Using Biometrics Signals to Predict Stress, Focus, and Awakeness of Knowledge Workers

Authors will be visible after double blind review

ABSTRACT

Knowledge workers face many challenges as they work: work is fragmented, disruptions are constant, tasks are complex, and work hours can be long. These challenges make it difficult for a knowledge worker to stay focused, impacting the quality of work performed and productivity. Biometric sensors make it possible to continuously monitor human aspects, such as focus, awakeness and stress providing opportunities to help knowledge workers combat workplace challenges. We investigate how biometric measures, such as heart rate, respiration rate, and galvanic skin response, can be used to create a machine learning model able to accurately predict stress, awakeness, and focus levels. In a field study with 14 professionals over an eight-week period we collected and compared biometric and self-reporting data. Our results show that, with only a few days of self-reporting data, biometrics can be used to predict the level of stress, focus, and awakeness of operators.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*;

KEYWORDS

Biometrics, Stress, Awakeness, Focus, Computer Interaction

ACM Reference Format:

Authors will be visible after double blind review. 2019. Using Biometrics Signals to Predict Stress, Focus, and Awakeness of Knowledge Workers. In *Proceedings of ACM CHI Conference on Human Factors in Computing Systems (CHI'19).* ACM, New York, NY, USA, Article 4, 13 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

'The most valuable asset of a 21st-century institution (whether business or non-business) will be its knowledge workers and their productivity' [22]. Yet, in today's workplaces, knowledge workers and their time are not necessarily treated as the most valuable asset, and they constantly face challenges, such as a high work fragmentation, continuous disruptions and distractions, highly complex and demanding tasks, and long working hours [19, 34, 50]. These challenges make it difficult for a knowledge worker to stay focused and can impact both the quality of work performed and productivity. For instance, the continuous disruptions that knowledge workers face in office work environments can lead to a higher error rate and a lower performance [4, 50]. In some cases, the consequences are even more severe; in control room settings, the consequences can be accidents that cost millions of dollars in losses. Less immediate, but just as insidious, stress at the workplace is a growing concern and one of the most common work-related health problems. It leads to fatigue, burnout and various other illnesses, ultimately resulting in work absences and marked productivity losses [29, 41, 71].

Being able to continuously measure human aspects — such as focus, awakeness, and stress — in the workplace bears huge potential for better supporting knowledge workers, combatting workplace challenges, and improving knowledge worker productivity. Such measures could be used to provide better management of disruptions [44, 85], to automatically adjust lighting to reduce sleepiness among control room operators, or to continuously monitor stress levels to trigger pro-active interventions and reduce health problems.

Fortunately, with recent advances in sensing technologies there is increasing hope that we can accurately and automatically measure such human aspects. A number of studies have already discussed and investigated the use of biometrics to measure aspects such as a person's cognitive and emotional states [5, 49, 66, 68, 83]. Studies have also looked more specifically at using biometrics to measure stress [21, 72], or awakeness / (energetic) arousal [25, 36, 53]. Yet, very few studies measure productivity-related factors in the workplace over an extended period of time, and little is known about measuring the more abstract concept of focus. Given the importance of focus, described as a combination of engagement and challenge in a work task by Mark et al. [51], to an office workers' productivity, this factor also deserves further study.

The objective of our research is to develop continuous (and automatic) measures of focus, awakeness, and stress in the workplace. These measures are targeted with an eye towards improving knowledge workers' productivity and well-being in the future. For instance, by automatically protecting the worker from audible interruptions during a high focus period by showering him/her with white noise or by recommending stress-reducing interventions during an extended period of high stress.

CHI'19, May 2019, Glasgow, UK
2019. ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

Our work builds upon previous work and extends it by performing an eight-week study in the workplace with 14 knowledge workers, collecting biometric and computer interaction data as well as experience samples on focus, awakeness, and stress. Our participants, who perform various job functions for a research and development group within a single large corporation, wore a single biometric armband sensor[1] with low invasiveness that captured heart- and skin-related measures. This modality was chosen to ease longitudinal deployment in the field. Using machine learning, we created classifiers and analyzed their ability to predict the three productivity-related human aspects. In addition, we compared biometric features with computer interaction features in their predictive power, we analyzed how the combination of these two training sets compares to the performance of each set individually.

