computer than biometric data (74.8% vs. 68.3% accuracy), and that combining both yields the best results (75.7% accuracy). We discuss our findings and their practical applicability also in light of collected qualitative data.

DOI: https://doi.org/10.1145/3173574.3174165

Originally published at:
Züger, Manuela; Müller, Sebastian; Meyer, André; Fritz, Thomas (2018). Sensing Interruptibility in the Office: A Field Study on the Use of Biometric and Computer Interaction Sensors. In: CHI 2018, Montreal, 21 April 2018 - 26 April 2018.
DOI: https://doi.org/10.1145/3173574.3174165

Sensing Interruptibility in the Office: A Field Study on the Use of Biometric and Computer Interaction Sensors

Manuela Züger, Sebastian C. Müller, André N. Meyer, Thomas Fritz
University of Zurich, Zurich, Switzerland
{zueger, smueller, ameyer, fritz}@ifi.uzh.ch

ABSTRACT

Knowledge workers experience many interruptions during their work day. Especially when they happen at inopportune moments, interruptions can incur high costs, cause time loss and frustration. Knowing a person's interruptibility allows optimizing the timing of interruptions and minimize disruption. Recent advances in technology provide the opportunity to collect a wide variety of data on knowledge workers to predict interruptibility. While prior work predominantly examined interruptibility based on a single data type and in short lab studies, we conducted a two-week field study with 13 professional software developers to investigate a variety of computer interaction, heart-, sleep-, and physical activity-related data. Our analysis shows that computer interaction data is more accurate in predicting interruptibility at the computer than biometric data (74.8% vs. 68.3% accuracy), and that combining both yields the best results (75.7% accuracy). We discuss our findings and their practical applicability also in light of collected qualitative data.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

interruptibility, field study, biometric sensors, computer interaction

INTRODUCTION

In today's collaborative work environments, knowledge workers are constantly facing interruptions, such as instant message alerts, emails or a co-worker asking a question in person [24, 15, 38]. Many of these interruptions are necessary to share knowledge and resolve problems quickly [39]. Yet, the timing of the interruption has a big impact on its disruptiveness [2, 7]. Several studies have demonstrated the negative effects that interruptions can have when they happen at inopportune moments, e.g. when a person is highly focused, ranging from a higher error rate and a lower overall performance to more stress and frustration [8, 16, 51]. To optimize the timing of interruptions and reduce the disruptiveness and negative effects, researchers have looked into measuring a person's interruptibility—the availability for interruptions. Such an interruptibility measure could then be used to postpone computer-based interruptions to more opportune moments [37], or to provide awareness to co-workers and prevent in-person interruptions at inopportune moments [87].

Prior research on measuring interruptibility can roughly be categorized by the kinds of sensors examined: computer interaction or biometric (*aka.* psycho-physiological) sensors. Studies investigating computer interaction use features such as keyboard/mouse input or application usage to find suitable moments for interruptions [20, 37]. Studies on biometric sensors are based on the assumption that physiological features, such as heart rate, pupil dilation or brain activity, can be linked to the user's cognitive states and task engagement and thus be used to determine interruptibility [14, 6, 88]. While study results have demonstrated the potential of features from both sensor types to determine a person's interruptibility, the studies were predominantly conducted on small and controlled tasks over short periods of time (less than three hours) and mostly limited to either computer interaction or biometric sensors.

In the presented research, we build upon and extend previous work by investigating the use of computer interaction and biometric sensors to determine a person's interruptibility at office work-places over a two-week period. Especially since computer interaction sensors are limited to a specific kind of interaction and work during the day and biometric sensors are more physically invasive and more sensitive to noise (e.g. movement artifacts), we are interested in examining the accuracy and feasibility of features of either one or a combination of both sensor types in the field and over a longer period of time. We conducted a two-week field study with 13 professional software developers from three companies, enabling us to study a homogeneous group with similar work patterns, including a variety of activities of which many are performed on the computer [74, 5, 78, 24]. We collected biometric data from several sensors including heart rate, physical activity and sleep measurements, as well as computer interaction data including mouse and keyboard interaction, the active application window, and time and calendar information. In addition, we collected interruptibility ratings through experience sampling using a pop-up displayed on the computer, that we then used as ground truth for predicting a participant's interruptibility.

With the study at hand, we aim to build a classifier that predicts a software developer's interruptibility accurately in the field. Therefore, we first examine the optimal *time window* to extract features from the continuous biometric and computer interaction data. Second, we examine the best combination of *sensors and features* using machine learning techniques and

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ACM 978-1-4503-5620-6/18/04...$15.00
http://dx.doi.org/10.1145/3173574.3174165

how these quantitative results align with the participants' *subjective perceptions* based on qualitative survey and interview data. Finally, we examine whether it is possible to create a *general classifier* rather than one per individual for predicting interruptibility with high accuracy for new people.

In our analysis we found that: (a) the optimal time windows vary per feature (e.g., 10-20min for user input and 2-3min for heart-related data); (b) computer interaction sensors had more predictive power than biometric sensors (74.8% accuracy compared to 68.3% on average), while a combination of both was most accurate (75.7%); (c) participants' perceptions overlap with quantitatively identified feature importance; and that (d) a general classifier can achieve a high accuracy (69.8%), yet a classifier trained for a single individual can outperform the general one even with few data points. Our main contributions are an analysis of predicting software developers' interruptibility in the field, and a comparison of the predictive power of various biometric and computer interaction features.

RELATED WORK

Related work in the area primarily focuses on studies on interruptions, in particular their effects and factors influencing their disruptiveness, and on approaches to measure interruptibility.

Interruptions at the Workplace

Several observational studies showed that a typical work day of knowledge workers is highly fragmented. On average, they switch activities every 2-3 minutes and get interrupted 13 times a day, e.g. through personal visits, emails or phone calls [24]. Solingen et al. found that people spend 15-20 minutes per interruption and a total amount of 15-20% of their time handling interruptions [77]. Sykes reported that the longest interruptions are personal visits from colleagues (ranging from 24 minutes up to 4 hours) [75].

Many interruptions are necessary in a collaborative work space, and often a short interruption can help a co-worker to solve a problem quickly and make progress on a task [77]. However, interruptions can also have multiple negative effects, such as long resumption lags and an increase in errors and frustration (e.g. [8, 16]). Often, knowledge workers do not even go back to their suspended task directly after an interruption [50], or compensate for interruptions by working faster which leads to more stress and frustration [51].

Not all interruptions are equally disruptive. Studies found the interruption moment, duration and frequency as well as the difficulty of the interrupting task and its relevance to current work to be important factors in the disruptiveness of interruptions [6, 16, 59, 12, 23]. Borst et al. developed a disruptiveness model of interruptions and found that the memory required for the interrupted and interrupting task is an important factor, explaining why interruptions are less costly at breakpoints and times of low mental work load compared to moments in the middle of tasks and during high mental workload [10]. With our research we contribute an analysis of automatic and continuous measures of interruptibility in the field that can be used to find opportune moments for interruptions and reduce their disruptiveness.

Finding Opportune Moments for Interruptions

There are primarily two ways to optimize the moment of interruptions: deferring interruptions to task boundaries or continuously measuring interruptibility even during tasks. Since working memory is usually low at task boundaries, the defer-to-boundary policy aims at determining these natural breakpoints during work and delaying interruptions, such as email notifications, to these more opportune moments [35, 7]. Another type of approaches aims at predicting interruptibility continuously [20, 88]. These approaches are particularly useful to reduce in-person interruptions at inopportune moments by indicating the current interruptibility state to potential interrupters [87], but can also be used to postpone computer-based interruptions from moments of low to high interruptibility.

Approaches to continuously measure interruptibility can broadly be categorized by the types of sensors used: biometric, computer interaction, or context sensors. Biometric sensors can be used to measure the body's activities and responses to external stimuli. Various studies have shown that biometric data such as heart rate (HR), heart rate variability (HRV), electro-dermal activity (EDA), eye tracking, skin temperature or electroencephalography (EEG) can be used to assess mental effort and cognitive load [85, 66, 28, 13], task difficulty [79, 22], emotions [48, 27, 62], or stress [29, 70, 84]. A few researchers have also investigated whether such measurements can be used to measure interruptibility. Mathan et al. used an EEG device to compute interruptibility during military training [54]. Goyal and Fussell used EDA data to find opportune moments for interruptions in a lab study [25]. Bailey and Iqbal as well as Katidioti et al. used measures of pupil dilation to find suitable moments for interruptions in lab studies [6, 44]. In a short lab and field study with software developers, a combination of HR, HRV, EDA, and EEG sensors has been used to predict interruptibility [88]. Furthermore, accelerometer data has been used in several studies to detect physical activity and to show that interruptions are better delivered during moments recognized as activity transitions, e.g. when walking to another location [30, 18, 45]. A further and not yet fully studied factor of interruptibility is sleep, which has been shown to have a big impact on productivity and mood [68, 80, 49].

