Noname manuscript No.
(will be inserted by the editor)

Detecting work stress in offices by combining unobtrusive sensors

Saskia Koldijk · Mark A. Neerincx · Wessel Kraaij

Received: date / Accepted: date

Abstract Employees often report the experience of stress at work. In the SWELL project we investigate how new context aware pervasive systems can support knowledge workers to diminish stress. The focus of this paper is on developing automatic classifiers to infer working conditions and stress related mental states from a multimodal set of sensor data (computer logging, facial expressions, posture and physiology). We address two methodological and applied machine learning challenges: 1) Detecting work stress using several (physically) unobtrusive sensors, and 2) Taking into account individual differences. A comparison of several classification approaches showed that, for our SWELL-KW dataset, neutral and stressful working conditions can be distinguished with 90% accuracy by means of SVM. Posture yields most valuable information, followed by facial expressions. Furthermore, we found that the subjective variable 'mental effort' can be better predicted from sensor data than e.g. 'perceived stress'. A comparison of several regression approaches showed that mental effort can be predicted best by a decision tree (correlation of 0.82). Facial expressions yield most valuable information, followed by posture. We find that especially for estimating mental states it makes sense to address individual differences. When we train models on particular subgroups of similar users, (in almost all cases) a specialized model performs equally well or better than a generic model.

Keywords Machine learning · mental state inference · stress · individual differences · computer logging · facial expressions · posture · physiology

Saskia Koldijk, E-mail: s.koldijk@cs.ru.nl
Intelligent Systems, Radboud University & TNO, The Netherlands.

Mark A. Neerincx, E-mail: mark.neerincx@tno.nl
Interactive Intelligence, Delft University of Technology & TNO, The Netherlands.

Wessel Kraaij, E-mail: wessel.kraaij@tno.nl
Intelligent Systems, Radboud University & TNO, The Netherlands.

1 Introduction

Employees often report the experience of stress at work, which can in the worst case lead to burn-out. Stress is a broad concept referring to psychological and biological processes during emotional and cognitive demanding situations. We follow a pragmatic approach and decompose stress in three factors that can be measured more precisely: (1) the task load, which poses demands on the worker, (2) the mental effort, which the worker needs to handle the task and (3) the emotional response that is raised, in terms of arousal and valence.

In the area of stress research, questionnaires are commonly used to get insight in the general working experiences (e.g. Zapf (1993)), but little is known on the immediate effects of stressors at work. Work in the area of affective computing investigates the possibility of inferring stress and emotion from sensor data (see e.g. Matthews et al. (2005)). To investigate the direct effect of different degrees of mental load, typically standardized tasks are used in a lab setting, like remembering digits. These tasks are very simple and not representative of 'real' office work. Furthermore, work on user state modeling is often performed in a process control context, e.g. on naval ships (Neerincx et al., 2009) or in flight control. Only little work is done on user state modeling in an office context.

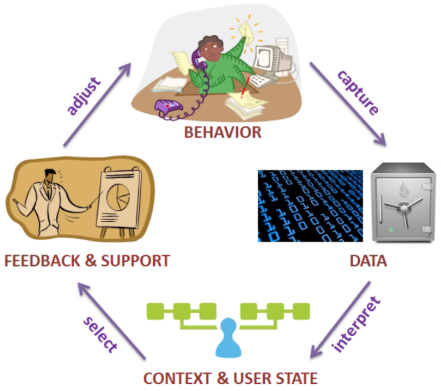

Fig. 1 SWELL approach.

In the SWELL project we investigate how unobtrusive and easily available sensors can be used in offices, to detect stress and the context in which it appears in real-time (see Figure 1; Koldijk (2012)). Based upon this information, we aim to develop pervasive supporting technology that is optimally adapted to the current working context and mental state of the user. Knowledge workers can then directly act, gaining a more healthy work style and preventing stress building up. Trends like 'quantified self' already show the potential of collecting personal sensor data (e.g. heart rate, activity patterns) for health

improvement. Personal sensor data is relatively easy to collect nowadays, the challenge is making sense of this data.

The focus of this paper is on developing automatic classifiers to infer working conditions and stress related mental states from a multimodal set of sensor data: computer logging, facial expressions, posture and physiology. We present related work in Section 2. The dataset that we use is presented in Section 3. We identified two methodological and applied machine learning challenges, on which we focus our work:

1. **Using several unobtrusive sensors to detect stress in office environments.** We found that state of the art research in stress inference often relies on sophisticated sensors (e.g. eye tracker, body sensors), and/or uses data collected in rather artificial settings. We see possibilities to build human state estimation techniques for use in office environments. We aim to combine information from multiple weak indicator variables based on physically unobtrusive measurements. We address the following **research questions**: Can we distinguish stressful from non-stressful working conditions, and can we estimate mental states of office workers by using several unobtrusive sensors? Which modeling approaches are most successful? Which modalities/ features provide the most useful information? This helps to configure a minimal sensor set-up for office settings. We address these questions in Section 4.
2. **Taking into account individual differences.** We found that, in affective computing, often one generic model is learned for all users. This may work for something universal, as the expression of emotions. However, in earlier work (Koldijk et al., 2011, 2015), we found that people differ in their (work) behavior: typical behavior of users already differs per person. Moreover, the way in which people express mental effort or stress may differ. This highlights a need to build personalized models for particular users or user groups, instead of one general model. We address the following **research questions**: How important are individual differences? Can we improve performance by building personalized models for particular user groups? We address these questions in Section 5.

Finally, we present our Conclusions and Discussion in Section 6 and 7.

2 Related work

Here we present related work on affective computing, more particularly on using physiology, facial expressions, postures and computer interactions (or preferably a combination of modalities) to infer the user's mental state, e.g. in terms of stress. We describe the sensors that are used, in which context data is collected, which machine learning approaches are applied, and how individual differences are addressed.

Physiology Most often, body sensors are used to measure the physiological stress response directly. For a general overview of psycho-physiological sensor

techniques we refer to Matthews et al. (2005). Most studies report on experimental environments, since this new technology is hardly deployed yet in work environments.

In research by Riera et al. (2012), for example, brain imaging (electroencephalography, EEG) and facial electromyography (EMG) data were collected. The authors show that these sensors can be used for monitoring emotion (valence and arousal) and stress. Although its great potential, we think deploying brain imaging in a daily office setting is not yet realistic.

Other common measurements in stress research are pupil diameter and heart rhythm (electrocardiogram, ECG). Mokhayeri et al. (2011), for example, collected such data in context of the Stroop color-word test. They state that pupil diameter and ECG have great potential for stress detection.

The question that arises is: can we make an estimate of affective and mental states outside the lab? We see some potential for heart rhythm measurements (ECG) and skin conductance (GSR or EDA), with the rise of wearable sensors, which are becoming more and more integrated into devices as watches and bracelets. Bakker et al. (2012) e.g. measured skin conductance of 5 employees during working hours.

Setz et al. (2010) present work in which they use EDA measurements to distinguish cognitive load and stress. 32 participants solved arithmetic tasks on a computer, without (cognitive load condition) or with time pressure and social evaluation (stress condition). To address individual differences, data was also normalized per participant by using a baseline period. However, the non-relative features turned out to work better. Leave-one-person-out cross validation yielded an accuracy of 82% to distinguishing both conditions. The authors 'suggest the use of non-relative features combined with a linear classification method' (p.416).