The results of our analysis show that biometrics of a single minimally invasive sensor can be used to predict productivity-related indicators accurately, with the abstract concept of focus (predictably) being the hardest to detect. The results also show that knowledge workers' self-reported levels of stress, focus and awakeness and their physiological manifestation and prediction can vary a lot between individuals. Our results further show that overall combining biometric features with computer interaction information results in an improvement over each of the training sets individually. The main contributions of our work are:

- The creation and analysis of measures for the automatic monitoring of knowledge workers' awakeness, stress, and focus in the workplace based on an eight-week field study with 14 office workers.
- A thorough analysis of ways to increase performance, including the training sample size, time window to collect data, and the inclusion of computer interaction features.
- A discussion on the impact of applying this research to possible target workspaces, besides a reflection on aspects that can be further improved in future studies.

2 FIELD STUDY

We conducted an eight-week field study with 14 participants to investigate the feasibility of predicting stress, focus, and awakeness based on biometric signals. On each workday for eight consecutive weeks, each of the 14 participants wore a biometric sensor while performing their normal work tasks and completed two electronic surveys to report their current self-perceived levels of stress, focus, and awakeness. During the eight week study period, a computer activity tracker was installed and active on 10 of the participants' company-issued laptops (note: a few declined for privacy reasons).

[1]Biovotion Everion sensor [10]

The goal of our study is to understand the efficacy of biometric measurements in predicting stress, focus, and awakeness. To that end we posed the following research questions.

RQ1: Can we use biometric measurements to accurately predict stress, focus, and awakeness?

RQ2: What is the minimum number of samples needed to accurately predict stress, focus, and awakeness?

RQ3: What is the minimum time window needed to accurately predict stress, focus, and awakeness?

RQ4: How do biometrics compare to computer interaction measurements for predicting stress, focus, and awakeness?

Participants

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

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

Data Collection

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

Biometric Sensors. Figure 1 illustrates Biovotion's Everion [10], 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. Previous studies [35, 38, 70, 81, 86, 87] have used similar devices [23, 27, 61] to capture psycho-physiological and biometric measurements.

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

Figure 1: Everion sensor used to collect biometric measures.

Biometric Measurement	Units of Measure
Physical Activity	[2, 30]
Intensity of motion	(No unit)
Energy Expenditure	Calories per second (cal/s)
Step counter	Steps
Heart	[36–38, 56]
Heart rate	Beats per Minute (bpm)
Blood pulse wave	(No Unit)
Heart rate variability	Milliseconds (ms)
Blood oxygenation	Percent (%)
Blood perfusion	(No unit)
Skin	[36, 38]
Galvanic skin response	kOhm
Skin temperature	Degrees Celsius (°C)
Respiration	[36, 38, 52, 56]
Respiratory rate	Breaths per Minute (bpm)

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

Surveys. We sent via text message a survey request to each participant two times per work day. 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). The same survey was sent each time:

(1) How awake are you right now?
(2) How stressed do you feel right now?
(3) How focused on work are you right now?

We used the phrase "right now" to capture each aspect in the moment (so as to permit later prediction of each aspect based on biometric data). The wordings of the questions are based on a previous survey of individuals in an organizational context [33]. The use of awakeness (rather than sleepiness) in Question 1 is inspired by previous work [82] and to some extent also captures the "arousal" aspect of the affective space [69].

Following guidelines from similar previous studies [28, 76], we asked the participants to respond to each question using a 5-point Likert scale ranging from 1 (not at all) 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.

Data Preparation

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

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

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

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

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

Figure 2 illustrates the number of survey responses with associated biometric data for each study participant. 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.

