Monitoring stress with a wrist device using context

Martin Gjoreski*, Hristijan Gjoreski, Mitja Luštrek, Matjaž Gams
Department of Intelligent Systems, Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova cesta 39, Ljubljana, Slovenia
{martin.gjoreski, hristijan.gjoreski, mitja.lustrek, matjaz.gams}@ijs.si

ABSTRACT

Being able to detect stress as it occurs can importantly contribute to dealing with its negative health and economic consequences. However, detecting stress in real life with an unobtrusive wrist device is a challenging task. The objective of this study is to develop a method for stress detection that can accurately, continuously and unobtrusively monitor psychological stress in real life. First, we explore the problem of stress detection using machine learning and signal processing techniques in laboratory conditions, and then we apply the extracted laboratory knowledge on real-life data. We propose a novel context-based stress-detection method. The method consists of three machine-learning components: a laboratory stress detector that is trained on laboratory data and detects short-term stress every 2 minutes; an activity recognizer that continuously recognizes the user's activity and thus provides context information; and a context-based stress detector that uses the outputs of the laboratory stress detector, activity recognizer and other contexts, in order to provide the final decision on 20-minute intervals. Experiments on 55 days of real-life data showed that the method detects (recalls) 70% of the stress events with a precision of 95%.

Keywords

Stress detection; real life; wrist device; machine learning; context; healthcare.

1. Introduction

In 1908, Yerkes and Dodson presented the Yerkes–Dodson law of empirical relationship between arousal and performance. According to the Yerkes-Dodson law, the human performs at a near-optimal level under a certain amount of stress. Consequently, stress is not necessarily a negative process, but when present continuously it can result in chronical stress. Chronical stress has negative health consequences, such as raised blood pressure, bad sleep, increased vulnerability to infections, decreased performance, and slower body recovery [1].

Work-related stress is defined as a harmful psycho-physiological response that occurs when the requirements of a job do not match the capabilities, resources or needs of a worker, which can lead to poor health and injury [2]. Regarding the economic costs of stress, the European Commission estimated the costs of work-related stress at €25 billion a year for 2013 [2]. This is because work-related stress leads to an increased number of accidents, absenteeism and decreased productivity. Therefore, having an automatic stress-monitoring system would be beneficial for the self-management of mental (and consequently physical) health of workers [3], students, and others in the stressful environment of today's world.

The three characteristics that make the problem of monitoring stress challenging and worth researching are:

- **Stress is highly subjective**. A stimulus that triggers the stress process in one person may not trigger it in another.
- **It is difficult to define the ground truth for the detection of stress.** Because of the high subjectivity and the continuous nature of the stress process, it is difficult to define the start, the duration and the intensity of a stress event.
- **Stress cannot be monitored directly.** The stress response consists of three components: physiological, behavioral and affective response [4]**.** A part of the physiological response (e.g., increased heart rate, increased sweating rate, etc.) can be monitored directly using wearable devices (e.g., Microsoft Band fitness tracker). However, there are no direct methods for monitoring the other two components (behavioral and affective response) of the stress response.

The recent technological advances brought wearable bio-sensors in the everyday life, e.g., ECG sensors [5], sweating-rate sensors [6], respiration-rate body sensors [7] etc. In our experiments, we chose a wrist device because users are mostly accustomed to wrist wearables (due to the habit of wearing watches), and it is one of the least obtrusive placements. However, the wrist is also subject to frequent movement due to the hands activity, which introduces noise in the bio-sensors' data and therefore additionally complicates the already challenging problem of stress detection.

The current state of the art studies for automatic stress detection in real life [8][9] propose a methodology using a chest sensor. In their approach, they first tune their machine-learning model in a laboratory and then apply it in real-life environments using some simplifications, e.g., they discard periods of moderate to high activity. As future work they suggest smartwatches as a source of physiological data, better handling of physical activity and including context information in the process of stress detection. In our study we tackle all of these issues by:

- Using only a wrist device as the source of physiological data.
- Recognizing the user's activity by analyzing the acceleration data from the wrist device using an award-winning machine-learning method [28].
- Using real-life contextual information in the machine-learning process to improve the performance of the method.

In addition, building upon the state-of-the-art studies, we analyze the problem of stress detection first in laboratory conditions using an off-the-shelf wrist device equipped with bio-sensors, and apply the extracted laboratory knowledge in real life, on data gathered completely in the wild. In addition to the laboratory knowledge, real-life context information is extracted for the method to be successfully applied on the real-life data. The context information is required to distinguish between psychological stress in real life and the many situations which induce a similar physiological arousal (e.g., exercise, eating, hot weather, etc.).