Cinaz et al. (2013) present research in which they used ECG. Three calibration tasks in a laboratory setting were used to induce low, medium and high workload. This was used to train models, which were then used on 1 hour data recorded during office work. Data was aggregated over 2 minutes. They find that linear discriminant analysis (LDA) performs best in predicting mental workload (classifying 6 of 7 participants correctly), followed by k-nearest neighbor (kNN) and support vector machines (SVM).

Moreover, Yang et al. (2008) present work in which they evaluate a wearable patch style heart activity monitoring system. They extract several signals from the ECG measurements, like heart rate and heart rate variability and demonstrate correlations with stress. Andreoli et al. (2010) present the SPINE-HRV tool, which is a full-fledged heart monitoring system specially designed to be non-restricted, non-aware and non-invasive, therefore suitable in daily life use. Measuring physiological stress reactions in office settings may thus be feasible.

Facial expressions There also is potential in using behavioral cues, such as facial expressions, postures or computer interactions as indicators for the user's mental state. For an overview of machine understanding of human behavior we

refer to the survey by Pantic et al. (2007). In related work, facial expressions are widely used for inferring emotions. The data are often recorded while emotions are induced in participants. The publicly available multimodal dataset described by Soleymani et al. (2012), for example, was collected in context of watching emotion inducing video clips and consists of: face videos, audio signals, eye gaze data and physiological signals (EEG, ECG, GSR, respiration amplitude, skin temperature). Although this dataset is very interesting, emotions in a daily computer work context are probably less intense than the valence or arousal experienced during watching a movie clip.

An interesting question is whether people show facial emotions during computer work, and whether their facial expressions are indicative of mental states. Preliminary results by Dinges et al. (2005) suggest that high and low stressor situations could be discriminated based on facial activity in mouth and eyebrow regions. They applied a Hidden Markov model. They state that their algorithm has 'potential to discriminate high- from low-stressor performance bouts in 75 - 88% of subjects'.

Moreover, Craig et al. (2008) looked at facial expressions while students worked with an online tutoring system. Association rule mining identified that frustration was associated with activity in facial action units (AU) 1, 2 (inner and outer brow raiser) and 14 (dimpler); confusion was associated with AU 4 (brow lowerer), 7 (lid tightener) and 12 (lip corner puller). So, facial expressions are an interesting modality for detecting mental states.

Postures Regarding postures, Kapoor & Picard (2005) present research in which posture data was collected together with facial expressions and computer information while children solved an educational computer puzzle. Sensors in the chair were used to extract posture features (like leaning back, sitting upright) and activity level (low, medium, high). Posture information yielded the highest unimodal accuracy (82.52%) with an SVM for estimating interest (vs. uninterest). Performance was further improved by adding facial expression and computer information in a multimodal Gaussian Process approach. We conclude that posture information and movement are an interesting source for estimating the users' mental state. We see potential for posture measurements in the office, as with the Kinect, recently an affordable 3D camera with skeleton detection has entered the market.

Computer interactions Finally, in some research, stress or emotions are estimated from computer interaction data. Vizer et al. (2009) provide an overview of related work, and they present own work on the effect of stress on keystroke and linguistic features. Participants first performed a mentally or physically stressful task (e.g. remembering digits or exercising) and were then asked to write an email. They applied the following classifiers: decision tree, SVM, kNN, AdaBoost and artificial neural networks. They state that 'These techniques were selected primarily because they have been previously used to analyze keyboard behavior (e.g. kNN), or they have shown good performance across a variety of applications (e.g. SVM)' (p.878). To address individual differences,

data was also normalized per participant by using baseline samples. Results indicate that stress can produce changes in typing patterns. With an accuracy of 75% kNN generated the best results for detecting cognitive stress (based on normalized features). In general, normalization improved the accuracy of all techniques, with an average increase of 13.1%. Vizer et al. (2009) conclude: 'individual differences should be taken into consideration when developing techniques to detect stress, especially if cognitive stress is of interest' (p.879).

In the work by Khan et al. (2008) the user's mood is inferred from their computer behavior. They aggregated computer activities within a 6 and 10 minute time window around mood ratings, and applied a correlation analysis. For 31% of the 26 participants they found significant correlations between keyboard/ mouse use and valence, and for 27% of the participants they found significant correlations with arousal. They further found that valence can better be predicted for users with more experience and less self-discipline, whereas arousal can better be predicted for users that are more dutiful.

Furthermore, Epp et al. (2011) recognize 15 emotional states based upon keystroke dynamics. For classification they used decision trees as a 'simple and low-cost solution' (p. 719). They did not create user specific models due to the large variation in responses per user.

Finally, van Drunen et al. (2009) did research on computer interaction data as indicator of workload and attention. The participants performed a task in which they were asked to search for an item on a website. They found a correlation between mouse data and heart rate variability. So, computer interactions may also be an interesting modality for detecting mental states.

To conclude, all four modalities have previously been used for mental state inference with some success, although most researchers collected data in a lab setting. Only some report on experiments that are more close to real-world situations. Several machine learning approaches are applied. In most cases, classification is used to distinguish 2 or more states. Sometimes, correlation analysis is performed to assess the strength of a relation. We aim to not only compare different classification approaches, but also apply regression models to make numerical predictions of e.g. mental effort. Several researchers find that individual differences play a role. In the models individual differences are not addressed or addressed by normalizing data per participant based upon a baseline sample. Making models for subgroups of similar users seems to be a new approach.

3 Dataset

To investigate which preferably physically unobtrusive and readily available sensors are most suitable to infer working conditions and mental states of knowledge workers, a data collection study was performed. We now first describe the study in more detail and then present the resulting dataset (for more details see Koldijk et al. (2014)).

3.1 Data collection study

To collect data, we created a realistic knowledge worker setting in which the effects of external stressors on subjective experience of task load, mental effort and emotions, as well as the effects on behavior could be investigated. 25 participants (8 female, average age 25, stdv 3.25) performed knowledge worker tasks, i.e. writing reports and making presentations. To manipulate the experienced task load, we chose two stressors that are relevant in the knowledge worker context: interruptions by incoming emails and time pressure to finish a set of tasks before a deadline. So each participant worked under the following 3 working conditions:

- Neutral: the participant was allowed to work on the tasks as long as he/she needed. After a maximum of 45 minutes the participant was asked to stop and told that enough data of 'normal working' was collected.
- Stressor 'Time pressure': the time to finish all tasks was 2/3 of the time the participant needed in the neutral condition (and maximally 30 minutes).
- Stressor 'Interruptions': 8 emails were sent to the participant during the task. Some were relevant to one of the tasks, others were irrelevant. Some emails required a reply, others did not. Examples are: "Could you look up when Einstein was born?" or "I found this website with lots of nice pictures for presentations.".

Each of the experimental blocks started with a relaxation phase of about 8 minutes (which is typical for stress research) in which a nature film clip was shown. Then the participants received instructions on the tasks to work on. In each block the participants were provided 2 topics to work on, and were instructed to write 2 reports, one on each topic, and make 1 presentation on one of the topics (participants could choose the topic). After each condition, the participants were asked to fill in a questionnaire on their subjective experience of stress, emotion, mental effort and task load. Between the conditions the subjects were allowed a short break and the total experiment took about 3 hours.