Feature Extraction. We extracted features from the biometric data to provide as input to machine learning models. Previous studies [8, 86] 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 [87], ranging from 10 seconds to 3 hours. Specifically, we considered the following time windows: *10sec, 20sec,*

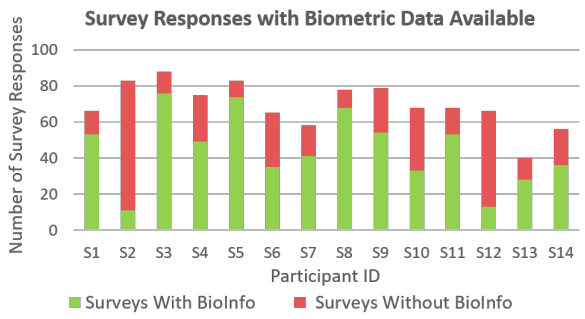

Figure 2: Availability of biometric data per participant. Green: survey responses with biometric data; red: responses with no biometric data.

30sec, 45sec, 1min, 2min, 3min, 5min, 7.5min, 10min, 20min, 30min, 45min, 1hour, 2hour, 3hour.

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

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

We binarized the survey responses as follows. For each participant, we calculated the median response value for each question. We classified each response below the median as 0 ('negative') and each response above the median as 1 ('positive'). The distribution for the stress question skewed

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

Table 2: Distribution of responses per participant over the 5-point Likert scale (1: not at all; 5: extremely) for each question.

left, so we included the median values in the 'positive' class, while the distributions for focus and awakeness skewed right, so we included those median values in the 'negative' class.

Oversampling. Even after binarizing the responses as described in the previous section, we found the distribution of responses was still quite imbalanced for many of our participants. This can be seen in the distribution columns in Table 3. To combat this, we applied random oversampling to our training sets, which artificially rebalances the dataset by creating randomly replicated data in the minority class. This has been a commonly used technique in previous studies on unbalanced datasets [12, 84].

3 ANALYSIS AND RESULTS

To evaluate the efficacy of continuously predicting a knowledge worker's stress, focus, and awakeness in the workplace, we trained and tested machine learning classifiers using a leave-one-out cross validation. For prediction, we used the features extracted from the collected data as the input data, and binarized the participants' self-reported responses on stress, focus, and awakeness into two classes each (e.g., 'stressed' and 'not stressed') to use as output measures.

Predicting Stress, Focus & Awakeness

As a first step, we compared multiple classifiers using the popular machine learning library scikit-learn [63] and performing a grid search analysis to determine the optimal hyperparameters for each classifier. Our analysis showed that random forest outperforms all other classifiers, including Naïve Bayes, decision trees, support vector machine, random forest, and a multilayer perceptron neural network. The optimal values for random forest and the three output measures are listed in Table 4. For the remainder of this paper, we will

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

Table 3: Prediction results individual models. The baseline row states averaged results of our baseline classifier, the general row states the averaged results of model trained on all participants. Distribution presents the distribution of the binarized outcome variable.

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

Table 4: Hyperparameters selected by grid search analysis for tuning our random forest models. K refers to the number of features selected.

be referring to random forest classifiers trained with these hyperparameters.

Results of Individual Classifiers

Since peoples' experience of stress, focus, and awakeness (as well as their physiological manifestations) can vary a lot (e.g., [40]), we first trained individual classifiers for each participant (rather than a general one for all participants). The results of our analysis are reported in Table 3. 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 can be more adequate metrics in this case [9, 40, 84].

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

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

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

Table 5: Averaged Gini importance of each feature category per response variable.

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 chose to use the top k features with the highest ANOVA F-value, as it is relatively simple and efficient to calculate. The values of k used were selected by grid search analysis, and are shown in Table 4.