The proposed method is evaluated on 55 days of real-life data for 5 subjects. Real-life evaluation is poorly explored in the related work. It poses numerous problems, which are discussed in this paper. Among them are: how to gather the real-life data, how to segment the data, and how to label the data. In this study we additionally provide guidelines how to tackle these issues, which is an additional improvement compared to the related work, since the majority of the related-work methods for stress detection are tested only on laboratory data.

The rest of the paper is organized as follows: in Section 2, an overview of the related work on stress detection is presented. In Section 3, the method for stress detection in constrained environments and its evaluation are presented. In Section 4, the context-based method for stress detection in unconstrained environments and its evaluation are presented. In Section 5, practical usage of the context-based method for stress detection is presented. Finally, Section 6 summarizes the study, and presents discussion and ideas for future work.

2. Related Work

The analysis of the related work on stress detection through the prism of computer science shows that the focus shifts from stress detection in a constrained environment using less comfortable sensors to stress detection in an unconstraint environment using more comfortable sensors. The pioneers in this field are Healey and Picard who showed in 2005 that stress can be detected using physiological sensors [10]. With the advance of the technological devices equipped with physiological sensors, the method, which in 2005 required intrusive wires and electrodes, can finally be implemented comfortably.

In the period 2005-2016, various studies were conducted to implement stress detection using a combination of signal processing and machine learning (ML). Most of them used data from a respiration sensor [8][10][11], ECG sensor [8][10][11], heartrate (HR) sensor [12], acceleration sensor [13][14], electrodermal activity (EDA) sensor [8][10][11][14][15], blood volume pulse (BVP) sensor [18] and electromyogram sensor [10] [19]. Some are more constrained, either physically (e.g., brain activity analysis [20]) or with respect to privacy (e.g., analyzing the user's audio or video [21]). In our study we use a device that provides acceleration, BVP, EDA, HR, inter-beat interval (IBI), and skin temperature (ST) data.

A key difference between previous approaches in the related work is the environment for which they are intended. As with many scientific problems, the problem is first analyzed in constrained environments, e.g., a laboratory [8][17], office [16], car (analysis while driving) [10], bed (analysis while sleeping) [11], and call center [15]. One step closer to real life are Ramos et al. [13], Mohino-Herranz et al. [22] and Lu et al. [23], who presented studies in which the subjects are allowed to be active based on a predefined scenario.

Very few approaches are tested in a completely unconstrained environment. Sano et al. [14] collected 5 days of data for 18 participants using wrist-worn sensors (accelerometer and EDA) and smartphone (calls, SMS, location and screen on/off) for stress detection in real-life environments. The reported accuracy for a 2 class problem is 74% by using 10-fold cross-validation. They did not present results for person-independent models, and the wrist-worn accelerometer data is not used for distinguishing EDA caused by physical activity or stressful event, which is something that we are proposing in our study.

Adams et al. [24] collected data from seven participants as they carried out their everyday activities over a ten-day period. They used smartphone audio-sensing and a wrist-worn EDA sensor. They analyzed correlations between stress self-reports and smartphone audio-sensing. They did not use machine learning to detect stress. They concluded that context information is needed to distinguish between pleasant and negative experiences. Our proposed machine-learning method exploits context information to detect stress.

Wang et al. [25] and Bauer at al. [26] presented studies in which smartphone data was analyzed to detect behavioral changes related to stress, but they did not build models for stress detection. In our previous work, we used the data from the Wang's et al. [25] study to build machine-learning models for stress detection based only on the smartphone data. The conclusion was that only person-dependent models perform accurately enough [27].

Finally, in 2015 Hovsepian et al. [8] proposed cStress, a method for continuous stress assessment in real life, and in 2016, cStress is used in another real-life study [9]. They proved that stress can be detected using a chest belt which provides respiration and electrocardiogram (ECG) data. Building upon their guidelines for future work, we used the Empatica wrist device as the source of data, and our proven activity-recognition algorithms [28] for handling user activity and providing context information for the stress detection in real life.

This paper extends our previous short papers [29][30] which present the main idea of the method for stress detection. As a significant improvement in this paper, the method and the experimental results are described thoroughly for the first time. We also present additional and improved experimental results.

Table 1 presents a summary of the related work described in the previous subsections. The studies are grouped with respect to the environment in which they are performed (constrained – a laboratory, a car, a bus; semi-constrained – a laboratory with physical activities; unconstrained – completely in real life). Additionally we present the sensors used in the studies, the type of stressor and the number of participants. This study falls into two categories. On one hand we have experiments performed in constrained environments (in laboratory) and on the other hand we have experiments performed in unconstrained environments (real-life). The sensors used are BVP, EDA, ST, ACC, which besides the raw data also provide HR and IBI data. The stressors we analyze are a cognitive task (in the laboratory experiments) and stressors from real life. The number of participants is 21 in the laboratory experiments, and 5 in the real-life experiments. The number of participants in the real-life experiments is low, however, the overall data gathered sums up to 55 days of real-life data.