Computer interaction sensors measure a user's interaction with task artifacts on the computer. They mainly consist of mouse, keyboard, and application usage data. Some studies went a step further to get more context from other sources such as audio and video recordings, calendar or network connection data. As an example, Fogarty et al. collected a total of 475 interruptibility ratings and IDE interaction data from 20 participants to measure interruptibility during software development tasks [20]. Other researchers identified breakpoints using computer interaction sensor features such as the frequency of window switches in studies ranging from a few hours [76, 37] to 2 weeks [63]. Kapoor and Horvitz developed BusyBody, an approach that calculates interruptibility using a rich set of computer interaction and contextual features from user input, calendar, time and wireless signal data [32, 41, 42]. Horvitz et al. built a query-able service to predict a user's presence and availability from user activity and proximity from multiple devices, calendar and time information [33]. Another body of research focused on indicating interruptibility or availability in messaging clients or physical indicator lights, and used computer or device interaction, location, speech, calendar, time, presence or network data, or a combination thereof as underlying sensing technique [87, 9, 46, 21].

In our study, we extend upon prior work by using a combination of biometric and computer interaction sensors in the field. We used two biometric sensors (a Fitbit Charge 2 and a Polar H7) to measure HR, HRV, physical activity and sleep and to capture a wide range of biometric data with little invasiveness, compared to e.g. EEG and eye tracking devices, which are more difficult to use in the field. For computer interaction, we recorded the user input (keystrokes and mouse interactions), application usage, and calendar data. To our knowledge this is the first study using this combination of sensors to investigate the continuous measurement of interruptibility in the field and for a longer period of time, in particular its accuracy, feasibility and the predictive power of various types of data.

STUDY DESIGN

To study the prediction of interruptibility in the field, we conducted a two-week field study with 13 professional software developers. For this study, we gathered a rich set of data, including a variety of biometric and computer interaction data as well as interruptibility ratings and qualitative data.

Participants. We recruited 14 software developers through professional and personal contacts from one large-sized and two medium sized companies in the software industry. We focused on software developers as one community of knowledge workers, to ensure our participants have similar work patterns including a wide variety of activities and extensive computer use to support both individual and collaborative tasks [74, 5, 78, 24]. We discarded the data of one participant due to a technical issue with the Polar sensor that led to no recordings from this sensor and thus an incomplete and incomparable dataset for this participant. Of the remaining 13 participants, 1 was female and 12 were male. At the time of the study, participants had an average age of 32.4 years (standard deviation, in the following denoted with ±, of 6.2), an average professional experience of 6.5 years (± 6.2) and total experience in software development of 11.8 years (± 6.6). Most participants were individual contributors (6), and the others had job roles such as architects (3), executives (1), lead (2) and other (1).

Procedure. At the beginning of the study, we explained the purpose and process of the study, and handed out, set up and introduced the two biometric sensors (Fitbit Charge 2 and Polar H7). We asked the participants to wear the Polar sensor during work hours, and the Fitbit sensor as much as possible including work and free time as well as nights, except when they did not feel comfortable wearing it or when swimming, showering or charging the device. The participants synced the data every one or two days. In addition, we installed a **monitoring tool** to collect computer interaction data. In case a participant worked on several computers, we installed the monitoring tool on all of them to collect a complete data set. We further automated the synchronization of the time for all devices (computers and sensors) participants used for the study period.

For the following two weeks (some participants also continued the study for a few more days), we asked participants to follow the same procedure every work day. We asked them to wear the biometric sensors, to rate their interruptibility when prompted by a pop-up on their computer with an experience sampling technique, and to fill out a short daily diary survey regarding their perception of the work day in the evening.

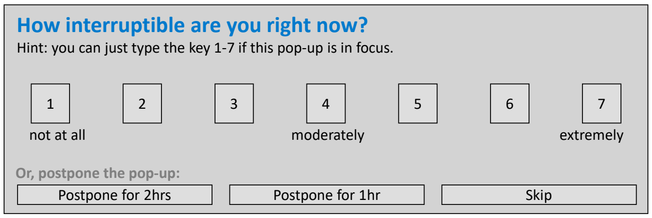
At the end of the study, we collected the sensors and data, conducted interviews on our participants' perception on interruptibility, and asked them to fill out our end survey with demographic questions. In the remainder of this section, each part of the study procedure is explained in detail.

Biometric Sensors. Based on prior research as well as invasiveness, we chose to use two biometric sensors for our field study: the Polar H7 for recording HR and HRV data, which both have been linked to stress and cognitive load by previous research [1, 28], and the Fitbit Charge 2 for recording HR (sampled every 10s), physical activity (sampled every 1min), and sleep (duration and quality metrics), which have been linked to interruptibility [30, 18, 45] and productivity [68, 80].

The Polar H7 [17] is a chest strap recording heart beats and interbeat-intervals, using an electrocardiograph (ECG) based sensor technique with medical grade accuracy [83]. The Polar's little invasive form-factor and long battery life make it feasible to be used in a field study. Since the sensor has no built in memory, we extended our monitoring tool with the capability to receive the measurements of the device via bluetooth, which limits the data collection with this sensor to the time spent within bluetooth range of the computer.

The Fitbit Charge 2 [19] is one of the most accurate wrist-worn activity trackers [26]. While the Fitbit's coarser sampling granularity does not allow measuring HRV [83] and tends to overestimate sleep duration, it has a high intra-device reliability [60] and can be worn constantly (except for charging, showering and swimming) thanks to the minimally invasive form-factor and the built-in memory. The Fitbit data was synced to Fitbit servers via bluetooth using the official smart phone or computer application, and then automatically downloaded by our monitoring tool. For this purpose, the participants granted our monitoring tool access to the Fitbit account during the study.

Monitoring Tool. To collect computer interaction data, we used our own monitoring tool for the Windows operating system that tracks a participant's mouse and keyboard interactions, the active window, and calendar information. For the mouse, the clicks (coordinates and button), the movement (coordinates and moved distance in pixels), and the scrolling (coordinates and scrolled distance in pixels) are tracked along with the corresponding timestamp. For the keyboard, we recorded the keystroke type (normal, navigating or delete key) along with the corresponding timestamp. We did not record specific keystrokes for privacy reasons. For the active window, we recorded the name of the active process and the window title, along with the timestamp at which the user switched to the window. For calendar data, the tool used the Microsoft Graph API of the Office 365 Suite [57] and recorded start time, duration and subject of meetings.



How interruptible are you right now?

Hint: you can just type the key 1-7 if this pop-up is in focus.

Figure 1: Screenshot of the interruptibility rating pop-up

Interruptibility Ratings. To collect the ground truth for the interruptibility classification, we prompted our participants with an experience sampling technique using a pop-up that was displayed on the computer. The prompts asked participants to rate their current interruptibility on a 7-point Likert scale and were displayed in random intervals between 10 and 40 minutes. We chose this time interval as a trade-off between annoyance and invasiveness while also collecting enough samples to apply machine learning. This decision was based on our experience from a pilot study with 8 software developers during 7 work days and from testing the final study procedure ourselves for several days. In the pilot study, we further observed that some participants tended to avoid the extreme or intermediate parts of the scale. Therefore we extended the original 5-point Likert scale (which has predominantly been used in related work [20, 76]) to a 7-point Likert scale to obtain a higher variety of ratings. The pop-up prompts were displayed in the bottom right corner of the main screen and were directly integrated into the monitoring tool (see Figure 1). With just one click, the prompt could be answered, skipped or postponed for the next one or two hours, preventing false answers caused by annoyance. For participants that used multiple computers simultaneously, we disabled, if desired, the prompts on all except the main computer to prevent fatigue from too many and frequent prompts. Our participants had the possibility to correct a rating by sending an email which occurred twice throughout the course of the study.