Overall, the features that were selected as the important ones for the individual models based on the random forest algorithm varied greatly across participants. Yet, some feature categories were considered to be important more frequently than others. Table 5 shows the averaged Gini importance for the feature categories used for predicting stress. Particularly important for stress were the feature categories heart rate variability (18.3%) and skin temperature (15.7%), both of which have been shown in several previous studies to be indicators for stress [21, 47, 54]. For predicting focus, the feature categories for blood pulse wave (14.2%) and heart rate variability (13%) showed to be the most important categories, while for awakeness the most important ones were heart rate variability (13.6%) and blood pulse wave (13.1%).

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 big individual differences in biometrics, training a general model to achieve an adequate accuracy for new individuals is not necessarily possible.

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

The bottom row of Table 3 presents the averaged performance results for this approach in terms of accuracy, precision, and recall. Although the averaged precision and recall are comparable or better than those of the averaged individual results, this was at the cost of a large decrease in overall accuracy. Upon closer investigation into the performance of the general model when testing on each participant, we found that individually trained models for each participant performed much better than a general model trained over all participants. Using stress as an example, for participant S12, for whom we saw the greatest increase compared to the baseline in individual models, the general model was unable to predict a single instance of 'stressed' correctly. This is consistent with our expectations because biometric features are highly specific to individuals.

Minimum Number of Training Samples

Collecting experience samples from users is expensive, since participants are being interrupted several times a day and have to answer the questions. To minimize the number of samples to be collected from participants, we examined how the performance of individual classifiers changes over the number of samples used to train the classifier.

Participants in our study had varying levels of responsiveness to the experience sampling ranging from 10 to 76 (see Figure 2). To maintain a certain generalizability while also being able to examine a range of sample numbers for training the classifier, we decided to include the ten participants with the most samples for this analysis. As a result, we had at least 34 survey responses per participant to examine the learning curve of the classifiers. For our analysis, we thus performed a leave-one-out cross validation with random sample sets of size 1 to 33 and calculated the average through all folds of the validation.

For each of the three productivity-related aspects, we are again more interested in predicting when a worker is stressed, not awake, or not focused, the less common class in all three cases. Since the less common class can be very small, we weighted each participants' classifier performance by the percentage of the samples in this smaller class.

Individual Classifiers for Binary Prediction

The averaged performance of all individually trained random forest classifiers for 'awake' with respect to the training

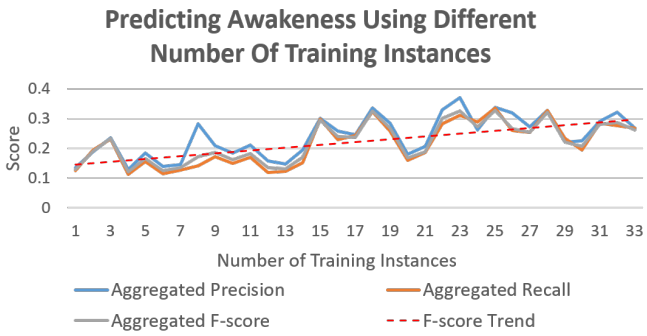

Figure 3: Performance of the per participant trained awakeness classifiers, measured in precision, recall (sensitivity), and F-score. The dotted red line represents the F-score trend.

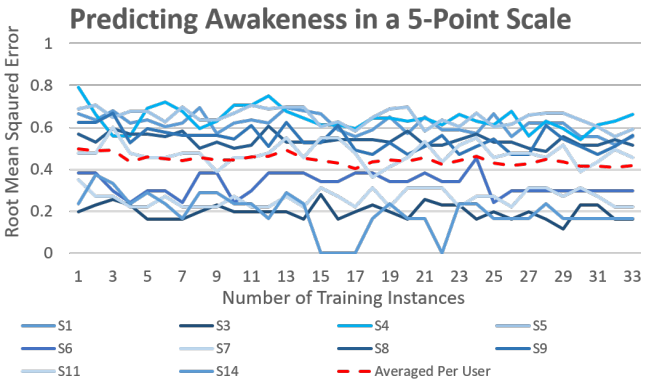

Figure 4: Performance of the per participant trained classifiers for predicting 5-point awakeness, measured in root mean squared error.

sample size is presented in Figure 3. The trend indicates a positive correlation between the number of samples in the training set and the classifiers' F-score performance, with an overall improvement of 114% (from 0.14 to 0.30) in the F-score between a training set of one sample to one with 33 samples. The trends for the remaining indicators are 29% for stress and no overall improvement for focus.