Table 1. Related-work summary.

Study	Sensors	Type of stressor	#Participants
Constrained environments			
Healey et al. [11]	Resp., ECG, EDA, EMG	Driving a car	24
Sierra et al. [12]	HR and EDA	Hyperventilation, speaking	80
Wijsman et al. [19]	Resp., ECG, EDA, and EMG	Cognitive tasks	30
Hernandez et al. [15]	EDA	Call center	9
Mellilo et al. [17]	ECG	Student exam	42
Zhai et al. [16]	EDA, BVP, PD*, ST	Cognitive tasks	32
Muaremi et al. [11]	Resp., ECG, EDA, ST	Analysis while sleeping	10
Rodrigue et al. [50]	ECG	Driving bus	36
Semi-constrained environments			
Ramos et al. [13]	Resp., HR, EDA, ST, ACC	Scenario & activity	20
Mohino-Herranz et al. [22]	ECG, TEB*	Scenario & activity	40
Lu et al. [23]	EDA, audio analysis	Scenario	14
Unconstrained environments			
Hovsepian et al. [8]	Resp., ECG, ACC	2 x Scenario, real-life	24, 26, 30
Sarker et al. [9]	Resp., ECG, ACC	Real life	38
Sano et al. [14]	EDA, ACC, smartphone	Real life	18
Adams et al. [24]	EDA and audio analysis	Real life	7
Bauer et al. [26]	Smartphone	Real life, student exams	7
Wang et al. [25]	Smartphone	Real life, student exams	48
Gjoreski et al. [27]	Smartphone	Real life, student exams	48
Gjoreski et al. [29][30]	BVP, EDA, ST, ACC	Cognitive tasks, real life	21, 5

* TEB – Thoracic electrical bioimpedance. *PD – pupil diameter

3. Stress detection in constrained environments

The term "constrained environments" implies that the method for stress detection is developed using constraints, e.g., stress detection while driving [10], stress detection while sleeping) [11], stress detection in a laboratory [19], etc. These constraints significantly simplify the detection by discarding real-life situations which induce a response of the human body similar to the stress response (e.g., physical exercise, eating, hot weather, etc.). Due to the simplifications, the use of these methods is limited to the environment for which they are developed. However, besides the limited use, developing a method in constrained conditions allows for detailed analysis of the stress response. For this reason, we performed laboratory experiments in our study.

In the next subsections, first the laboratory data is described, then the method for stress detection and finally the experimental results are presented.

3.1 Laboratory Experimental Setup

For collecting the laboratory data we used a standardized stress-inducing experiment [31]. Additionally, baseline (no-stress) data was recorded on a separate day when subjects were relaxed. For the stress-inducing experiments, a web application was developed in collaboration with psychologists. The application implements a variation of the stress-inducing method presented by Dedovic et al. [31]. The main stressor is solving a mental arithmetic task under time and evaluation pressure. In short, a series of randomly generated equations were presented to subjects, who provide answers verbally. The time given per equation was dynamically changing. For each two consecutive correct answers the time was shortened by 10%, and for each two consecutive wrong answers the time was increased by 10%. Each session consisted of three series of equations with increasing difficulty: easy, medium and hard. Each series of equations lasted for five minutes. For motivation, a reward was promised to the top three participants. After each stage, the participant was shown a false ranking score, positioning him/her in the top five, this way motivating him/her to try harder in the next stage and try to win the reward. The application is available on-line: http://dis.ijs.si/thestest/.

The experiments were organized with respect to the four pre-conditions for a situation to induce a stress response as presented by Lupien et al. [32]. These are:

- **Novel** for the subjects because none of them had attended similar experiments.
- **Unpredictable** because none of them knew how exactly the experiments were organized.
- **Not controllable** because the subjects had to follow strict instructions which were given while the experiments were being executed.
- **A social evaluative threat** because rankings were available on-line and the subjects were competing against each other.

During the experiment, there were no movement constraints, making it as close as possible to real-life sedentary situations.

3.2 Statistical analysis and labelling of the laboratory data

Four Short STAI-Y anxiety questionnaires [26] were filled by each participant: before the experiment (1), and after the easy (2), medium (3) and hard session (4). The mean STAI score is presented in Table 2. We performed statistical analysis using repeated measures ANOVA [33]. The resulting p-value was 0.0014, confirming that there is a statistically significant difference between the answers of the different stages (before, easy, medium and hard stage).