Surveys and Interviews. We collected qualitative data to gain insights on participants' perceptions of interruptibility and related factors, complementary to the quantitative data. At the end of each work day, the participants answered the same short **diary survey** containing items regarding their work day. The items in the survey were all rated on a 7-point Likert scale and included productivity, sleepiness, challenge, engagement, arousal, valence, stress, interruption frequency, and daily interruptibility. We included these items to analyze their relation to interruptibility and chose them based on literature and their potential impact on interruptibility (e.g. [68, 6, 62, 51]). At the end of the study period, we further conducted **interviews** to ask openly about factors that influence participants' interruptibility, and about their experience with the biometric sensors and the monitoring tool. The study concluded with an **end survey** to collect demographic data.

DATA COLLECTION AND PREPROCESSING

In our two-week field study, we collected a rich set of quantitative and qualitative data (see details in Table 1). Prior to the main analysis of the data, we performed multiple preprocessing steps that are summarized in the remainder of this section.

	Total	**per Participant**
Polar data	808 hours	62 hours (± 12)
Fitbit data	5532 hours	426 hours (± 76)
Fitbit sleep data	197 nights	15 nights (± 4)
Computer monitoring data	3552 hours	273 hours (± 143)
Calendar entries	746 meetings	57 meetings (± 37)
Interruptibility ratings	2515 samples	193 samples (± 88)
Interviews	525 minutes	40 minutes (± 8)
Diary survey	151 responses	12 responses (± 1)

Table 1: Collected data

Our preprocessing and analysis scripts along with more detailed explanations and information are available online[1].

Basic Preprocessing. Before analyzing the computer interaction data, we anonymized the data by replacing identifying text fragments with placeholders. We further merged the computer interaction data for participants that worked on several computers in parallel, mostly by adding all data points into one common database. For two participants that used Remote Desktop Connection to switch between computers, we further had to delete entries representing the Remote Desktop window, and only used the user input from the main machine to prevent duplicates.

Feature Extraction. A first step towards building a reliable interruptibility classifier is to extract meaningful features of the raw data. We extracted features that have previously been linked to cognitive states such as cognitive load, stress or emotions, and also interruptibility. Table 2 provides an overview of all 85 features that we extracted along with the corresponding references where they have been defined or used previously.

From the computer interaction data, we extracted user input features, in particular frequency and duration measures of keystroke and mouse events that capture if a person is actively producing content or being idle, e.g. thinking, reading or away from the computer. We further extracted application window features that capture window switching events and time spent in specific activity categories. We define an application window as a unique combination of the process name and window title. An application window switch can therefore refer to a switch between two different applications as well as, for example, a switch between two different tabs in a web browser. We obtained activity categories from the window switching events by mapping window and process names to a general activity category such as *Coding*, *Reading or Writing Documents*, or *Email or Planning* (for all categories see Table 2). We used common categories typical for software developers that had previously been identified by Meyer et al. [56]. We mapped the data semi-automatically in two stages. First, an automatic algorithm developed by Meyer et al. mapped obvious programs and activities, such as *Microsoft Visual Studio* belonging to the activity category *Coding* [55]. In a second step, we manually mapped the remaining entries using the window titles that provided valuable contextual information, e.g. to distinguish between *Work Related Browsing* and *Work Unrelated Browsing*. We further extracted features related to focus duration and activity / category switching frequency inspired by Sarkar and Parnin who used these features to predict

[1] `http://dx.doi.org/10.5281/zenodo.1118965`

Feature Group	**Importance**	**Features**	***References***
User Input	**29.6%**	*Sensor: Computer Monitoring*	*[36, 72, 3, 31, 21, 87, 71, 41, 42, 32, 33]*
Keystrokes	11.3%	Number of all (2min, 10min) / normal (20min) / navigation (20min) / delete (20min) keystrokes per min, percentage of time spent typing (10min)	
Mouse Clicks	8.2%	Number of all (10min) / left (10min) / middle (45min) / right (30min) / other (20s, 45min) mouse clicks per min, percentage of time spent clicking (10min)	
Mouse Scrolls	3.2%	Scrolled distance per min (30min), percentage of time spent scrolling (30min)	
Mouse Moves	4.2%	Moved distance per min (20min), percentage of time spent mouse moving (10min)	
Keystrokes & Mouse	2.6%	Percentage of time being idle (10min)	
Application Window	**44.6%**	*Sensor: Computer Monitoring*	*[36, 63, 58, 3, 41, 42, 32, 33]*
Activity Category	30.4%	Percentage of time spent in the following activity categories and sub categories: Software development (2min) (coding (3min, 10min), debugging (5min), version control (10min), reviewing (2min)), communicating (3h) (email (1h), instant message (2min, 2h)), reading or editing documents (1min, 20min), web browsing (30s, 20min) (work related browsing (10min), work unrelated browsing (3h)), work unrelated activities (10s) (work unrelated browsing (10s), work unrelated apps (1min, 3h)), planning (10s, 20min, 45min, 3h), navigating and other (45min)	
Focus Duration	5.9%	Max. time in one application window (20min) / category (10s, 20min)	
Activity Switches	8.2%	Number of application window (20min) / category (10s, 5min, 20min) switches per min, number of distinct categories (10s, 20min)	
Calendar	**2.8%**	*Sensor: Computer Monitoring*	*[73, 31, 21, 87, 41, 42, 32, 33]*
Past Meetings	1.8%	Number of past meetings per hour (3h), percentage of time spent in meetings (7.5min, 3h), meeting now (boolean)	
Upcoming Meetings	1.0%	Number of upcoming meetings per hour (1min, 45min), percentage of time planned in meetings (30s, 1h)	
Heart	**14.2%**	*Sensors: Polar and Fitbit*	*[85, 28, 88, 61, 13, 29, 1, 86, 43, 4, 27]*
HR	9.8%	Polar HR mean (20s, 3min) / std. dev. (45s), Fitbit HR mean (20s) / std. dev. (10min), Fitbit resting HR, Fitbit percentage of time spent in HR zones (45min)	
HRV	4.4%	Polar SDNN (3min), Polar RMSSD (3min), Polar pNN50 (2min)	
Movement	**2.3%**	*Sensor: Fitbit*	*[30, 18, 45]*
Steps	2.3%	Number of steps per min (2min), percentage of time spent walking (3min)	
Circadian Rhythm	**6.5%**	*Sensors: Computer Monitoring and Fitbit*	*[81, 52, 65, 68, 80, 49, 41, 42, 32, 33]*
Time	2.1%	Hour of day, day of week, hour arrived at work	
Sleep	4.4%	Duration, sleep efficiency, hour of midpoint of sleep, hour of wakeup, number and minutes being awake / restless	

Table 2: Features analyzed in our study and grouped by sensor together with the feature's importance for the interruptibility classifier, the used time window per feature (colored and in brackets), and references to prior related work on these features.

mental fatigue of software developers [71]. From the calendar entries we extracted features indicating whether the person had scheduled a meeting for the recent past or future. Finally, to capture data related to the circadian rhythm we extracted time related features, e.g. the hour arrived at work based on the first interaction with the computer per day.

From the biometric data we extracted HR and HRV related features from both the Polar and the Fitbit sensors by taking advantage of the higher accuracy of the Polar and the larger amount of data available from the Fitbit. For HRV, we used three standardized metrics: the standard deviation of the successive differences of heart beats (SDNN), the root mean square of the successive differences (RMSSD) and the proportion of pairs of successive intervals that differ by more than 50 ms (pNN50) [86]. To calculate the heart rate zones, we used the Karvonen method, using the mean of the daily resting heart rates measured by the Fitbit Charge 2 throughout the whole study period and the age the participants reported [43]. We use heart rate zones as suggested by the American Heart Association and used by Fitbit: up to 49% of the maximum heart rate is regarded as being out of zone, 50% to 69% is labeled with low activity, 70% to 84% high activity and 85% and more is peak activity [4]. Steps and sleep measurements were extracted as indicated in Table 2.