3.2 Collected data

The following data was captured with sensors:

- Computer interactions, via a computer logging tool
- Facial expressions, via a webcam
- Body postures, via a Kinect 3D camera
- Physiology (ECG and skin conductance), via body sensors

The raw sensor data was preprocessed and features per minute were extracted. The computer logging files were used to compute several relevant mouse, keyboard and application characteristics per minute, resulting in 12 computer interaction features. The recorded videos were analyzed with facial

expression software to determine the orientation of the head and looking direction, the amount of activation in several facial action units and the detected emotion, resulting in 40 facial expression features, averaged per minute. The 3D recordings of body posture were analyzed with the Kinect SDK to extract a skeletal model. From this model we determined joint angles between bones of the upper body, for example the angle between the upper and lower arm. Moreover, we determined bone orientations of the upper body relative to the x, y and z axis, for example the angle between the left shoulder and the up pointing y axis. From the depth image also the average distance of the user was determined. This resulted in 88 body posture features, of which 44 are minute averages on posture, and 44 are standard deviations of those minutes for body movement. The raw ECG signal was filtered and peak detection was applied to determine the heart rate and heart rate variablity (RMSSD). The skin conductance signal was averaged per minute. This resulted in 3 physiological features. Table 1 provides an overview of all features (for more detailed information see `http://cs.ru.nl/~skoldijk/SWELL-KW/Dataset.html`).

Per participant ca. 45 minutes of working under normal conditions were collected, ca. 45 minutes working with email interruptions and ca. 30 minutes working under time pressure. The feature dataset is annotated with the conditions under which the data was collected. The possibly chaotic minutes at the very beginning and very end of each condition were removed. The resulting SWELL-KW dataset contains 149 features and 2688 instances in total (on average 107 minutes data per participant). All our data is available for access by the scientific community at `https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:58624`.

Moreover, the SWELL-KW dataset includes a ground truth for subjective experience (see Table 2). This was assessed by means of validated questionnaires on

– Task load (NASA-TLX, Hart & Staveland (1988))

Table 1 SWELL-KW feature dataset. Data is preprocessed and aggregated per minute. The dataset contains 149 features and 2688 instances.

Modality (#features)	Feature type (#features)
Computer interactions (18)	Mouse (7) Keyboard (9) Applications (2)
Facial expressions (40)	Head orientation (3) Facial movements (10) Action Units (19) Emotion (8)
Body postures (88)	Distance (1) Joint angles (10) Bone orientations (3x11)) (as well as stdv of the above for amount of movement (44))
Physiology (3)	Heart rate (variability) (2) Skin conductance (1)

Table 2 Subjective experience data (3 ratings per participants). Average values for the Neutral, Interruption and Time pressure condition can be found in the last 3 columns.

Type	Feature	N	I	T
TaskLoad (NASA-TLX)	MentalDemand (0: low - 10: high)	4.9	5.4	4.9
	PhysicalDemand	1.9	2.3	2.7
	TemporalDemand	5.7	5.9	7.1
	Effort	5.2	5.9	6.1
	Performance	4.8	6.1	6.0
	Frustration	3.5	3.6	3.5
Mental Effort (RSME)	MentalEffort (0: no - 10: extreme effort)	5.5	6.5	6.3
Emotion (SAM)	Valence (1: unhappy - 9: happy)	4.8	5.7	5.3
	Arousal (1: calm - 9: excited)	3.3	3.9	4.6
	Dominance (1: submissive - 9: dominant)	5.2	6.2	5.9
Stress (VAS)	Perceived stress (0: not - 10: very stressed)	2.9	3.2	3.8

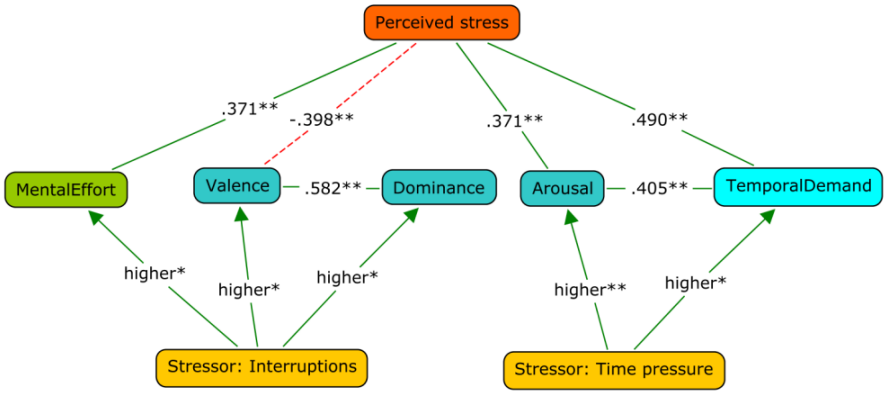

Fig. 2 Bottom part: The stressors affected several aspects of subjective experience. Top part: These aspects of subjective experience correlated with perceived stress. (* significant at the .05-level, ** significant at the .001-level.)

- Mental effort (RSME, Zijlstra & van Doorn (1985))
- Emotion (SAM, Bradley & Lang (1994))
- Perceived stress (own visual analog scale)

25 participants each rated 3 working conditions, which yielded 75 ground truth ratings in total. Note that one rating corresponds to 30-45 minutes of working behavior data. In our dataset therefore, we repeatedly annotated each row of one minute data with the ground truth of that condition. In previous work (Koldijk et al., 2013) we found that the stressor working conditions indeed affected subjective experience, see Figure 2. Therefore, we can use this dataset for user state modeling in stress related terms.

In our work we make the following assumptions about the dataset: 1) "Facial expressions, postures and physiology were reliably inferred from the raw sensor data". The data that we used here, was captured in a realistic office setting in an experimental context, which means that the quality of all record-

ings was high. Moreover, specialist equipment for capturing physiology was used.

2) "Aggregated data over 1 minute yields valuable information". There are many potential choices on how to handle the aspect of time. Here we chose to aggregate data per minute and classify each minute separately. Alternatively, different time-frames can be considered. Moreover, a model that takes into account relations between time-frames may be suitable. We started with the most simple approach.

3) "Subjective ratings provide a good ground truth". There is debate on whether subjective ratings provide a good ground truth. An alternative would be to use e.g. physiology as ground truth for stress. Here we chose to use subjective ratings, because we expected that physiological stress reactions in office setting would not to be very strong.

4) "The subjective rating given to the entire condition can be used as ground truth for each separate minute". It may be argued that stress experienced due to time pressure or incoming emails, may become stronger as the deadline comes closer or more emails have interrupted the user. Moreover, behavior may fluctuate over time. Not each minute may include signs of stress, whereas others do. Our analysis will need to show whether aggregating data per minute and classifying each minute is a good approach.

We further discuss the validity of these claims in the discussion of this paper.

4 Using several unobtrusive sensors to detect stress in office environments

In this section, we aim to answer our first main research question: Can we distinguish stressful from non-stressful working conditions, and can we estimate mental states of office workers by using several unobtrusive sensors? To distinguish stressful working conditions from neutral ones, we use *classification models* (Section 4.1). Moreover, to estimate mental states, like the amount of mental effort and stress, we use *regression models* (Section 4.2). We compare the performance of several machine learning approaches and also investigate which modalities (computer interactions, facial expressions, posture, physiology) and which particular features are most suitable.

4.1 Inferring the working condition

We first investigate whether we can distinguish stressful from non-stressful working conditions, i.e. we try to predict whether a minute of data is from the normal working condition (N, 1028 instances), or from a stressful working condition (T&I, i.e. time pressure (664 instances) or email interruptions (996 instances)).