Table 2. Mean STAI score for the laboratory data.

Type	Before	Easy	Medium	Hard
STAI score	10.95	13.33	14.05	13.81

The answers of the STAI questionnaire were used for subject-specific labelling of the data. For each subject, the period before answering the STAI questionnaire in which they achieved the lowest score is labelled as low stress, and for each +3 STAI points (the statistical tests showed that difference of 2.38 is enough), the stress label is increased by one, thus we get no stress (baseline data), low stress (lowest STAI score), medium stress (lowest STAI score +3) and high stress (lowest STAI score +6). In the final experiments the medium and high stress were merged because only two subjects achieved a high level of stress, so we had three degrees of stress: no stress, low and high. Table 3 presents an overview of the labelled laboratory data.

Table 3. Laboratory data overview. The number of participants, age (mean and standard deviation) and duration in minutes for the three levels (No Stress, Low Stress and High stress).

	Data
# Participants	21
Age	28±4
No Stress - overall duration per sensor	840 minutes
Low Stress - overall duration per sensor	356 minutes
High Stress - overall duration per sensor	368 minutes

3.3 Machine-learning method for stress detection in constrained environments

For the creation of the laboratory stress-detection classifier, we used the machine-learning pipeline presented in Figure 1. First, the data was stored locally on the Empatica device, then transferred to a computer where the rest of the processing was performed. The method contains six phases: segmentation, filtering, feature extraction, feature selection, model learning and evaluation of the models. In the next subsections, each stage is described in detail.

The segmentation refers to the segmentation of the data into windows used for extracting the features. We experimented with window lengths from 30 s to 300 s. The experimental results are presented in Subsection 3.3.7.

The filtering methods are specific for each data type (e.g., the BVP signal requires a different filtering technique than the EDA signal), and are described in Subsections 3.3.1 – 3.3.6 together with the feature-extraction phase.

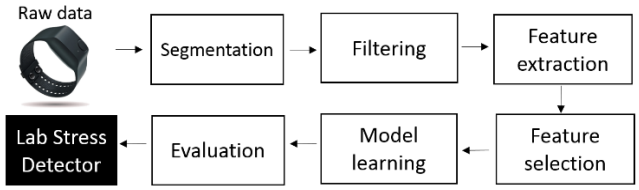

Figure 1. The method for learning laboratory the stress detector.

In the feature extraction phase, numerical features are extracted from each data window using statistical functions, regression analysis and frequency and time analysis, depending on the type of the signal.

3.3.1 Feature extraction from the BVP data

The Empatica device provides BVP data extracted from a PPG (Photoplethysmography) sensor for which they use a proprietary algorithm. The measurement unit is a fraction of nanoWatt (nW) which represents the difference of light absorption observed by a light receiver in the PPG sensor. The sampling frequency is 64 Hz. Figure 2 shows an example 5 seconds of clean (black) and noisy (grey) BVP data provided by the Empatica device. The clean data was collected during a low-movement period (probably sleeping). It can be clearly noticed that it is a periodical signal, which represents the activity of the heart. The Empatica device exploits this periodic feature to extract the duration between heartbeats. However, the noisy data is BVP provided during a high-movement period. These signals present the problem when analyzing data from wearable devices: the noise in the data and the disturbances from the environment. For this reason, prior to the feature extraction phase, the BVP signal is filtered using winsorization method [34]. Winsorization is a statistical method for removing the outlier values over the n^{th} and $100 - n^{th}$ percentile (experimentally n was set to 2). From the filtered signal, statistical features were computed: standard deviation, 20^{th} percentile, 80^{th} percentile and quartile deviation (75^{th} percentile – 25^{th} percentile).

3.3.2 Feature extraction from the HR data

The Empatica device provides the average heart rate (extracted from the BVP signal) with a sampling frequency of 1 Hz. Figure 3 presents example HR data provided by the device. The HR signal is already filtered by the device. From this signal the following features were calculated: mean, standard deviation, percentiles, quartile deviation (75^{th} percentile – 25 percentile), and slope and intercept of a fitted regression line.

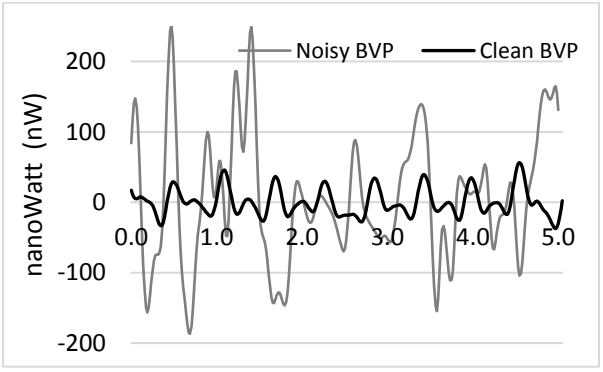

Figure 2. Clean and noisy BVP signal from Subject 5.