Outcome Measure. As outcome measure we used the interruptibility ratings collected with experience sampling. Figure 2 shows that prompts were answered throughout the whole work day, though less often early in the morning, at lunch and in the evening and that most prompts were answered quickly (50% within 8s, and 83% within 15 minutes). To predict if a person is interruptible, we reduced the 7-point Likert scale to two states (splitting at 123 | 4567), similarly to previous studies predicting interruptibility based on experience sampling ratings, which split a 5-point Likert scale between 2 and 3, counting the middle rating to the interruptible samples [20, 88]. For our more fine-grained analysis, we used the full 7-point Likert scale and additionally split it into three states (splitting at 12 | 345 | 67). As one participant never used a rating of 1 or 2 and thus had a highly imbalanced dataset using this splitting method (for two states: 91.4% being interruptible - 8.6% being non-interruptible), we accommodated for the imbalance by using a different splitting mechanism (1234 | 567 and 1234 | 5 | 67) for this participant.

Machine Learning Tuning. We used scikit-learn [64], a widely used machine learning library for Python, to predict

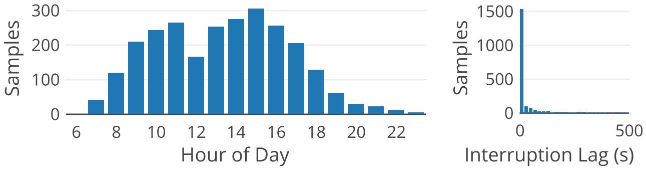


Figure 2: Distribution of self-reports and interruption lags (truncated after 500s for better readability).

interruptibility from biometric and computer interaction data. We evaluated several classifiers by applying them to our feature set and testing different parameter values. A random forest classifier (500 estimators, no prior feature selection) outperformed all other approaches, including a gradient boosting classifier (500 estimators, max. depth=3, no prior feature selection), support vector machine (kernel=RBF, C=1, gamma=0.03, selected 30 best features prior to classification), neural network (solver=LBFGS, alpha=0.0001, hidden layers=100, no prior feature selection) and Naïve Bayes classifier (selected 30 best features prior to classification) [69]. Therefore, for the remainder of this paper, we will present results obtained with a random forest classifier. A random forest classifier is an ensemble learning method that creates a multitude of decision tree classifiers and aggregates their predictions with a voting mechanism [11, 47]. It is noteworthy that this classifier does not require preselecting features, and can deal with a large feature space that also contains correlated features. In all our machine learning experiments, we first imputed missing values by replacing them with the mean, and normalized the features to comparable scales using a *StandardScaler*. These are common initial steps in a machine learning pipeline and a requirement for many classifiers to work properly [64].

ANALYSIS AND RESULTS

To examine whether we can use the collected sensor data to accurately predict interruptibility in the field and which combination of computer interaction and biometric features achieves the highest accuracy, we applied machine learning to our preprocessed features using the self-reports as the outcome measure. In the following, we first examine which time windows to use for each extracted feature, followed by an analysis and findings of the best features and combinations thereof. To complement the quantitative results, we further analyze how participants' perceptions of their interruptibility overlap with our findings. Finally, we investigate how well a general classifier of interruptibility can be used across participants in the field compared to an individually trained classifier and examine whether the features can also be used to predict interruptibility on a more fine-grained level.

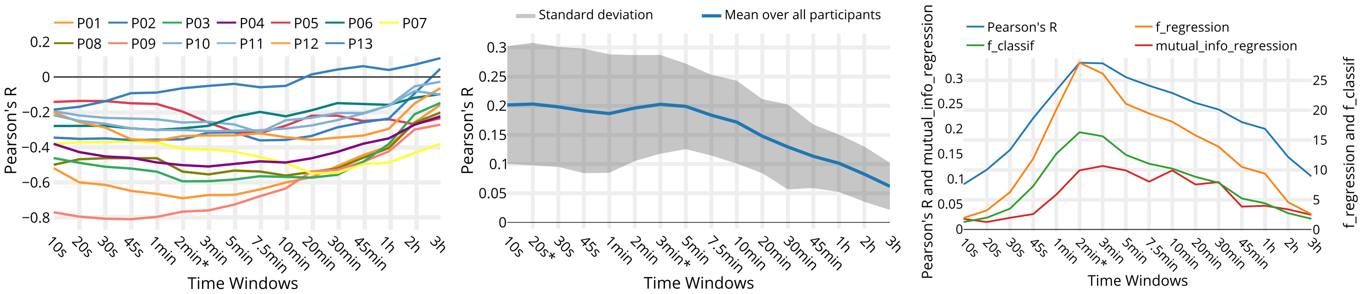
Time Windows

As a first step in determining a classifier for interruptibility, we have to decide on the time windows that are being used for each of the extracted features so that we can transform the continuous data streams of each feature into discrete variables. The time window can have a significant impact on the classifier as previous research has shown [82, 88]. While previous researchers have used a variety of time windows for predicting interruptibility, predominantly between 1s and 5mins [34, 20, 88], there is no general guideline on which time windows to use for which feature. In our analysis, we take advantage of the longitudinal nature of our study and the variety of features examined and analyze which time windows are optimal to predict interruptibility. In particular, we analyze an extensive set of time windows ranging from 10 seconds all the way to 3 hours: *10s, 20s, 30s, 45s, 1min, 2min, 3min, 5min, 7.5min, 10min, 20min, 30min, 45min, 1h, 2h, 3h*. We put a focus on shorter time windows due to their use in prior studies, but we also include longer time windows that have not been examined in earlier studies, especially due to the short and controlled nature of the tasks used in these studies.

To determine the optimal time window(s) for predicting interruptibility, we used three commonly used univariate statistics (Pearson's R, ANOVA's F and Mutual Information (MI)) to calculate a set of four metrics for each combination of time window and feature. We chose these three statistics to capture a broad range of possible dependencies between the outcome measure and the feature values with Pearson's R and ANOVA's F capturing linear relationships and MI to capture non-linear dependencies. Based on the three statistics, we calculated a total of four metrics since we calculated the F-score for both classifications (two states of interruptibility) using `f_classif` and regression (7 states of interruptibility) using `f_regression`. As we did not have enough samples to compute MI for classification reliably, we only computed it for regression using `mutual_info_regression`. We used scipy.stats [40] to calculate Pearson's R and scikit-learn [64] for the other metrics.

We visually inspected the graphs that we generated for each feature, metric and each participant (see an example in Figure 3 (a)) and found that the line graphs from different participants have similar trends and slopes (see Figure 3 (a)). We therefore aggregated the data from all participants by calculating the mean and standard deviations of each metric's absolute values and generated a graph for each feature (see an example in Figure 3 (b)). Finally, we compared the four different metrics with each other by generating graphs for each feature including all metrics (see an example in Figure 3 (c)). We found that all four metrics were highly correlated, even the mutual information metric (Pearson's R and `f_classif`: Pearson r=.92, p<.000001, Pearson's R and `f_regression`: Pearson r=.95, p=.0, Pearson's R and `mutual_info_regression`: Pearson r=.84, p<.000001) and that they have similar peaks (see Figure 3 (c)). We ended up choosing the time windows that maximized the absolute mean of Pearson's R over all participants through manual visual peak detection. When there were several peaks or in the rare cases where the metrics had substantially different peaks (e.g. due to a non-linear dependency), we added each peak as a time window. The latter occurred for 15 of the 55 features for which we determined a time window.

The selected time windows per feature are listed in (blue) in Table 2. For the biometric features (heart and movement), shorter time windows between 10s and 3min were generally better than longer ones, whereas for user input and application window features, longer windows between 10min and 20min



(a) Pearson correlation between interruptibility ratings and the feature *percentage of time spent in software development* over all time windows and per participant.

(b) Overall Pearson correlation between interruptibility ratings and the feature *Polar mean of HR* extracted over all time windows.

(c) All four metrics used to compare time windows for predicting interruptibility using the feature *number of steps per min* averaged over all participants.

Figure 3: Selection of graphs generated to determine the optimal time window for predicting interruptibility (chosen time window denoted with *).

were better. Exceptions were communication (3h) and software development (2min). Our results show that the optimal time window varies per feature and suggest a range of time windows which work well for certain feature groups.

Sensors, Features and Perceptions

To evaluate the accuracy of predicting interruptibility in the field and compare the predictive power of the various features, we applied machine learning to the collected features as well as groups of features. To add to the understanding of participants' perception on interruptibility and in particular how and why specific features might relate to their interruptibility, we further complement the quantitative findings with an analysis of our diary survey and interviews.