4.1.1 Comparison of different classifiers

We start with our first subquestions: Which modeling approaches are most successful? We selected the following types of classifiers (we used their implementations in the machine learning toolkit Weka (Hall et al., 2009)):

- Nearest neighbors: IBk (uses euclidean distance, we tested the following number of neighbors: 1, 5, 10, 20), K-star (uses entropic distance)
- Bayesian approaches: Naive Bayes, Bayes Net
- Support vector machines (SVM): LibSVM (we tested the following kernels: linear, polynomial, sigmoid, radial basis function)
- Classification trees: J48 (decision tree), Random Forest
- Artificial neural networks: Multilayer perceptron

(For a comprehensive and concise explanation of the different approaches, we refer to Novak et al. (2012)). Based on the presented related work we expect that nearest neighbors, trees and SVM perform well.

As features we always used the entire SWELL-KW dataset with features on 4 modalities: computer interactions, facial expressions, physiology and postures. For nearest neighbor, SVM and neural network we first standardized the features to 0 mean 1 std (as was done in Caruana & Niculescu-Mizil (2006), for Bayes and trees scaling of the features is not necessary). We evaluated our models on accuracy, i.e. the percentage of correctly classified instances. We used 10-fold cross-validation. (This means that the data is randomly split into 10 equally large folds. 90% of the data (9 folds) is then used for training the model, the remaining fold is used for testing. This is repeated 10 times, and scores are averaged.)

The results for the different classifiers are presented in Table 3. In general, a performance of about 90% accuracy can be reached in distinguishing neutral form stressful working conditions based on sensor data, which is high.

Regarding the different classification approaches, we find the following: The *Bayesian approaches* score only somewhat above baseline (61.7600%): naive Bayes (64.7693%), and Bayes net (69.0848%). This means the data is not well-modeled in terms of chance distributions, which is what we expected. Regarding the *nearest neighbor classifiers*, KStar does not work well (65.8110%). However, IBk (which uses an euclidean distance measure) reaches a really good performance with 10 neighbors (84.5238%). Looking up nearby data points thus seems to work for this dataset, which was expected. Also as expected, *classification trees* seem to work on our data: decision tree (78.1994%), and random forest (87.0908%). The advantage of a decision tree approach is that the model is insightful. The *artificial neural network* also yields a good result: 88.5417%. However, it takes very long to train a neural network model (1 hour in our case). Finally, the best results were obtained with an *SVM* (using a radial basis function kernel): 90.0298%. That SVM perform well was also found in previous work.

Classifier	Accuracy
BASELINE Majority class: Stressful (I&T)	61.76%
Naive Bayes	64.7693%
K-star (nearest neighbor with entropic distance)	65.8110%
Bayes Net	69.0848%
J48 (decision tree)	78.1994%
IBk (nearest neighbor with euclidean distance), 10 neighbors	84.5238%
Random Forest	87.0908%
Multilayer perceptron (neural network)	88.5417%
LibSVM (support vector machine), radial basis function kernel	90.0298%

Table 3 Comparison of classifiers. Predict working conditions (N vs. I&T) from 4 modalities (Computer, Facial, Physiology, Posture).

4.1.2 Comparison of different feature sets

Now, we address our second subquestion: Which modalities/ features provide the most useful information? We continued our analyses with the SVM classifier, as this performed best.

First, we tested the following subsets: 1 modality (computer, facial, physiology or posture), a combination of 2 modalities, 3 modalities or all 4 modalities. We hypothesize that combining different modalities improves classification performance. However, for an office setting, excluding sensors to yield a minimal set-up would be preferable.

Our results are presented in Table 4. When using only a single modality, posture features yield best performance (83.4077%). When adding a second modality, facial features yield improvements (88.6905%). Only minor improvements can be reached when adding a third modality (physiology: 89.2857% or computer features: 89.1369%), and only a slightly higher accuracy is reached when using all four modalities (90.0298%). So, as expected, combining more than one modality improved performance, although the gains due to additional modalities are modest. In an office setting thus the most important modality to use would be posture, possibly combined with facial information.

We also ran feature selection in Weka[1], yielding a subset of the 26 best features for predicting the working condition. This set includes 17 posture features (13 features related to sitting posture, 4 related to body movement), 5 facial expression features (LidTightener, rightEyebrowRaised, Dimpler, looking surprised, and happy), 2 physiological features (heart rate, and skin conductance) and 2 computer interaction features (applicationChanges, and leftClicks). This confirms our hypothesis that features from different modalities can complement each other. With only these 26 features, still a performance of 84.5238% can be reached.

[1] CfsSubsetEval with BestFirst search

Features	Accuracy
BASELINE: ZeroR (majority class I&T)	61.76%
1 modality (Computer, 18 features)	65.5134%
1 modality (Physiology, 3 features)	64.0997%
1 modality (Facial, 40 features)	75.4092%
1 modality (Posture, 88 features)	**83.4077%**
2 modalities (Computer & Physiology, 21 features)	67.8943%
2 modalities (Computer & Facial, 58 features)	79.1295%
2 modalities (Facial & Physiology, 43 features)	79.9479%
2 modalities (Posture & Computer, 106 features)	83.7798%
2 modalities (Posture & Physiology, 91 features)	83.7798%
2 modalities (Posture & Facial, 128 features)	**88.6905%**
3 modalities (Computer & Facial & Physiology, 61 features)	81.2872%
3 modalities (Posture & Computer & Physiology, 109 features)	84.0402%
3 modalities (Posture & Facial & Computer, 146 features)	**89.1369%**
3 modalities (Posture & Facial & Physiology, 131 features)	**89.2857%**
4 modalities (Computer & Facial & Physiology & Posture, 149)	**90.0298%**
Only 26 best features	84.5238%

Table 4 SVM with radial basis function kernel. Comparison of using feature subsets to predict working conditions (N vs. I&T).

4.1.3 Conclusions

In this section we investigated the research question: Can we distinguish stressful from non-stressful working conditions by using several unobtrusive sensors? We found that a performance of about 90% accuracy can be reached, which is reasonably high. SVM, neural networks and random forest approaches yield the best performance. Also the rather simple nearest neighbor and decision tree approaches seem to provide reasonable accuracy. On the other side, Bayesian approaches seem less suitable for this data. With respect to the most useful modalities we find that posture yields most valuable information to distinguish stressor from non-stressor working conditions. Adding information on facial expressions can further improve performance. Computer interactions and physiology, however, showed no gains, which was unexpected given the presented related work on stress recognition based upon these modalities.

4.2 Predicting mental states

In the data collection experiment also information of subjective experience was collected after each working condition. We have information on: perceived stress, mental effort, emotion (i.e. valence, arousal, dominance), and task load (i.e. mental demand, physical demand, temporal demand, performance, effort, frustration). We address the following question now: Can we estimate mental states of office workers by using several unobtrusive sensors?

4.2.1 Predicting different subjective variables

First of all, we investigated which subjective variable can best be predicted from our sensor data. Our main aim is to infer stress from sensor data. However, other relevant subjective variables may be more directly related to the recorded sensor data.

We used Weka to train linear regression models. For evaluation, we used the correlation r between the predicted values and the true values. We also used RMSE (root mean squared error) as measure of error in the predictions. We applied 10fold cross-validation again. As features we always used the entire SWELL-KW dataset with features on 4 modalities: computer interactions, facial expressions, physiology and postures.

The results for predicting different subjective variables from sensor data are presented in Table 5. The variable mental effort can best be predicted from our sensor data, yielding a reasonably high correlation of 0.7920 and the lowest RMSE. Other related subjective variables (perceived stress, arousal, frustration, valence, task load and temporal demand) can all be predicted equally well, with a lower correlation of around .7 (and a worse RMSE). This shows that mental effort is easier to read from facial expressions, posture, computer interactions and physiology, than e.g. stress.