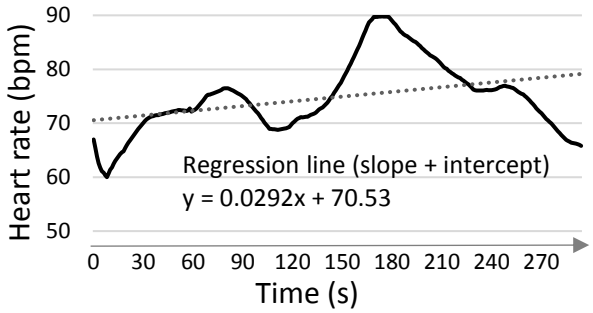

Figure 3. HR signal provided by Empatica for Subject 5.

3.3.3 Feature extraction from the IBI data

The Empatica device provides the time between individual heartbeats, IBI data, extracted from the BVP data. This data contains a timestamp and duration of the detected heartbeats. For detection they are a using proprietary algorithm, which provides IBI data only in moderate to low-movement periods. Since the Empatica device does not provide IBI data in high-movement periods, the IBI data is not continuous, thus it requires additional filtering and segmentation.

To perform the frequency analysis, continuous stream of IBI data points are detected in each data window. If the stream is not continuous (e.g., the time between two IBI data points is higher than 2s), the data is disregarded. This introduces a tradeoff between the number of missing values for the features calculated using frequency analysis (the longer the continuous stream of neighboring intervals has to be, the longer the person needs to be still to gather the data), and the quality of the spectrum used for the frequency analysis (the longer the stream, the better the resolution of the power spectrum). We used a window of 32 continuous samples, which – for an average heart rate of 64 bpm – requires the person to be still for around 30 seconds. Since the data window can be bigger than 30 s, the power spectrum is calculated as the average of the power spectrums over all continuous IBI samples in one data window. Figure 4 presents a power spectrum of the IBI samples for one data window of 300 seconds. The frequency domain features were the total spectral power of all IBI samples in power bands up to 0.04 Hz, between 0.003 and 0.04 Hz, between 0.04 and 0.15 Hz, and between 0.15 and 0.4 Hz, and the ratio of low (0.04 Hz – 0.15 Hz) to high (0.15 Hz – 0.4 Hz) frequency power.

The time-domain features were the mean of the IBI samples, the standard deviation of the IBI samples, the square root of the mean of the squares of the differences between adjacent IBI samples, and the percentage of the differences between adjacent IBI samples that are greater than x ms (x = 20, 50, 70).

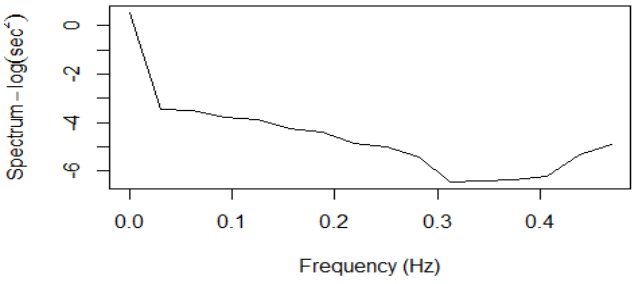

Figure 4. Power spectrum of the IBI data in a data window of 300 seconds for Subject 5.

3.3.4 Feature extraction from the EDA data

Electrodermal activity refers to electrical changes that arise when the skin receives specific signals from the brain. These changes may be due to emotional activation, cognitive workload or physical exertion. The electrical change is enough for the EDA sensor to capture it. The Empatica E4 device captures the change in the electrical conductance using two electrodes attached to the wristband. By flowing a minuscule amount of current between the electrodes, the device measures the conductivity on the wrist. The data unit is micro Siemens (μS) and the frequency is 4 Hz.

From the EDA signal the following features were calculated: mean, standard deviation, 20th percentile, 80th percentile, quartile deviation (75th percentile – 25 percentile), and the slope and intercept of a fitted regression line. Additionally, an algorithm for peak detection [35] was used to detect the EDA responses – peaks in the EDA signals. This enabled additional features: the number of responses, the power of responses, the number of significant responses (responses which have a value over some 1.5 μS) and the power of significant responses.