Interruptibility Prediction

The goal of our research is to predict a person's interruptibility in a specific moment with high accuracy using the features extracted from the collected biometric and computer interaction data. We use the ratings from the participants' experience samples split into two states as ground truth. Since biometric and computer interaction data is highly individual and trained models can often not easily be transferred to new participants [22, 88, 81], we trained models individually for each participant, similarly to Haapalainen et al. [28]. For each participant, we predicted interruptibility using ten trials of stratified ten-fold cross-validation, which keeps the class proportions consistent in each fold, and a random forest classifier pipeline (500 estimators) with initial feature imputation and standard scaling.

Table 3 presents the accuracy scores for each sensor and combinations thereof. As baseline accuracy we report the accuracy that a majority classifier would achieve that always predicts the class containing more samples. The results are obtained training individual models for two states of interruptibility. While all sensors were better than the baseline, the features of the computer interaction sensors (accuracy=74.8%) were more predictive compared to the features from the biometric sensors (accuracy=68.3%). Adding one or both biometric sensors slightly improves the classifier (accuracy=75.7%). When comparing the Polar and the Fitbit sensors, for 9 of 13 participants the Fitbit yielded better results, while for the remaining 4 the Polar was more accurate (accuracy Fitbit = 66.2%, accuracy Polar = 62.5%). Note that the Fitbit comprises a wider variety of features, e.g. step count, than the Polar.

Table 2 contains the feature importance attributed by the random forest classifier using all features and averaged over all participants' individual models. For the feature importance metric we used the Gini impurity measure from scikit-learn [64] that is attributed to each feature by the random forest classifier and captures the feature's ability to avoid misclassification [67]. The most important features are the application window group and user input, followed by heart and sleep measurements. Calendar (2.8%), movement (2.3%) and time related features (2.1%) are the least important, contributing only 7.2%.

Developers' Perceptions of Interruptibility

To complement our quantitative comparison of features and sensors, we analyzed the interview and diary survey data to learn more about how software developers perceive interruptibility and related factors, and whether their perception matches our feature model. We analyzed the interview audio recordings by transcribing and applying open and axial coding and the diary survey data using multiple regression analysis.

Similar to our classification results that show that application window and user input features are most predictive, all participants stated in the interview that their interruptibility changes with certain activities on the computer, such as coding or writing emails, but only a few also explicitly mentioned the user input (15% of participants).

> *When I do development or code reviews I am very focused and not interruptible. During email writing on the other hand, I am more interruptible.* (P04)

> *If I am typing something, sure I might forget what I was typing when I get interrupted.* (P12)

In addition and consistent with prior work (e.g. [6]), participants stated that they are more interruptible at task boundaries (69%). While this is not explicitly covered in our examined features, this is somewhat implicitly captured by user input and application window features while participants are working at the computer as previous research has shown [76, 63,

	Interruptibility Prediction Accuracy (2 States, Individual Models)													
	P01	**P02**	**P03**	**P04**	**P05**	**P06**	**P07**	**P08**	**P09**	**P10**	**P11**	**P12**	**P13**	**All**
Baseline Accuracy	66%	63%	53%	58%	58%	53%	71%	82%	53%	61%	52%	56%	57%	**60.2%**
Fitbit	64%	72%	72%	64%	66%	57%	65%	79%	61%	74%	63%	65%	59%	**66.2%**
Polar	66%	67%	56%	59%	58%	55%	69%	77%	73%	62%	59%	54%	59%	**62.5%**
Computer Monitoring	78%	69%	80%	73%	70%	74%	74%	85%	85%	76%	74%	72%	62%	**74.8%**
Fitbit + Polar	68%	76%	70%	65%	61%	58%	69%	81%	72%	76%	67%	63%	61%	**68.3%**
Fitbit + Computer Monitoring	79%	73%	80%	75%	70%	74%	74%	86%	85%	78%	74%	72%	64%	**75.7%**
Polar + Computer Monitoring	78%	72%	80%	74%	69%	72%	73%	85%	86%	77%	73%	73%	62%	**75.0%**
Fitbit + Polar + Computer Monitoring	79%	76%	79%	74%	69%	72%	74%	85%	86%	78%	74%	72%	62%	**75.3%**

Table 3: Prediction results using different sensors and combinations thereof per participant and averaged over all (the darker the color the higher the accuracy).

		Predicted Label						
		1	2	3	4	5	6	7
True Label	1	15%	2%	1%	1%	0%	0%	1%
	2	3%	5%	4%	1%	1%	0%	1%
	3	1%	3%	6%	2%	1%	1%	1%
	4	3%	2%	4%	5%	1%	1%	1%
	5	1%	1%	2%	2%	2%	2%	1%
	6	1%	0%	1%	1%	1%	3%	2%
	7	1%	0%	1%	1%	0%	2%	10%

Table 4: Aggregated confusion matrix for seven states from individual models of all participants.

Obs.: 151, Adj. R^2=.28, R^2=.32, F(8, 142)=8.34, p<.0.00000001

int. frequency*	(β=.14, p=.025)	engagement*	(β=-.41, p=.000)
productivity	(β=-.12, p=.23)	challenge	(β=-.08, p=.44)
stress	(β=-.16, p=.07)	sleepiness	(β=.11, p=.12)
valence	(β=.11, p=.21)	arousal	(β=-.06, p=.45)

Table 5: Linear regression results with daily interruptibility as dependent and feature ratings collected in the daily survey as independent variables (* denotes significance at p<.05).

37], as well as with the features related to being idle, calendar entries and physical movement, e.g. when changing location and coming back from a meeting [30, 18, 45, 73].

> *... [I am more interruptible] between tasks, when I organize myself and plan my next step.* (P03)

> *... [more interruptible] around meetings, because it takes me a bit of time to get back in the flow of things.* (P09)

Participants further mentioned that their interruptibility depends on internal states such as sleepiness (85%), focus (77%), mood (46%), challenge (38%), productivity (38%), stress (38%), health (23%), and engagement (15%).

> *When [last night] was relatively short, I have a hard time to concentrate anyways, and want to be disturbed less.* (P03)

> *When I am kind of frustrated or nervous, I am more annoyed if someone interrupts me.* (P13)

> *When I was doing a complicated code review, where I first had to understand the dependencies, it would not be good to be interrupted.* (P05)

This overlaps with the sensors we chose, especially the biometric ones, as they have the potential to measure a variety of internal states such as stress, mood or mental load [29, 27, 62, 28].

When asked about temporal patterns of interruptibility over the course of the day, many participants stated that they do not necessarily think that there is a direct link to interruptibility, but rather that the routine of activities and external factors such as background noise and interruption frequency is linked and might vary throughout the day.

> *There is nothing specific about the time of day, it is just how my routine is laid out.* (P06)

> *Around lunch time is the busiest time of the day.* (P02)

Most participants find it easier to focus, which would result in lower interruptibility, when the office is quieter (46%).

> *After 5pm many go home and then it's very quiet, then it is easier to concentrate.* (P02)

To further examine temporal patterns of interruptibility, we visually analyzed the interruptibility ratings in relation with the time of day. Similar to the interview responses, we could not find any consistent and significant patterns across participants, which is also supported by the fact that time related features were only weighted by 2.1% in the interruptibility classifier.

In our daily diary survey that we performed throughout the study period, we asked participants to rate their relative overall interruptibility for the whole day. We further asked them to rate several features (listed in Table 5) that were referenced in prior work in relation to interruptibility and work focus [68, 6, 62, 51, 52] and that are to some extent captured by our sensors, especially the biometric ones. We found that the interruptibility ratings per day collected with the experience sampling prompts and the daily interruptibility rating from the diary survey correlate significantly (Pearson r=0.42, p<.000001), which provides support for the validity of the measures. We then performed a multiple linear regression analysis with the daily interruptibility rating as the dependent variable using all 151 recorded responses from all participants. The results (shown in Table 5) show that participants were more interruptible when they had many interruptions, and less when they were engaged, and that there is a trend (not significant though) that participants were more interruptible when they were sleepy, and less when they were stressed or productive.

Overall, our results indicate that there is a strong overlap between the features determined as particularly predictive in our analysis of the sensor data and the perceptions of participants.

Interruptibility Prediction in the Field

To investigate the general use and sensitivity of our interruptibility classification in the field, we first create and compare a general model trained across several participants with our individually trained models, and second, examine the classification of a more fine-grained interruptibility.