Prediction variable	Correlation	RMSE
Mental effort	0.7920	0.6115
Valence	0.7139	0.7024
Arousal	0.7118	0.7044
Frustration	0.7117	0.7048
Perceived stress	0.7105	0.7054
Task load	0.6923	0.7241
Temporal demand	0.6552	0.7592

Table 5 Linear regression. Predicting different subjective variables from our 4 modalities (computer, facial, physiology and posture features). Data was standardized to 0 mean 1 std, for fair comparison of RMSE.

Best features for each of the subjective variables We also ran feature selection in Weka, to see which set of features is used to predict a specific subjective variable (see Tables 6 and 7). In general, most often several facial and posture features are selected for predicting mental states, sometimes combined with a physiological feature. It is interesting to see that the algorithm selects different specific features for different subjective variables. These selected features seem to make sense: e.g. skin conductance to predict stress and frustration, or the amount of error keys to predict arousal.

4.2.2 Comparison of different regression models

As the variable mental effort seemed to be best predictable from our sensor data we focus on this subjective variable for the remainder of our analyses.

Type	MentalEffort	Stress
Facial	sad surprised scared rightEyebrowLowered gazeDirectionLeft Au06_CheekRaiser Au07_LidTightener Au10_UpperLipRaiser Au17_ChinRaiser	surprised Au05_UpperLidRaiser Au06_CheekRaiser Au15_LipCornerDepressor Au26_JawDrop Au43_EyesClosed
Posture	2 posture features (leaning, left shoulder) 1 movement feature (right elbow)	3 posture features (left shoulder, left wrist, head) 2 movement features (average body movement, right upper arm)
Computer	-	-
Physiology	-	skin conductance level

Table 6 Feature selection for the subjective variables stress and mental effort. (CfsSubsetEval with BestFirst, features were standardized to 0 mean and 1 stdv).

Type	Arousal	Frustration
Facial	angry scared zHeadOrientation rightEyeClosed leftEyebrowRaised Au06_CheekRaiser Au09_NoseWrinkler Au10_UpperLipRaiser Au25_LipsPart	(quality) gazeDirectionRight Au05_UpperLidRaiser Au09_NoseWrinkler Au14_Dimpler Au15_LipCornerDepressor Au23_LipTightener Au43_EyesClosed
Posture	6 posture features (head, shoulder center, left shoulder, left upper arm, left wrist, right upper arm) 1 movement feature (right lower arm)	4 posture features (left shoulder, left upper arm, left wrist, right upper arm) 5 movement features (leaning, left upper arm, left wrist, right upper arm, right lower arm)
Computer	nErrorKeys	nRightClicked
Physiology	heart rate variability	skin conductance level

Table 7 Feature selection for the subjective variables arousal and frustration. (CfsSubsetEval with BestFirst, features were standardized to 0 mean and 1 stdv).

We now address our first subquestion again: Which modeling approaches are most successful? We selected the following types of regression models (we used their implementations in the machine learning toolkit Weka):

- Linear regression
- Nearest neighbors: IBk (uses euclidean distance, we tested the following number of neighbors: 1, 5, 10), K-star (uses entropic distance)
- Support vector machine: SMOreg (we tested the following kernels: polynomial, radial basis function)
- Regression trees: REPTree (regression tree, i.e. sample mean at each leaf), M5P (model tree, i.e. function at each leaf)

– Artificial neural networks: Multilayer perceptron

We expected that a simple linear regression or tree approach, or the more complex SVM would work well with our data. As features we always used the entire SWELL-KW dataset with features on 4 modalities: computer interactions, facial expressions, physiology and postures.

The results on predicting mental effort from sensor data are presented in Table 8. Several models reach reasonable performance far better than baseline. The simple linear regression model reaches a comparable performance to the more complex support vector machine (correlation of 0.7920 vs. 0.7990; RMSE of 0.6115 vs. 0.6035). Furthermore, a model tree approach seems to work best with a correlation between predicted and real mental effort of 0.8221 and the lowest RMSE (0.5739). This is in line with our expectations.

Classifier	Correlation	RMSE
ZeroR (baseline)	-0.0703	1.0004
Kstar (nearest neighbor with entropic distance)	0.5875	0.9104
IBk (nearest neighbor with euclidean distance), 5 neighbors	0.7330	0.7229
REPTree (regression tree)	0.7577	0.6534
Multilayer Perceptron (neural network)	0.7763	0.7064
Linear regression	0.7920	0.6115
SMOreg (SVM), with radial basis function kernel	0.7990	0.6035
M5P (model tree)	0.8221	0.5739

Table 8 Comparison of regression models. Predict mental effort from 4 modalities (Computer, Facial, Physiology, Posture). Data was standardized to 0 mean 1 std, for fair comparison of models.)

4.2.3 Comparison of different feature sets

For its good performance and speed, we decided to continue our analyses with the model tree. We now address our second subquestion again: Which modalities/ features provide the most useful information?

We tested the following subsets: 1 modality (computer, facial, physiology or posture), a combination of 2 modalities, 3 modalities or all 4 modalities. We hypothesize that combining different modalities improves classification performance. However, for an office setting, excluding sensors to yield a minimal set-up would be preferable.

Our results are presented in Table 9. When using only a single modality, facial features yield the best performance with a correlation between predicted and true values of 0.8091. When adding a second modality, only posture features yield a slight improvement (0.8300). No real improvement is gained when adding more modalities. Contrary to our expectations, it seems best to merely use facial features, or just add posture information to predict mental effort.

We also ran feature selection in Weka[2], yielding a subset of the 25 best features for predicting mental effort with a model tree. This subset includes 10

[2] Wrapper for M5P with BestFirst search

Features	Correlation
BASELINE: ZeroR	-0.0637
1 modality (Computer, 18 features)	0.1545
1 modality (Physiology, 3 features)	0.5715
1 modality (Facial, 40 features)	**0.8091**
1 modality (Posture, 88 features)	0.5896
2 modalities (Computer & Physiology, 21 features)	0.5527
2 modalities (Computer & Facial, 58 features)	0.8027
2 modalities (Facial & Physiology, 43 features)	0.7891
2 modalities (Posture & Computer, 106 features)	0.6254
2 modalities (Posture & Physiology, 91 features)	0.7644
2 modalities (Posture & Facial, 128 features)	**0.8300**
3 modalities (Computer & Facial & Physiology, 61 features)	0.7909
3 modalities (Posture & Computer & Physiology, 109 features)	0.7718
3 modalities (Posture & Facial & Computer, 146 features)	0.8182
3 modalities (Posture, Facial, Physiology, 131 features)	0.8295
4 modalities (Computer, Facial, Physiology, Posture, 149)	**0.8309***
Only 25 best features	**0.8416**

Table 9 Decision tree (M5P). Comparison of feature subsets to predict mental effort. (* slightly differs from result in Table 8, as for the tree here non-standardized features were used.)

facial features (BrowLowerer, UpperLidRaiser, LidTightener, UpperLipRaiser, looking sad, angry, left and right eyebrows lowered, xHeadOrientation, and gazeDirectionLeft), 9 posture features (5 features related to sitting posture, and 4 related to body movement), 2 physiological features (heart rate, and skin conductance), and 4 computer interaction features (right click, double click, direction keys, and error keys). This confirms our hypothesis that features from different modalities can complement each other. With only these 25 features a performance of 0.8416 can be reached, which is slightly better than the best accuracy, which was reached with all 149 features (0.8309). Although for a real-world office setting it might be more interesting to restrict the setup to only facial expressions, which worked well as single modality (0.8091).