Predicting Five Classes

In a second step, we analyzed a more fine-grained prediction using the initial 5-point Likert scale responses rather than the binarized ones as output measure. Figure 4 depicts the performance of the individually trained classifiers in terms of the root mean squared error. The root mean squared error represents the distance of the predicted from the actual value, which provides a more nuanced measure of the performance in the fine-grained prediction case. The figure shows a similar trend as for the binarized prediction, in that the root mean square error averaged over all ten participants decreases with more samples (from 0.49 to 0.41 root mean square error) and

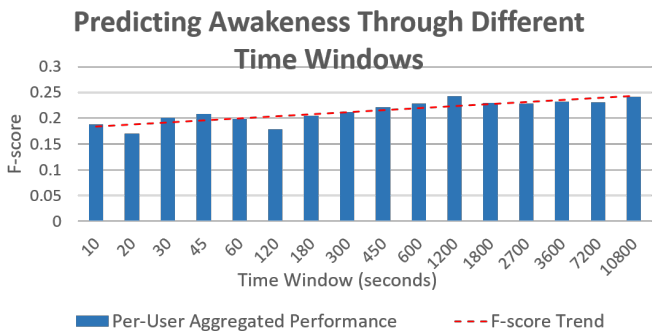

Figure 5: Performance (F-score) of individual classifiers trained on the different time windows to predict 'awake'.

thus the performance increases. At the same time, the figure also shows that the performance results for the fine-grained prediction, again, vary substantially across participants.

Minimum Time Window

In general, the less biometric data is needed to accurately predict a certain outcome measure, the easier and faster the analysis and data collection. To examine the optimal and minimum time window for the prediction of stress, focus, and awakeness, we used 16 different time windows from 10 seconds to 3 hours.For our analysis, we then trained individual classifiers for each of the 16 time windows, using only features that had a time window smaller or equal to the time window rather than using all combinations of $\{BiometricMeasures\}X\{StatisticalMetrics\}X\{TimeWindows\}$. We again used random forest and a leave-one-out cross validation to train individual classifiers. Since the number of features used for the training changed with each time window, we did not apply our feature selection in this case, but used all features available. Finally, due to the imbalance in the data, we again weighted each participants' classifier performance by the number of instances in the smaller class to calculate the average.

Figure 5 shows how the F-score changes for predicting 'awake' over the 16 different time windows. The figure shows an increasing trend in the F-score, i.e. the higher the number of included time windows included, the higher the F-score. However, there is one exception, the time window of 1200 seconds that achieves a performance close to the one for the time window 10,800 seconds (3 hours), at which point all features are included. Overall, our results thus show that while using all time windows up to 3 hours performs best, and outperforms the feature set that is solely based on a 10 second time window by 28% (from 0.18 to 0.24), the performance for a time window of 1200 seconds is already very close to optimal (everything up to 3 hours in our case).

Computer Interaction Data

Given our focus on knowledge workers (i.e., workers who generally spend a lot of time interacting with information on their computer at work), we also analyzed the use of computer interaction features

Feature collected by tool	Description
Total keystrokes per min	Sum of all types of keystrokes
Normal keystrokes per min	F[h] Not backspace and navigation
Backspace keystrokes per min	Backspace keystrokes
Navigation keystrokes per min	Arrow key keystrokes
Total clicks per min	Sum of all click types
Other clicks per min	Not right and left clicks
Left clicks per min	Left clicks
Right clicks per min	Right clicks
Scrolled distance per min	Scrolled distance in pixels
Moved distance per min	Mouse movements in pixels
Activity switches per min	Browser window title changes
Category switches per min	Activity performed category

Table 6: List of computer interaction features.

to predict focus, awakeness and stress. To collect computer interaction data, we used an open source computer interaction monitor (reference omitted for double-blind.) to track participant's mouse and keyboard activity, as well as details about their active window. The specifics of the features tracked are listed in Table 6. The tracker was installed on the computers of 10 of the 14 participants, with participants S6, S10, S12, and S13 opting out of this part of the study due to privacy concerns. Therefore, we limited this analysis to the 10 participants for whom we could calculate all features.