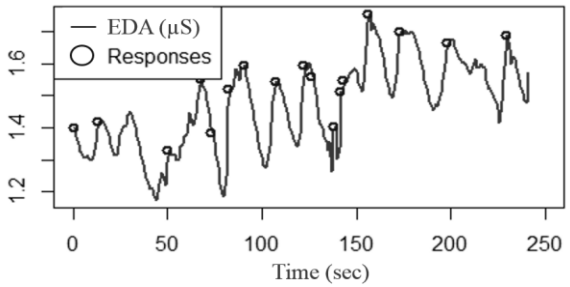

Figure 5. EDA signal and detected responses for Subject 5.

3.3.5 Feature extraction from the ST data

The "fight-or-flight" response restricts the blood flow from the extremities and increases the blood flow to the vital organs. This peripheral vasoconstriction produces changes in the skin temperature on the extremities including hands, which can indicate stress and its intensity [36]. The Empatica device provides peripheral skin temperature data with a frequency of 4 Hz. From the ST signal we extracted the features: mean temperature, the slope and intercept of a fitted regression line.

3.3.6 Feature extraction from the acceleration data

The Empatica device provides 3-axis accelerometer data with a sampling frequency of 32 Hz. This data was only used in the real-life experiments for recognizing the subject's activities ("lying", "sitting", "standing", "walking", "running" and "cycling"). The data was left out in the laboratory experiments because the subjects had to sit in front of a computer, whereas during the "no-stress" scenario, the subjects were allowed to behave normally (most of them were working on a computer and moving around freely). The difference in the scenarios/activities may be reflected in the acceleration data.

3.3.7 Feature selection

The idea of the feature-selection method is to remove correlated and "non-informative" features. Correlation between the features is expected because several features are extracted from one data source (e.g., several features are extracted from the BVP signal, several from the HR signal, etc.). Additionally, non-informative features are considered those that have a low information gain. The information gain metric evaluates the worth of a feature by measuring the information they carry about the class [37].

Algorithm 1: Feature selection method

```
Input: Instances, ML algorithm
Result: Best performing feature-set
currentFeatureSet←getFeatures(Instances)
bestFeatureSet←currentFeatureSet
bestPerformance ← evaluate(Instances, ML algorithm, currentFeatureSet)
rankings←calculateInfoGain(Instances)
sortedRankings←sort(Rankings)                    // rank features by increasing info-gain
size←size(sortedRankings*0.9)                    // check the lowest 90%
counter←0
while counter< size do
    currentFeature← sortedRankings[counter]
    correlationCoefficients←pearsonsCorrelation(currentFeatureSet, currentFeature)
    for each correlationCoefficient in correlationCoefficients:
        if correlationCoefficient > 0.8          // remove the feature if there is a strong correlation
            currentFeatureSet←removeFeature(currentFeatureSet, currentFeature)
            currentPerformance ← evaluate(Instances, ML algorithm, currentFeatureSet)
            if currentPerformance >= bestPerformance
                bestPerformance ← currentPerformance
                bestFeatureSet← currentFeatureSet    // save the best performing feature set
            break
    counter++
end
return bestFeatureSet.
```

The proposed method includes several main steps: rank features by the information gain, calculate the correlation coefficients between features, evaluate different subsets of features (selected based on information-gain rankings and correlation coefficients) using leave-one-subject-out (LOSO) technique, and finally provide the best performing feature set. The pseudocode for the feature selection method is provided in Algorithm 1**.**

3.3.8 Model learning and evaluation

For the model learning we used machine-learning algorithms as implemented in the WEKA machine-learning toolkit. We experimented with a variety of ML algorithms:

- **Majority classifier** – always predicts the majority label. This is a baseline model.
- **J48** – an algorithm for building a decision tree (DT) [38].
- **Naïve Bayes** – an algorithm for building a simple probabilistic classifier based on the Bayes' theorem with strong independence assumptions between the features [39].
- **KNN** – an algorithm which provides a prediction based on k training instances nearest to the test instance [40].
- **SVM** – an algorithm for building a classifier where the classification function is a hyperplane in the feature space [41].
- **Bagging** – an ensemble algorithm which learns base models on subsets of the training data with the purpose of reducing variance and avoiding overfitting [42].
- **Boosting** – an ensemble algorithm which learns models on subsets of the training data and "boosts" the weights of misrecognized instances allowing for the models in the ensemble to focus on the misclassified instances [43].
- **Random Forest** – an ensemble algorithm which learns base decision trees by sub-setting the feature set [44].
- **Ensemble Selection** – an ensemble algorithm for combining ML models built with various ML algorithms (e.g., in our experiments we used a combination of J48, KNN, Naïve Bayes, Boosting, SVM, and RF [45]).