4.2.4 Conclusions

In this section we investigated the research question: Can we estimate mental states of office workers by using several unobtrusive sensors? First of all, we found that mental effort seems to be the variable that can be best predicted from our sensor data (better than stress, arousal, frustration, valence, task load or temporal demand). A comparison of different regression models showed that a performance of 0.8221 can be reached with a model tree, which is reasonably high. Also linear regression models, or SVMs provide good accuracy. With respect to the most useful modalities we find that facial expressions yield most valuable information to predict mental effort. Adding information on posture can slightly improve performance.

5 Taking into account individual differences

Until now, we built one generic model over all users. Now, we address our second main research question: How important are individual differences? First we investigate the role of individual differences regarding working conditions (Section 5.1), then regarding mental states (Section 5.2). Finally, we investigate the performance of models that were built for specific user groups (Section 5.3).

5.1 Individual differences regarding working condition

We start with investigating the role of individual differences in the classification problem, i.e. predicting whether a data point is from the normal working condition, or from a stressful working condition.

Participant ID as feature To investigate in how far the models benefit from participant information, we add the participant ID as feature to the dataset. Recall that a SVM predicting stressor vs. non-stressor working conditions based on all 4 modalities reached a performance of 90.0298%. When we add participant ID to the set of features, the SVM reaches a comparable performance: 89.6577%. This means that knowledge on which specific user needs to be classified, yields no valuable information to the model.

We performed feature selection to test the relative importance of the participant ID feature. We find that the participant ID feature has a very low information gain and gain ratio, and is not selected in the subset of best features for predicting the working condition.

As decision trees provide most insight, we decided to also apply this model to our data. When we inspect the built decision tree, we see that participant ID is a feature which occurs relatively late in the tree. It is thus not the case that the model builds different sub-trees for different users. However, it is the case that towards the end of a branch describing particular behavior, a split-point based on the participant can be found. So the same behavior may need to be interpreted differently, depending on the user at hand. This may indicate that different users display stress in different ways.

Test on unseen user Furthermore, we investigated how general models perform on new, unseen users. Therefore, we trained a SVM on 24 user's data and test it on a left out, unseen, user (leave-one-subject-out cross validation). What we see is a drop in performance. It differs per user how bad this drop is. Remember that 10-fold cross-validation yielded a performance of 90.0298%. When testing on an unseen user, the model reaches an average accuracy of only 58.8887%. (Recall that the baseline for our dataset was 61.76%). The worst performance was reached for participant 21, namely 37.5000%. The best performance was reached for participant 2, namely 88.3495%. The standard deviation between performances on different users was with 11.6376%-points

relatively high. Whether a model performs well on a new, unseen user may depend on the similarity of the new user to previous users.

Conclusions With respect to distinguishing stressor form non-stressor working conditions, we see that information on the particular participant does not improve classification accuracy. We also find, that the participant ID is not important enough to be selected as one of the best features for this classification task. In the decision tree, the participant ID only appears late in the branches, helping to interpret the same behavior differently for different users. When we test a generic model on an unseen user, we see a drop in performance. It differs per user how big this drop is. This may depend upon the similarity of the new user to previous users.

5.2 Individual differences regarding mental states

We now investigate the role of individual differences in the regression problem, i.e. predicting the amount of mental effort based on sensor data.

Participant ID as feature To investigate in how far the models benefit from participant information, we add the participant ID as feature to the dataset again. Recall that a decision tree predicting mental effort based on all 4 modalities reached a performance of 0.8221 (RMSE was 0.5739). When we add participant ID to the set of features, the decision tree reaches a higher performance: 0.9410 (RMSE is 0.3383). This means that knowledge on the specific user yields valuable information to the model.

We performed feature selection to test the relative importance of the participant ID feature. We find that the participant ID is selected in the subset of 13 best features for predicting mental effort (besides 9 facial expression and 3 posture features).

When we inspect the built decision tree we see that the participant ID is included in the regression formulas: for groups of users specific weights are added or subtracted.

Test on unseen user Furthermore, we also investigated how generic models perform on new, unseen users. Therefore, we trained a decision tree model on 24 user's data and test it on a left out, unseen, user (leave-one-subject-out cross validation). What we see is a drop in performance. Remember that 10-fold cross-validation yielded a performance of 0.8221 (RMSE 0.5739). When testing on an unseen user, the model reaches an average correlation of only 0.0343 (average RMSE: 1.1684). (Recall that the baseline for our dataset was a correlation of -0.0703, with an RMSE of 1.0004). Predicting the mental effort of an unseen user is thus difficult. In terms of correlation, the worst performance was reached for participant 20, namely a negative correlation of -0.4311. The best performance was reached for participant 5, namely a correlation of 0.7241. The standard deviation between performances on different users was

with 0.2800 reasonably high. Whether a model performs well on a new, unseen user may depend on the similarity of the new user to previous users.

Conclusions With respect to estimating mental states, we see that information on the particular participant improves the mental effort estimates. The participant ID is even important enough to be selected as one of the best features for this regression task. In the regression formulas we see that specific weights are added or subtracted for groups of users. A general model tested on a new user does not perform well. This suggests that especially for the task of estimating mental states it makes sense to address individual differences.

5.3 Addressing individual differences

Finally, we investigate how individual difference can be addressed. We test whether the performance of the regression models for a single modality can be improved when distinct models are made for groups of similar users.

Clustering of users In previous work (Koldijk et al., 2015) we clustered users into groups, with respect to their average level of: computer activity, facial expressions or postures. Hierarchical clustering was used to reveal the amount of clusters (k) in the data and then k-means clustering was applied. We addressed each sensor separately and found that for each sensor the users were grouped differently. This yielded, for each modality, particular groups of similar users.

Computer activity groups We found that, based on average computer activity, 2 groups of users can be discriminated: the 'writers' (16 participants (PP), many keystrokes) and the 'copy-pasters' (9 PP, much mouse activity and special keys). Recall that the performance of a decision tree with only computer activity features for predicting mental effort yielded a performance of 0.1545. When training and validating a model only on the 'writers', we find an equal correlation of 0.1668 for predicting mental effort. When training and validating a model only on the 'copy-pasters', we find a higher correlation of 0.3441.

Furthermore, we applied feature selection to find the features most predictive of mental effort for both groups. For 'writers', the best features to predict mental effort are: amount of right clicks and scrolling. For 'copy-pasters', however, the best features to predict mental effort are: the amount of dragging, shortcut keys, application and tabfocus changes, as well as the error key ratio.

Facial action unit groups We found that, based on average facial action unit activity, 3 groups of users can be discriminated: The 'not very expressive' ones (16 PP), the 'eyes wide & mouth tight' group (3 PP), and the 'tight eyes & loose mouth' group (6 PP). Recall that the performance of a decision tree with only facial features for predicting mental effort yielded a performance of 0.8091. When training and validating a model only on the 'not very expressive', we find a sightly worse correlation of 0.7892. When training and validating a model

only on the 'eyes wide & mouth tight' group, we find an equal correlation of 0.8091. When training and validating a model only on the 'tight eyes & loose mouth' group, we find a higher correlation of 0.8742.