The main advantage of a general model is that no initial training phase is needed to use it on a new subject in practice. For our analysis, we used leave-one-out cross-validation for which we iterated over all participants and trained a classifier with data from all participants except one and tested it

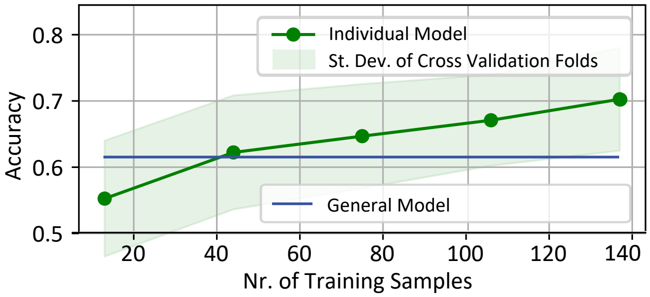


Figure 4: Learning curve for participant P06.

on the remaining one [81]. The results show that the general model achieves equal or better accuracy than the baseline for all except the two participants P07 and P08 (see last column of Table 6). At the same time, and not surprising, the individual models performed better for almost all participants, except for three (P03, P05, and P12), with a 75.3% averaged accuracy over all participants compared to 69.8% for the general model.

To investigate how many training samples per individual are approximately needed to build an individual interruptibility classifier that is as good or better than the general model, we produced learning curves for each individual using shuffle split cross-validation (100 splits, test size of 20% of the available samples). Figure 4 depicts an example of a learning curve for one participant (P06). The illustrated example shows that already with few samples (approximately 40 in this case), the individual classifier starts outperforming the general model and improves with increasing sample size. Over all participants, an average sample size of 20 to 80 was sufficient to train an individual interruptibility classifier that is close or outperforms the general one.

In a real-world application, a more fine-grained classification of interruptibility might also be valuable, e.g. when indicated to co-workers, it might enable a more informed decision whether to interrupt someone or not, also with respect to the priority and kind of the interruption. We therefore examined the accuracy of a more fine-grained classification by splitting the outcome measure—the interruptibility rating—into three and seven states. For this analysis, we used the best feature set, i.e. all features, that we determined earlier. Table 6 presents the results for interruptibility predictions into several granularities for each participant. Average prediction accuracies were 75.3%, 65.5% and 42.5% for prediction into two, three and seven states of interruptibility, which is an average improvement of 26.6%, 25.7% and 36.9% compared to a majority classifier. The aggregated confusion matrices for prediction into three and seven states reveal that mis-classifications rarely fall into distant classes, but often into adjacent ones (see Table 4 for seven states). These results indicate that a classifier trained on the collected computer interaction and biometric features is able to predict interruptibility with reasonable accuracy not only into two, but also three and even seven states of interruptibility in the field.

DISCUSSION

In the following, we discuss our findings, in particular implications from the time window analysis and feature comparison, practical applications of the interruptibility classifier as well as limitations and threats to validity.

Time Windows and Features. Our results suggest that a developer's interruptibility is not only affected by the few seconds and minutes before an interruption, but that there are features, such as the activities or sleep, that can have a longer lasting effect on interruptibility. While most prior work focuses on features calculated for short time windows of up to 5min [34, 82, 88], we analyzed a wider range of features and time windows spanning from 10s to 3h for most features and a whole day for some, such as sleep and resting HR. Our results show that for certain feature groups, longer time windows are more informative and that even daily features have an importance for predicting interruptibility, e.g. 4.4% importance for sleep (see Table 2). For example, communication related activities were most correlated to interruptibility using large time windows of 1 to 3h. This longer lasting effect of communication related activities was also mentioned in a previous study that found that office workers feel less productive after spending a longer amount of time with email activity [53], which in turn might impact their interruptibility. We also found that there were several 'good' time windows for certain features. A possible explanation is that these time windows refer to different notions of the feature. An example is the *max. time in an activity category*. For a short time window (10s) it might indicate whether the person is at a breakpoint or task switch, while for longer time windows (20min) it is more indicative of extended focus. In general, our findings show that there are certain ranges of time windows for certain feature categories, but that there is not necessarily just one best time window for each feature. Future studies should therefore further analyze how the feature under investigation varies over time.

Sensor Comparison. Previous research has already linked both, computer interaction and biometric sensors to mental load and interruptibility [28, 45, 80, 37, 63, 20]. To the best of our knowledge, our study is the first to compare these types of sensors in the field over a longer period of time. While our study demonstrates that computer interaction features can be used to accurately predict interruptibility at the computer and that they are more predictive than the biometric features used in our study, the results also show that biometric sensors already have a great potential in accurately predicting interruptibility in the field despite the noise. For our study, we focused on two biometric sensors that we selected due to their little invasiveness, cost and availability. Especially with the rapid advances in technology in combination with biometric sensors being less limited to a specific workstation and being able to capture more of a person's work day, our results demonstrate the potential of these types of sensors for the future. Overall, participants perceived the computer interaction sensors as less invasive, but thought that the captured data was more sensitive than the biometric data in the work context. Biometric sensors can thus serve as a complement or substitute to improve accuracy, or respect privacy preferences for now.

Practical Use. Our findings show that using a general interruptibility classifier is accurate enough to successfully break

	Valid Samples	Skipped Samples	Histogram	Individual Models									General Models	
				2 States			3 States			7 States			2 States	
				Base	Acc.	Impr.	Base	Acc.	Impr.	Base	Acc.	Impr.	Acc.	Impr.
P01	217	3		66%	79%	19%	46%	64%	39%	31%	41%	31%	67%	1%
P02	142	2		63%	75%	18%	59%	69%	16%	33%	40%	21%	66%	4%
P03	195	4		53%	80%	50%	53%	67%	27%	23%	39%	71%	82%	54%
P04	200	8		58%	75%	30%	52%	60%	17%	24%	36%	55%	74%	28%
P05	127	1		58%	69%	18%	43%	48%	11%	20%	27%	30%	71%	22%
P06	172	16		53%	73%	36%	40%	60%	51%	28%	38%	38%	64%	20%
P07	135	0		71%	73%	3%	49%	70%	43%	44%	51%	17%	70%	-1%
P08	152	0		82%	85%	4%	65%	73%	12%	33%	48%	46%	67%	-18%
P09	191	0		53%	86%	62%	43%	75%	75%	39%	68%	73%	76%	45%
P10	484	4		61%	78%	28%	56%	77%	38%	51%	69%	36%	64%	5%
P11	162	0		52%	73%	40%	62%	68%	9%	26%	39%	51%	73%	39%
P12	145	29		56%	71%	28%	63%	62%	-2%	29%	32%	10%	73%	31%
P13	193	17		57%	63%	10%	60%	59%	-2%	25%	25%	0%	60%	5%
Totals:	**2515**	**84**		**60.3%**	**75.3%**	**26.6%**	**53.1%**	**65.5%**	**25.7%**	**31.2%**	**42.5%**	**36.9%**	**69.8%**	**18.0%**

Table 6: Results for predicting 2, 3 and 7 states of interruptibility along with the size and distribution of the available samples. The last column reports results from general models trained on all but one and tested on the one participant. *Legend: "Base": Baseline accuracy obtained by a majority classifier, "Acc.": Accuracy, "Impr.": Percentage improvement over majority classifier*

the cold start problem. For practical use, we therefore suggest using a general model as a default and allowing the user to improve the classifier by training it. Even with few individual samples one is able to achieve a high accuracy with this approach. In general, such a classifier can then be used to indicate a knowledge worker's interruptibility to potential co-workers, which has been explored with physical indicators [87], or indicators displayed on the computer [9, 46, 21]). Similarly, such an interruptibility classifier can be used to mediate interruptions directly by postponing computer-based interruptions while a person is non-interruptible to a more opportune moment, which has also been investigated in prior work [37, 32, 41, 42, 33]. A further potential use of the data is to display the current and historical interruptibility state to the knowledge worker herself. Given the strong links between interruptibility and states such as focus or stress, increased awareness about one's interruptibility patterns might help knowledge workers to reflect on their work patterns and potentially improve their work experience. Several of our participants already enjoyed the biometric data by itself a lot.