For model evaluation we LOSO cross validation technique. The models are learned from all the data except the data of one subject, which is used as the test data. This procedure is repeated for each subject in the dataset (21 in the laboratory dataset), and the results are averaged. The LOSO cross-validation provides information about the performance of the models on data from a new subject, who has not been included in the training. Thus, it evaluates the generalization performance of the model.

The evaluation metrics used for comparing the models are: accuracy (1), precision (2), recall (3) and F1 score (4). The accuracy provides information about the percentage of instances that were classified correctly by the model. The precision and the recall are metrics that provide label-specific information. The precision provides information about how many instances the model classified correctly when predicting the label X (where X can be "no stress", "low stress", and "high stress"). The recall provides information about how many instances out of all instances labeled with the label X were correctly classified by the model. The F1 score combines precision and recall, since the precision and recall provide different information (e.g., one model can have a high precision and a low recall for the label "low stress" if it classifies each instance as "low stress").

$$(1)\ \text{Accuracy} = \frac{\text{Correctly predicted stress level}}{\text{Total number of instances}}$$

$$(2)\text{Precision} = \frac{\text{Correctly predicted stress level X}}{\text{Total predictions of level X}}, \text{X} = (0,1,2)$$

$$(3)\ \text{Recall} = \frac{\text{Correctly predicted stress level X}}{\text{Total instances of level X}}, \text{X} = (0,1,2)$$

$$(4)\ \text{F1 score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

3.4 Laboratory experimental results

On the laboratory dataset we performed two types of experiments, windowing experiments and feature-selection experiments. The windowing experiments provide a performance comparison between the models by varying the size of the data window used for the feature extraction. The feature-selection experiments provide performance comparison between the models using different subsets of the extracted features.

3.4.1 Windowing experiments

For the windowing experiments we used the data from all sensors (except the accelerometers) and all extracted features. The experiments were performed with a varying data-window size. We started the experiments with a data-window size of 30 seconds and increased it up to 360 seconds (6 minutes, which was the duration of one session in the laboratory experiments) in increments of 30 seconds. The overlap between the data windows was set to the size of the data window decreased by 25 seconds, which is maximum overlap size with respect to the minimum window size of 30 seconds.

presents the results for the experiments with varying data-window size. The rows represent the data-window size and the columns represent the accuracy of the models. It can be seen that all of the models except those built with Bagging achieve a higher accuracy for bigger data-window size. In addition, the best performing algorithm is SVM for each data-window size.

3.4.2 Feature-selection experiments

For the feature-selection experiments we compared the performance of the SVM model using subsets of features. The subsets were selected on a sensor-specific base (BVP, ST, EDA, HR and IBI), one feature set is generated for the PPG sensor which includes BVP, HR and IBI data, and one additional feature set was generated using the feature-selection algorithm described in Section 3.3.7.

Figure 6 presents the results for the feature selection experiments. Each bar represents different feature set used for the experiments. The best performance is achieved when the algorithm uses the combination of all sensors. When sensor-specific features are used, the IBI and HR sensor data perform slightly better than the other sensor-specific feature sets.

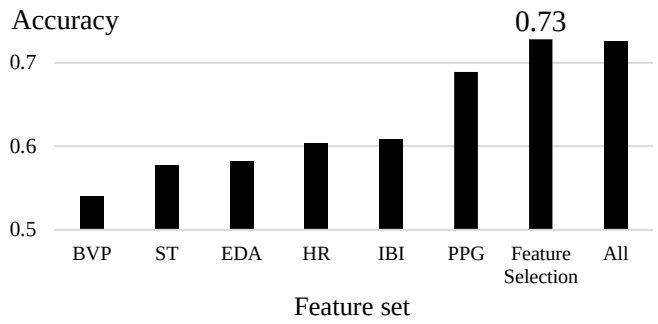

Figure 6. Accuracy for different feature sets using LOSO evaluation.

In addition, we present the confusion matrix for the FS-SVM model built using the selected features and a data window of 5 minutes with 2.5 minutes overlap. The confusion matrix in Table 5 presents the results for a 3-class problem ("No Stress" vs. "Low Stress" vs. "High Stress"). It can be seen that the "Low stress" is almost equally confused with "No stress" and "High stress". This is expected since the data is analyzed as a continuous stream using a sliding window of 5 minutes with 2.5 minutes overlap, so two neighboring data windows with (possibly) different labels always have 50% equal data. Additionally, it is almost impossible to define a strict border between different stress events.

Table 5. Confusion matrix for the FS-SVM model using LOSO evaluation. Each number represents an instance with a data-window duration of 5 minutes.