Furthermore, we applied feature selection to find the features most predictive of mental effort for both groups. For 'not very expressive' users, the best features to predict mental effort include the action units: Dimpler and LipsPart. For 'eyes wide & mouth tight' users, the best features to predict mental effort include the action units: LidTightener, UpperLipRaiser, LipCornerPuller and ChinRaiser. For 'tight eyes & loose mouth' users, the best features to predict mental effort include the same action units, but additionally also: BrowLowerer, Dimpler, MouthStretch and EyesClosed.

Body movement groups We found that, based on average body movement, 3 groups of users can be discriminated: the group that 'sits still & moves right arm' (5 PP), the group that 'moves body a lot & wrist less' (6 PP) and the group that 'moves average' (14 PP). Recall that the performance of a decision tree with only posture features for predicting mental effort yielded a performance of 0.5896. When training and validating a model only on the group that 'sits still & moves right arm', we find a higher correlation of 0.7564. When training and validating a model only on the group that 'moves body a lot & wrist less', we find a higher correlation of 0.8488. When training and validating a model only on the group that 'moves average', we find a higher correlation of 0.6917.

Conclusions When we train models on particular subgroups of similar users, (in almost all cases) a specialized model performs equally well or better than a general model. With respect to computer activity, the model for 'writers' performs similar to a general model, whereas a model for 'copy-pasters' outperforms our general model. With respect to facial activity, the model for 'not very expressive' users performs slightly worse than a general model. However, the model for the 'eyes wide & mouth tight' group performs the same as our general model. And the model for the 'tight eyes & loose mouth' group really outperforms our general model. Finally, with respect to posture, all models for the sub-groups really outperform our general model. We also find that for different user groups, different features are selected. To apply models for subgroups of users in office settings, data of an initialization phase may be necessary to categorize a user into one of the subgroups based upon his average behavior.

6 Conclusions

In this paper we investigated different machine learning approaches to infer working conditions and mental states from a multimodal set of sensor data (computer logging, facial expressions, posture and physiology).

We addressed two methodological and applied machine learning challenges: 1) Detecting work stress using several (physically) unobtrusive sensors. We first answered the following research question: Can we distinguish stressful from non-stressful *working conditions* by using several unobtrusive sensors? We found that on our dataset a performance of about 90% accuracy can be reached. SVM, neural networks and random forest approaches work best. Also the rather simple nearest neighbor and decision tree approaches seem to provide reasonable accuracy. With respect to the most useful modalities we find that posture yields most valuable information to distinguish stressor from non-stressor working conditions. Adding information on facial expressions further improves performance.

Moreover, we answered the research question: Can we estimate *mental states* of office workers by using several unobtrusive sensors? Mental effort seems to be the variable that can be best predicted from our sensor data (better than e.g. stress). A comparison of different regression models showed that a performance of 0.8221 can be reached on our dataset. Model trees yield the best performance. Also linear regression models, or SVMs provide good accuracy. With respect to the most useful modalities we find that facial expressions yield most valuable information to predict mental effort. Adding information on posture can slightly improve performance.

Then, we addressed the second methodological and applied machine learning challenge: 2) taking into account individual differences. We first answered the research question: How important are individual differences? With respect to distinguishing stressor form non-stressor working conditions, we see that in our dataset information on the participant is not important enough to be selected as one of the best features. In the decision tree, the participant ID only appears late in the branches. When we test a generic model on an unseen user, we see a drop in performance. It differs per user how big this drop is. This may depend upon the similarity of the new user to previous users. With respect to estimating mental states, we see that information on the participant is important enough to be selected as one of the best features. In the regression formulas we see that specific weights are added or subtracted for groups of users. We further find that a general model tested on a new user does not perform well. This suggests that especially for the task of estimating mental states it makes sense to address individual differences. It should be investigated in future work why individual differences seem to play a bigger role in estimating mental states than in distinguishing neutral form stressor working conditions.

Finally, we answered the research question: Can we improve performance by building personalized models for particular user groups? When we train models on particular subgroups of similar users, (in almost all cases) a specialized model performs equally well on our dataset or better than a general model. Especially with respect to facial activity, the model for the group 'tight eyes & loose mouth' really outperforms our general model. Also, with respect to posture, all models for the sub-groups really outperform our general model. We also find that for different user groups, different features are selected. We have

to note that, a good approach to address individual differences could also be to build models for single users. However, the amount of data we had available per participant here was not enough (only 3 different subjective ratings).

To conclude, the four modalities were successfully used in an office context. Several classification and regression models were compared to find the most suitable approach. We also investigated which modalities and features were most informative. Besides applying generic models, we investigated the role of individual differences. We showed how models for subgroups of similar users can be made.

7 Discussion

Our work was based on several assumptions, on which we will comment now: 1) "Facial expressions, postures and physiology were reliably inferred from the raw sensor data". The data that we used here, was captured in a realistic office setting in an experimental context, which means that the quality of all recordings was high. In a real-world office setting recordings may be more noisy, e.g. facial expression recognition may be less reliable with bad lighting, or when the user is not positioned well in front of the camera. Moreover, specialist equipment for capturing physiology was used here. In real-world settings, devices like smart measuring watches may provide less reliable data.

2) "Aggregated data over 1 minute yields valuable information". There are many potential choices on how to handle the aspect of time. Here we chose to aggregate data per minute and classify each minute separately. Alternatively, different time-frames can be considered. Moreover, a model that takes into account relations between time-frames may be suitable. Finally, we have to note that consecutive minutes may be very similar. A random split for cross-validation may thus contain test cases that are very similar to training cases. This effect may be stronger when data of less different participants is used. On the one side, learning from similar examples is exactly what machine learning aims to do. On the other side, we have to note that such very similar minutes may cause high accuracy in evaluation.

3) "Subjective ratings provide a good ground truth". There is debate on whether subjective ratings provide a good ground truth. An alternative would be to use e.g. physiology as ground truth for stress. Here we saw that the link between physiology and subjective ratings was not particularly strong. It may be the case that physiological stress reactions in office setting are not strong enough to be reliably measured by sensors.

4) "The subjective rating given to the entire condition can be used as ground truth for each separate minute". It may be argued that stress experienced due to time pressure or incoming emails, may become stronger as the deadline comes closer or more emails have interrupted the user. Therefore, one could argue to only use the data from the last part of the condition. As we do not have too much data per participant, however, we decided to include the entire condition. Moreover, behavior may fluctuate over time. Not each

minute may include signs of stress, whereas others do. The good accuracy of the classification and regression approaches, however, indicates that not too much noise was introduced into our models in this way.

We have to note that we did all analyses on one specific dataset, the SWELL-KW dataset. Our results may be dependent on specific characteristics of this dataset. First of all, the participants' behavior is dependent on the specific tasks we gave them. This may be especially reflected in our computer interaction data: part of the interactions we record are related to the tasks of writing reports and making presentations.[3]. Note however, that the tasks themselves stayed the same for all 3 working conditions, the only thing that may have changed due to our stressors is the manner of working. Computer logging can capture, besides task related aspects, general computer interaction characteristics that change under stress, e.g. a faster typing speed or quicker window switching. These may generalize to other office working contexts, and are thus independent of our chosen tasks. Second, the specific stressors we chose, time pressure and email interruptions, may have a specific influence on the participants behavior, like a quicker work pace or more effort to concentrate on the task at hand. This may explain why stress itself was harder to predict from our sensor data. Mental effort may be more closely related to the behavior displayed under these stressors. Finally, the (behavioral) signs of stress may be intertwined with a specific way of working. An interesting question is in how far the results found in our knowledge work context hold for stress detection in general. Throughout our analyses, the facial expression features proved to be well suited. We think that facial expressions are a rather general expression of mental effort, which holds among different contexts. Moreover, posture features proved suitable. These have a little less generalizability, as the participant's posture is clearly related to the task of working behind a computer. However, it may be independent of the exact tasks that are performed. All in all, we can conclude that in future work our analyses should be applied to another dataset to prove the generalizability of the findings presented in this paper.