Limitations. We conducted our study with software developers working in offices, which limits our results to this context. While we can assume that the results can be generalized to similar job roles and environments, more research needs to be conducted to study interruptibility in other areas. We further prompted our participants to rate their interruptibility using a pop-up displayed on the computer, which limits the times of responses to times spent at the computer. Therefore, we were not able to collect self-reports during times spent away from the computer. However, some of our features (e.g. heart and movement data) were collected at all times, even when the participant was away from the computer, and we have several data points from prompts that were answered shortly after returning to the computer. In fact in 17% of our data samples participants answered the prompt less than 1 minute after an idle period without computer interaction.

Threats to Validity. The interruptibility rating pop-up is, ironically, an interruption in itself and could have potentially disrupted our participants in their work flow. As participants usually only needed a very short time to answer the prompts (in 53% of all cases the pop-up was answered within 10s) and as they only rarely postponed a prompt (3% of all prompts), we are confident that the pop-ups did not disrupt our participants from their work flow notably. Another threat to validity is that participants might not be able to assess their interruptibility correctly or that they might not have understood the question. We ensured that we spent enough time to explain the pop-ups at the beginning of the study to mitigate this risk. Furthermore, not every one of the collected samples in our dataset contains full data from all sensors, which might influence the comparison of the sensors and features, e.g., computer interaction data is inherently limited to times spent at the computer. Missing values were imputed by replacing with the mean before classification, as this technique can lead to better results than discarding them which would decrease the sample size.

CONCLUSION AND FUTURE WORK

In this paper, we presented the results of a two-week field study with 13 professional software developers in which we examined the use of a wide variety of biometric and computer interaction features to predict interruptibility. Our analysis shows that we are able to predict interruptibility at the computer with 75.3% accuracy (a 26.6% improvement over the baseline) and that computer interaction features are more accurate than the biometric ones (74.8% vs. 68.3%). We further show that the best time windows to extract features vary across feature categories and that certain features can affect interruptibility over long periods of time. Finally, we show that even a generally trained model can accurately predict interruptibility for new subjects to overcome the cold start problem, and that even small sets of samples can be used to rapidly improve the classifier.

As a next step, we plan to generalize our model to a broader range of knowledge workers and explore its potential to actively reduce interruption cost by indicating the interruptibility status to co-workers and fostering undisrupted work.

ACKNOWLEDGMENTS

The authors would like to thank all study participants. This work was funded by SNF.

REFERENCES

1. U Rajendra Acharya, K Paul Joseph, Natarajan Kannathal, Choo Min Lim, and Jasjit S Suri. 2006. Heart rate variability: a review. *Medical and biological engineering and computing* 44, 12 (2006), 1031–1051.
2. Piotr D Adamczyk and Brian P Bailey. 2004. If not now, when?: the effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 271–278.
3. Ernesto Arroyo and Ted Selker. 2011. Attention and intention goals can mediate disruption in human-computer interaction. In *Proceedings of the 13th IFIP TC 13 international conference on Human-computer interaction-Volume Part II*. Springer-Verlag, 454–470.
4. American Heart Association. 2016. Target Heart Rates. (Oct. 2016). `http://www.heart.org/HEARTORG/HealthyLiving/PhysicalActivity/FitnessBasics/Target-Heart-Rates_UCM_434341_Article.jsp`
5. Alberto Bacchelli and Christian Bird. 2013. Expectations, outcomes, and challenges of modern code review. In *Proceedings of the 2013 international conference on software engineering*. IEEE Press, 712–721.
6. Brian P Bailey and Shamsi T Iqbal. 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 (2008), 21.
7. Brian P Bailey and Joseph A Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior* 22, 4 (2006), 685–708.
8. Brian P Bailey, Joseph A Konstan, and John V Carlis. 2001. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *Proceedings of INTERACT*, Vol. 1. 593–601.
9. James Bo Begole, Nicholas E Matsakis, and John C Tang. 2004. Lilsys: sensing unavailability. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. ACM, 511–514.
10. Jelmer P Borst, Niels A Taatgen, and Hedderik van Rijn. 2015. What makes interruptions disruptive?: A process-model account of the effects of the problem state bottleneck on task interruption and resumption. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 2971–2980.
11. Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
12. David M Cades, Deborah A Boehm Davis, J Gregory Trafton, and Christopher A Monk. 2007. Does the difficulty of an interruption affect our ability to resume?. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 51. SAGE Publications Sage CA: Los Angeles, CA, 234–238.
13. Daniel Chen, Jamie Hart, and Roel Vertegaal. 2007. Towards a physiological model of user interruptability. In *IFIP Conference on Human-Computer Interaction*. Springer, 439–451.
14. Daniel Chen and Roel Vertegaal. 2004. Using mental load for managing interruptions in physiologically attentive user interfaces. In *CHI'04 extended abstracts on Human factors in computing systems*. ACM, 1513–1516.
15. Jan Chong and Rosanne Siino. 2006. Interruptions on software teams: a comparison of paired and solo programmers. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM, 29–38.
16. Mary Czerwinski, Edward Cutrell, and Eric Horvitz. 2000. Instant messaging: Effects of relevance and timing. In *People and computers XIV: Proceedings of HCI*, Vol. 2. British Computer Society, 71–76.
17. Polar Electro. 2017. Equine H7 heart rate sensor belt set. `https://www.polar.com/en/products/equine/accessories/equine_H7_heart_rate_sensor_belt_set`. (2017). [Online; accessed 19-September-2017].
18. Robert Fisher and Reid Simmons. 2011. Smartphone interruptibility using density-weighted uncertainty sampling with reinforcement learning. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, Vol. 1. IEEE, 436–441.
19. Inc. Fitbit. 2017. Fitbit Charge 2. `https://www.fitbit.com/de/charge2`. (2017). [Online; accessed 19-September-2017].
20. James Fogarty, Andrew J Ko, Htet Htet Aung, Elspeth Golden, Karen P Tang, and Scott E Hudson. 2005. Examining task engagement in sensor-based statistical models of human interruptibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 331–340.
21. James Fogarty, Jennifer Lai, and Jim Christensen. 2004. Presence versus availability: the design and evaluation of a context-aware communication client. *International Journal of Human-Computer Studies* 61, 3 (2004), 299–317.
22. Thomas Fritz, Andrew Begel, Sebastian C Müller, Serap Yigit-Elliott, and Manuela Züger. 2014. Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th International Conference on Software Engineering*. ACM, 402–413.
23. Tony Gillie and Donald Broadbent. 1989. What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological research* 50, 4 (1989), 243–250.

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

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

26. Fangfang Guo, Yu Li, Mohan S Kankanhalli, and Michael S Brown. 2013. An evaluation of wearable activity monitoring devices. In *Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia*. ACM, 31–34.

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

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

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

30. Joyce Ho and Stephen S Intille. 2005. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 909–918.

31. Eric Horvitz and Johnson Apacible. 2003. Learning and reasoning about interruption. In *Proceedings of the 5th international conference on Multimodal interfaces*. ACM, 20–27.

32. Eric Horvitz, Paul Koch, and Johnson Apacible. 2004. BusyBody: creating and fielding personalized models of the cost of interruption. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. ACM, 507–510.

33. Eric Horvitz, Paul Koch, Carl M Kadie, and Andy Jacobs. 2002. Coordinate: Probabilistic forecasting of presence and availability. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 224–233.

34. Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. 2003. Predicting human interruptibility with sensors: a Wizard of Oz feasibility study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 257–264.

35. Shamsi T Iqbal, Piotr D Adamczyk, Xianjun Sam Zheng, and Brian P Bailey. 2004. Changes in mental workload during task execution. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology*.

36. Shamsi T Iqbal and Brian P Bailey. 2007. Understanding and developing models for detecting and differentiating breakpoints during interactive tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 697–706.

37. Shamsi T Iqbal and Brian P Bailey. 2008. Effects of intelligent notification management on users and their tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 93–102.

38. Shamsi T Iqbal and Eric Horvitz. 2007. Disruption and recovery of computing tasks: field study, analysis, and directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 677–686.

39. Ellen Isaacs, Steve Whittaker, David Frohlich, and Brid O'Conaill. 1997. Informal communication re-examined: New functions for video in supporting opportunistic encounters. *Video-mediated communication* 997 (1997), 459–485.

40. Eric Jones, Travis Oliphant, Pearu Peterson, and others. 2001–. SciPy: Open source scientific tools for Python. (2001–). `http://www.scipy.org/` [Online; accessed 08.01.2018].