For calculating computer interaction features, we again used the aforementioned 16 time windows and scaled the computer interaction values if the time windows did not align. For our comparative analysis of the different sensing techniques—biometrics vs computer interaction—we then created two new feature sets for each participant in addition to the biometric one: one with only computer interaction features, and one with computer interaction features plus biometric features.

Table 7 lists the results of our analysis. The results show that in all cases, the computer interaction based model was able to improve upon the biometric model in terms of precision and recall, but not in accuracy. Further, we found that the combined model was the most effective model in terms of precision and recall for predicting stress and awakeness overall, but performed slightly worse than the model using only computer interaction features for focus.

As with the biometric models, the individual performance of both the computer interaction only models and the combined models varied quite a bit between participants. Using stress as an example again, in the computer interaction models 5 of the 10 participants saw improvements compared to the baseline, with a maximum improvement of 128% in precision, and 78% in recall. In the combined model for stress, 5 of the 10 participants saw improvements compared to the baseline, with a maximum improvement of 117% in precision and 95% in recall. Neither model was capable of correctly predicting any instances of 'stressed' for participant S7.

Since the number of features changes depending on which feature set is used, we adjusted the feature selection parameter for each of the computer interaction and combined computer interaction/biometric models. The values reported in this section were achieved using the optimal feature selection parameters we found, which are shown in Table 8.

Model/Feature Set	Accuracy	Precision	Recall	F-Score
Awakeness				
Biometrics only	0.808	0.269	0.314	0.289
CI	0.758	0.425	0.362	0.391
Biometrics + CI	0.791	0.390	0.404	0.400
Stress				
Biometrics only	0.775	0.270	0.260	0.265
CI	0.698	0.290	0.272	0.281
Biometrics + CI	0.712	0.317	0.286	0.301
Focus				
Biometrics only	0.716	0.251	0.256	0.253
CI	0.742	0.332	0.342	0.337
Biometrics + CI	0.745	0.340	0.316	0.328

Table 7: Comparison of prediction performance for the 3 different features sets for 10 participants. CI stands for computer interaction. Precision/recall refer to prediction of more important classes, i.e. stressed, not awake, not focused.

Model/Feature Set	Number of Features Selected
Stress	
CI	400
Biometrics + CI	800
Focus	
CI	20
Biometrics + CI	300
Awakeness	
CI	All
Biometrics + CI	50

Table 8: Optimal number of features used for each of the three models. CI: computer interaction

4 RELATED WORK

We review two categories of related work for this study. First, we review work related to the productivity-related indicators that we considered in our study, including how they are detected and prevented in several risk-inducing scenarios. Second, we review work related to biometrics, including how they have been studied in the office environment or applied to assess cognitive states and processes.

Productivity-related Indicators

Previous studies have sought to measure, detect, and predict productivity indicators.

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 [64]. 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 [11, 39].

The second approach, which we have chosen for our study, is to measure autonomic nervous system (ANS) activity by analyzing biometric signals of the human body, such as blood pressure, heartbeat, and temperature [47, 77, 78]. In particular, changes in heart rate variability are associated with cognitive and emotional stress [21, 54]. This second approach has been used successfully by several past studies [32, 55, 60].

Stress in the workplace and its effects on service providers has been analyzed in call centers [40] and in the context of the perceived imbalance between resources and demands [13]. These studies considered several factors such as personality traits, career-related goals and attitudes, and life outside of work, and these factors were correlated with stress levels and burnout.