In general, the affective computing community often uses (black-box) machine learning algorithms to classify sensor data into mental states. In this work, we also investigated which behavior (e.g. typical facial activity, leaning forward, sitting still) that can be captured with sensors, is indicative of mental states related to stress. We see potential in building inference models that use an intermediate behavioral layer. This is in line with what Scherer et al. (2012) propose. We expect that a model with a more abstract intermediate behavior layer is more robust to individual differences and generalizes better over different users. This should be investigated in future work. In previous work (Koldijk et al., 2015), we e.g. applied a supervised Self-organizing Map (SOM) to find typical facial expressions related to high mental effort, which could be used as intermediate behavior layer. The same analyses could be applied to

[3] We also did research on automatic task recognition to investigate the manner of working. A visualization of these results can be seen here: `http://cs.ru.nl/~skoldijk/Visualization/ExperimentBrowser/Generic/Gantt_and_Numeric2.html`

posture or computer interaction data, to yield more behavioral patterns for a layered model.

As a final note, implementing such a stress detection system in real world settings brings additional challenges. Not only sensors have to be installed to collect data in the workplace, but also the signals need to be processed, features extracted and analyzed to yield meaningful data. Aspects like processing speed and privacy play an important role in practice. In their paper on body sensor networks, Fortino et al. (2013) elaborate on how a sensor system can handle such challenges. Moreover, ideally one system collects data and infers context information and then makes this available to several applications that want to make use of this context data. In their paper Fortino et al. (2014) describe different middleware systems that aim to accomplish this.

8 Acknowledgments

This publication was supported by the Dutch national program COMMIT (project P7 SWELL).

References

Andreoli, A., Gravina, R., Giannantonio, R., Pierleoni, P., & Fortino, G. (2010). Spine-hrv: A bsn-based toolkit for heart rate variability analysis in the time-domain. In *Wearable and autonomous biomedical devices and systems for smart environment* (pp. 369–389). Springer.

Bakker, J., Holenderski, L., Kocielnik, R., Pechenizkiy, M., & Sidorova, N. (2012). Stress work: From measuring stress to its understanding, prediction and handling with personalized coaching. In *Proceedings of the 2nd acm sighit international health informatics symposium* (pp. 673–678).

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49–59.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning* (pp. 161–168).

Cinaz, B., Arnrich, B., La Marca, R., & Tröster, G. (2013). Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and ubiquitous computing*, *17*(2), 229–239.

Craig, S. D., D'Mello, S., Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with autotutor: Applying the facial action coding system to cognitive–affective states during learning. *Cognition and Emotion*, *22*(5), 777–788.

Dinges, D. F., Rider, R. L., Dorrian, J., McGlinchey, E. L., Rogers, N. L., Cizman, Z., ... Metaxas, D. N. (2005). Optical computer recognition of facial expressions associated with stress induced by performance demands. *Aviation, space, and environmental medicine*, *76*(Supplement 1), B172–B182.

Epp, C., Lippold, M., & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 715–724).

Fortino, G., Giannantonio, R., Gravina, R., Kuryloski, P., & Jafari, R. (2013). Enabling effective programming and flexible management of efficient body sensor network applications. *IEEE Transactions on Human-Machine Systems*, *43*(1), 115–133.

Fortino, G., Guerrieri, A., Russo, W., & Savaglio, C. (2014). Middlewares for smart objects and smart environments: Overview and comparison. In *Internet of things based on smart objects* (pp. 1–27). Springer.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10–18.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, *52*, 139–183.

Kapoor, A., & Picard, R. W. (2005). Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual acm international conference on multimedia* (pp. 677–682).

Khan, I. A., Brinkman, W.-P., & Hierons, R. M. (2008). Towards a computer interaction-based mood measurement instrument. *Proc. PPIG2008, ISBN*, 971–978.

Koldijk, S. (2012). Automatic recognition of context and stress to support knowledge workers. In *Proceedings of the 30th european conference on cognitive ergonomics.*

Koldijk, S., Bernard, J., Ruppert, T., Kohlhammer, J., Neerincx, M., & Kraaij, W. (2015). Visual analytics ofwork behavior data - insights on individual differences. In *Proceedings of eurographics conference on visualization (eurovis).*

Koldijk, S., Sappelli, M., Neerincx, M., & Kraaij, W. (2013). Unobtrusive monitoring of knowledge workers for stress self-regulation. In *Proceedings of the 21st conference on user modeling, adaptation, and personalization.*

Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014). The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction.*

Koldijk, S., van Staalduinen, M., Raaijmakers, S., van Rooij, I., & Kraaij, W. (2011). Activity-logging for self-coaching of knowledge workers. In *2nd workshop on information access for personal media archives.*

Matthews, R., McDonald, N. J., & Trejo, L. J. (2005). Psycho-physiological sensor techniques: an overview. In *11th international conference on human computer interaction (hcii)* (pp. 22–27).

Mokhayeri, F., Akbarzadeh-T, M.-R., & Toosizadeh, S. (2011). Mental stress detection using physiological signals based on soft computing techniques. In *Biomedical engineering (icbme), 2011 18th iranian conference of* (pp. 232–237).

Neerincx, M. A., Kennedie, S., Grootjen, M., & Grootjen, F. (2009). Modeling

the cognitive task load and performance of naval operators. In *Foundations of augmented cognition. neuroergonomics and operational neuroscience* (pp. 260–269). Springer.

Novak, D., Mihelj, M., & Munih, M. (2012). A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing. *Interacting with computers*, *24*(3), 154–172.

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2007). Human computing and machine understanding of human behavior: a survey. In *Artifical intelligence for human computing* (pp. 47–71). Springer.

Riera, A., Soria-Frisch, A., Albajes-Eizagirre, A., Cipresso, P., Grau, C., Dunne, S., ... aStarlab Barcelona, S. (2012). Electro-physiological data fusion for stress detection. *Studies in health technology and informatics*, *181*, 228–32.

Scherer, S., Glodek, M., Layher, G., Schels, M., Schmidt, M., Brosch, T., ... Palm, G. (2012). A generic framework for the inference of user states in human computer interaction. *Journal on Multimodal User Interfaces*, *6*(3-4), 117–141.

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

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *Affective Computing, IEEE Transactions on*, *3*(1), 42–55.

van Drunen, A., van den Broek, E. L., Spink, A. J., & Heffelaar, T. (2009). Exploring workload and attention measurements with ulog mouse data. *Behavior research methods*, *41*(3), 868–875.

Vizer, L. M., Zhou, L., & Sears, A. (2009). Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies*, *67*(10), 870–886.

Yang, H.-K., Lee, J.-W., Lee, K.-H., Lee, Y.-J., Kim, K.-S., Choi, H.-J., & Kim, D.-J. (2008). Application for the wearable heart activity monitoring system: analysis of the autonomic function of hrv. In *2008 30th annual international conference of the ieee engineering in medicine and biology society* (pp. 1258–1261).

Zapf, D. (1993). Stress-oriented analysis of computerized office work. *The European Work and Organizational Psychologist*, *3*(2), 85–100.

Zijlstra, F., & van Doorn, L. (1985). *The construction of a scale to measure subjective effort* (Unpublished doctoral dissertation). Delft University of Technology.