41. Ashish Kapoor and Eric Horvitz. 2007. Principles of lifelong learning for predictive user modeling. *User Modeling 2007* (2007), 37–46.

42. Ashish Kapoor and Eric Horvitz. 2008. Experience sampling for building predictive user models: a comparative study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 657–666.

43. Juha Karvonen and Timo Vuorimaa. 1988. Heart rate and exercise intensity during sports activities. *Sports Medicine* 5, 5 (1988), 303–311.

44. Ioanna Katidioti, Jelmer P Borst, Douwe J Bierens de Haan, Tamara Pepping, Marieke K van Vugt, and Niels A Taatgen. 2016. Interrupted by Your Pupil: An Interruption Management System Based on Pupil Dilation. *International Journal of Human–Computer Interaction* 32, 10 (2016), 791–801.

45. Kyohei Komuro, Yuichiro Fujimoto, and Kinya Fujita. 2017. Relationship Between Worker Interruptibility and Work Transitions Detected by Smartphone. In *International Conference on Human-Computer Interaction*. Springer, 687–699.

46. Jennifer Lai, Sachiko Yoshihama, Thomas Bridgman, Mark Podlaseck, Paul B Chou, and Danny C Wong. 2003. MyTeam: Availability Awareness Through the Use of Sensor Data.. In *INTERACT*.

47. Andy Liaw, Matthew Wiener, and others. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

48. Choubeila Maaoui, Alain Pruski, and F Abdat. 2010. *Emotion recognition through physiological signals for human-machine communication*. INTECH Open Access Publisher.

49. Gloria Mark, Mary Czerwinski, Shamsi Iqbal, and Paul Johns. 2016. Workplace Indicators of Mood: Behavioral and Cognitive Correlates of Mood Among Information Workers. In *Proceedings of the 6th International Conference on Digital Health Conference*. ACM, 29–36.

50. Gloria Mark, Victor M Gonzalez, and Justin Harris. 2005. No task left behind?: examining the nature of fragmented work. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 321–330.

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

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

53. Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, Paul Johns, Akane Sano, and Yuliya Lutchyn. 2016. Email duration, batching and self-interruption: Patterns of email use on productivity and stress. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1717–1728.

54. Santosh Mathan, Stephen Whitlow, Michael Dorneich, Patricia Ververs, and Gene Davis. 2007. Neurophysiological estimation of interruptibility: Demonstrating feasibility in a field context. In *In Proceedings of the 4th International Conference of the Augmented Cognition Society*.

55. Andre N Meyer, Laura E Barton, Gail C Murphy, Thomas Zimmermann, and Thomas Fritz. 2017. The Work Life of Developers: Activities, Switches and Perceived Productivity. *IEEE Transactions on Software Engineering* (2017).

56. André N Meyer, Thomas Fritz, Gail C Murphy, and Thomas Zimmermann. 2014. Software developers' perceptions of productivity. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM, 19–29.

57. Microsoft. 2017. Microsoft Graph. `https://graph.microsoft.io`. (2017). [Online; accessed 19-September-2017].

58. Hamid Turab Mirza, Ling Chen, Gencai Chen, Ibrar Hussain, and Xufeng He. 2011. Switch detector: an activity spotting system for desktop. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2285–2288.

59. Christopher A Monk, J Gregory Trafton, and Deborah A Boehm-Davis. 2008. The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied* 14, 4 (2008), 299.

60. Hawley E Montgomery-Downs, Salvatore P Insana, and Jonathan A Bond. 2012. Movement toward a novel activity monitoring device. *Sleep and Breathing* 16, 3 (2012), 913–917.

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

62. Sebastian C Müller and Thomas Fritz. 2015. Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress. In *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, Vol. 1. IEEE, 688–699.

63. Rahul Nair, Stephen Voida, and Elizabeth D Mynatt. 2005. Frequency-based detection of task switches. In *Proceedings of the 19th British HCI Group Annual Conference*, Vol. 2. 94–99.

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

65. June J Pilcher, Douglas R Ginter, and Brigitte Sadowsky. 1997. Sleep quality versus sleep quantity: relationships between sleep and measures of health, well-being and sleepiness in college students. *Journal of psychosomatic research* 42, 6 (1997), 583–596.

66. Peter Richter, Thomas Wagner, Ralf Heger, and Gunther Weise. 1998. Psychophysiological analysis of mental load during driving on rural roads-a quasi-experimental field study. *Ergonomics* 41, 5 (1998), 593–609.

67. Lior Rokach and Oded Maimon. 2005. Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 35, 4 (2005), 476–487.

68. Mark R Rosekind, Kevin B Gregory, Melissa M Mallis, Summer L Brandt, Brian Seal, and Debra Lerner. 2010. The cost of poor sleep: workplace productivity loss and associated costs. *Journal of Occupational and Environmental Medicine* 52, 1 (2010), 91–98.

69. Claude Sammut and Geoffrey I Webb. 2011. *Encyclopedia of machine learning*. Springer Science & Business Media.

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

71. Saurabh Sarkar and Chris Parnin. 2017. Characterizing and predicting mental fatigue during programming tasks. In *Proceedings of the 2nd International Workshop on Emotion Awareness in Software Engineering*. IEEE Press, 32–37.

72. Tammar Shrot, Avi Rosenfeld, Jennifer Golbeck, and Sarit Kraus. 2014. Crisp: an interruption management algorithm based on collaborative filtering. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 3035–3044.

73. Hermann Stern, Viktoria Pammer, and Stefanie N Lindstaedt. 2011. A preliminary study on interruptibility detection based on location and calendar information. *Proc. CoSDEO* 11 (2011).

74. Margaret-Anne Storey, Alexey Zagalsky, Fernando Figueira Filho, Leif Singer, and Daniel M German. 2017. How social and communication channels shape and challenge a participatory culture in software development. *IEEE Transactions on Software Engineering* 43, 2 (2017), 185–204.

75. Edward R Sykes. 2011. Interruptions in the workplace: A case study to reduce their effects. *International Journal of Information Management* 31, 4 (2011), 385–394.

76. Takahiro Tanaka and Kinya Fujita. 2011. Study of user interruptibility estimation based on focused application switching. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 721–724.

77. Rini van Solingen, Egon Berghout, and Frank van Latum. 1998. Interrupts: just a minute never is. *IEEE software* 15, 5 (1998), 97.

78. Bogdan Vasilescu, Kelly Blincoe, Qi Xuan, Casey Casalnuovo, Daniela Damian, Premkumar Devanbu, and Vladimir Filkov. 2016. The sky is not the limit: multitasking across github projects. In *Proceedings of the 38th International Conference on Software Engineering*. ACM, 994–1005.

79. JA Veltman and AWK Gaillard. 1998. Physiological workload reactions to increasing levels of task difficulty. *Ergonomics* 41, 5 (1998), 656–669.

80. Stjepan Vidaček, Ljiljana Kaliterna, Biserka Radošević-Vidaček, and Simon Folkard. 1986. Productivity on a weekly rotating shift system: circadian adjustment and sleep deprivation effects? *Ergonomics* 29, 12 (1986), 1583–1590.

81. Aku Visuri, Niels van Berkel, Chu Luo, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. 2017. Predicting Interruptibility for Manual Data Collection: A Cluster-Based User Model. (2017).

82. Peter Vorburger, Abraham Bernstein, and Alen Zurfluh. 2011. Interruptability Prediction Using Motion Detection. In *1st Int. Workshop on Managing Context Information in Mobile and Pervasive Environments MCMP-05*.

83. Robert Wang, Gordon Blackburn, Milind Desai, Dermot Phelan, Lauren Gillinov, Penny Houghtaling, and Marc Gillinov. 2017. Accuracy of wrist-worn heart rate monitors. *Jama cardiology* 2, 1 (2017), 104–106.

84. Jacqueline Wijsman, Bernard Grundlehner, Hao Liu, Hermie Hermens, and Julien Penders. 2011. Towards mental stress detection using wearable physiological sensors. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 1798–1801.

85. Glenn F Wilson. 2002. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology* 12, 1 (2002), 3–18.

86. Borejda Xhyheri, Olivia Manfrini, Massimiliano Mazzolini, Carmine Pizzi, and Raffaele Bugiardini. 2012. Heart rate variability today. *Progress in cardiovascular diseases* 55, 3 (2012), 321–331.

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

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