Evans and Johnson [24] investigated the correlation between noise in the workplace and stress levels. They found that workers exposed to open-office noise showed aftereffects that indicate motivational deficits. The population for their experiment comprised 40 female clerical workers, who were randomly assigned to a control condition or to three-hour low-intensity noise room designed to simulate typical open-office noise levels.

Hovsepian et al. [42] 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.

Awakeness. Sleepiness (lack of awakeness) and its associated risk of serious injury to passengers has been studied in the context of automobile accidents [15, 59]. These studies show a strong association between the level of acute driver sleepiness and the risk of injury crash. Connor et al. [15] 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 [59] 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.

Focus. Focus refers to the allocation of limited cognitive processing resources [3]. Mark et al. [51] studied engagement in workplace activities by analyzing the digital activity of 32 information workers in situ for 5 days to understand how attentional states change with context. They found that boredom is highest in the early afternoon and focus peaks in the middle of the afternoon. They also found that doing work that requires focus correlates with stress, while rote work correlates with happiness.

Interruptions in the office are a common barrier keeping workers from sustaining focus on their work related activities, particularly when the interruptions occur at inopportune moments. Such interruptions may include emails, alerts, or interactions with co-workers[14, 34, 45]. Interruptions in inopportune times can have negative effects that range from higher error rate and lower overall performance to an increase in stress and frustration [4, 18, 50]. External interruptions may cause workers to enter a "chain of distraction" [45]. This chain is composed by stages of preparation, diversion, resumption and recovery that result in time away from an ongoing task.

Other studied constructs that relate to focus in the workplace include cognitive absorption, cognitive engagement, flow, and mindfulness. Cognitive absorption describes periods of time in which a person experiences total immersion in an activity. This state is also accompanied by a sense of deep enjoyment, a feeling of control, curiosity, and not realizing the passing of time. It has been associated with ease of use and perceived usefulness of information technology [1]. Cognitive engagement is described [79] as a period of strong focus in an activity without the feeling of a sense of control of the situation. Flow [17], and mindfulness [20, 80] 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.

Biometrics

Our study investigates the prediction of multiple productivity-related indicators (stress, focus, and awakeness) using two different types of measurements, biometric signals and computer interaction, over eight weeks in a real-life office setting. Existing work [35, 38, 58, 62, 65, 70, 81, 86] analyzes a broad array of biometric signals and correlates them with individual's cognitive states and processes. For example, Zuger et al. [87] used biometrics to sense interruptibility in the office [87]. Biometric signals have also been studied in the context of technology users. For example, Parnin [62] analyzes electromyography to measure sub-vocal utterances, and how these might be correlated with the programmer's perceived difficulty of programming tasks. Similarly, biometrics have been used to measure code difficulty by using biometric sensors [31] and using Near Infrared Spectroscopy to measure developer's cerebral blood flow [58].

Eye tracking technology [6, 16, 67] and brain activity [43, 73] have been used in previous studies to analyze different tasks in an office environment. Eye tracking has been used to analyze memory load and processing load by inspecting task-evoked pupillary response and pupil size [5]. Similar studies have shown high correlation between pupil size and mental workload of subtasks [5] and cognitive load [48]. Brain activity has been associated with different mental states [7] 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 [74, 75]. 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.

5 DISCUSSION

Target Workspaces: Currently, the cost of biometric sensors and necessary infrastructure, such as automated light and sound systems for adjusting the environment, makes our approach most appropriate for high-value workspaces, such as control rooms,

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

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

Ground Truth and Self-Reporting: One of the key points for developing a good classifier for awakeness, focus and stress, is to collect a valid ground truth to be used as the output measure. When designing the study, we therefore spent an extensive amount of time on determining the exact questions to ask in the experience sampling, consulting experts in the area, and basing the questions and wording on previous research and studies. Yet, the reliability and validity of self-reports have been questioned in the past due to subjective biases, lack of care in reporting, and the highly individual nature of reporting aspects such as stress [40, 42]. In addition, in contexts such as the workplace, employees might be afraid to genuinely report levels of aspects, such as sleepiness. Hence, there is a chance that the collected experience samples do not adequately reflect the ground truth of the underlying variable 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 THREATS TO VALIDITY

There are numerous threats to validity to our study.

External Validity: It is a threat that the results of this study will not generalize to a broader population of office workers. To address this concern, we collected participants from a wide variety of different departments with different age ranges, genders, work experience, and working in different positions.

Another threat is that our results might not generalize to a different office environment. We have conducted this study in a typical office environment, common among technology workers across the world. These office environments control for a series of variables to make them standard world wide such as controlled temperature and lighting.

Internal Validity: This study tries to find correlations between biometric features and different productivity-related indicators (stress, focus, and awakeness). Nonetheless, biometric signals are influenced by far more variables than the ones this study comprehends. Therefore, trying to draw a strong causality between the biometric features and the productivity-related aspects would be inaccurate. To address this concern 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.

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 threat we used an state-of-the-art device that captures a large number of highly accurate biometric signals. We collected the most commonly analyzed biometrics that historically have shown correlation with productivity aspects from each user. We also perform a grid analysis in our machine learning model to pick the number of features that is more predictive of the productivity-related indicators, therefore removing the biometric features that may be unrelated to the indicators.

7 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 examine the ability of biometrics for predicting these producitivity-related aspects. The longitude and in situ placement of the study—two key components of the study—as well as the breadth of human aspects examined, support and extend previous work. Based on twice collected daily survey responses and continually collected biometrics data we were able to create a model that was able to predict awakeness, stress, and focus, outperforming the baseline by as much as 84.6% in the case of awakeness. The analysis further shows that there is a continuous increase in performance with the number of training instances, suggesting that there is more potential to improve the model's performance with further data points. Interestingly, while aspects such as stress have previously been predominantly linked to biometrics, our comparison with a model build from computer interaction features shows that even these types of features might bear big potential in predicting some of these productivity-related human aspects in the workplace.

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

REFERENCES

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

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

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

[4] 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.

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

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

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

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

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

[10] Biovotion. 2018. Everion. https://www.biovotion.com/everion/. [Online; accessed 20-September-2018].

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

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

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

[14] 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.

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

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

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

[18] 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.

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

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

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

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

[23] 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. [Online; accessed 19-September-2017].

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

[25] Anja Exler, Andrea Schankin, Christoph Klebsattel, and Michael Beigl. 2016. A Wearable System for Mood Assessment Considering Smartphone Features and Data from Mobile ECGs. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16).* 1153–1161.

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

[27] Inc. Fitbit. 2017. Fitbit Charge 2. https://www.fitbit.com/de/charge2. [Online; accessed 19-September-2017].

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

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

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

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

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

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

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

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

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

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

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

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

[40] Javier Hernandez, Rob R. Morris, and Rosalind W. Picard. 2011. Call Center Stress Recognition with Person-Specific Models. In *International Conference on Affective Computing and Intelligent Interaction*. 125–âĂŞ134.

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

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

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

[44] Shamsi T. Iqbal and Brian P. Bailey. 2005. Investigating the Effectiveness of Mental Workload As a Predictor of Opportune Moments for Interruption. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, 1489–1492.

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

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

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

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

[49] Arthur F. Kramer. 1990. Physiological metrics of mental workload: A review of recent progress. (1990).

[50] 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.

[51] 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.

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

[53] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: An Intelligent System for Emotional Memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, 849–858.

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

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

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

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

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

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

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

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

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

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

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

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

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

[68] Dennis Rowe, John Silvert, and Don Irwin. 1998. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 480–487.

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

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

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

[72] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Troster, and Ulrike Ehlert. 2010. Discriminating Stress From Cognitive Load Using a Wearable EDA Device. *Trans. on Information Technology in Biomedicine* 14, 2 (2010).

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

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

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

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

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

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

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

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

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

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

[83] Glenn F. Wilson. 2002. An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychphysiological Measures. *International Journal of Aviation Psychology* 12, 1 (2002).

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

[85] 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.

[86] 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.

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