	No Stress	**Low Stress**	**High Stress**
No Stress	308	15	14
Low Stress	33	68	40
High Stress	28	38	73
Precision	0.84	0.56	0.58
Recall	0.91	0.48	0.53
F1 score	0.87	0.52	0.55
Accuracy		0.73	

Table 4. Accuracy on the laboratory data for varying data-window size and varying ML algorithms using LOSO evaluation.

Data-window size in minutes	SVM	Random Forest	Boosting	Bagging	KNN	Naïve Bayes	Decision Tree	Ensemble Selection	Majority
0.5	0.67	0.60	0.59	0.62	0.53	0.59	0.52	0.53	0.55
1.0	0.67	0.62	0.56	0.59	0.53	0.60	0.58	0.53	0.55
1.5	0.67	0.63	0.60	0.59	0.52	0.61	0.59	0.53	0.55
2.0	0.67	0.65	0.61	0.61	0.53	0.61	0.55	0.53	0.55
2.5	0.67	0.66	0.61	0.61	0.55	0.62	0.56	0.53	0.55
3.0	0.67	0.65	0.62	0.65	0.56	0.62	0.56	0.54	0.55
3.5	0.68	0.67	0.62	0.65	0.56	0.62	0.56	0.54	0.55
4.0	0.68	0.66	0.64	0.64	0.58	0.62	0.59	0.54	0.55
4.5	0.69	0.66	0.64	0.64	0.63	0.62	0.53	0.54	0.55
5.0	0.70	0.68	0.67	0.61	0.64	0.62	0.59	0.54	0.55
5.5	0.71	0.69	0.66	0.61	0.64	0.62	0.58	0.55	0.55
6.0	0.71	0.69	0.66	0.58	0.65	0.62	0.58	0.55	0.55

4. Stress Detection in Unconstrained Environments

4.1 Real-life experimental setup

For collecting real-life data we used a combination of a stress log and Ecological Momentary Assessment (EMA) prompts implemented on a smartphone. The EMA prompts are questionnaires displayed at a random time of the day. The subjects had to answer 4–6 EMA prompts per day (with at least 2 hours between consecutive prompts), and in the case of a stressful situation, they logged the start, the duration and the level of stress on a scale from 1 to 5 (1 to 2 - no stress, 3 to 5 - stress). The answers of the EMA prompts and the stress log were used to label the real-life data. Table 6 presents an overview of the real-life data. The first two rows present the number of the participants and the age structure. The final two rows present the duration of the data in minutes after labeling it with the corresponding label, "No Stress" or "Stress".

Table 6. Real-life data overview. Participant information and duration of labeled data for No Stress / Stress.

	Data
# Participants	5
Age Mean	28±4.3
No Stress - overall duration per sensor	1216 hours
Stress - overall duration per sensor	111 hours

To evaluate the method on the real-life data we had to address the well-known problem of subjective stress labeling [8]. In addition to the perception of stress being subjective, a time lag is often a problem. For example, the user marked that a stressful event occurred from 14:00 to 15:00, but this happened to be a scheduled exam, and the physiological arousal (which the sensors capture) started at 13:00. So, if we run the laboratory stress-detection classifier, it would start to predict stress at 13:00 (which is correct), but the labels of the data would say that the stress event started at 14:00. This also goes the other way around – users may mark that a stressful situation started before it actually did when labelling retroactively.

Figure 7 depicts two scenarios for splitting the real-life data into events. In scenario 1, the user answered an EMA questionnaire at time X with a stress level higher than 1. The period X – 10 minutes to X + 10 minutes is labeled as stress and one event is created from this data. In scenario 2, the user logged a stressful situation that started at time Y and ended at time Z. The period Y – 10 minutes to Z + 10 minutes is labeled as stress and one event is created from this data. In both scenarios, the rest of the data is split into events with a duration of 10 minutes and labeled as no stress. There is no information about the duration of the events in the features used by the context-based classifier, thus the event splitting does not implicitly indicate the type of the event.

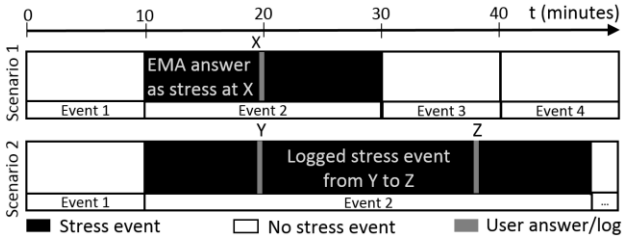

Figure 7. Splitting the real-life data into events.

4.2 Method for Stress Detection in Unconstrained Environments

The proposed context-based method is presented in Figure 8. The method consists of three main ML components: the laboratory stress detector, an activity-recognition classifier and a context-based stress detector. The following subsections explain each ML component in detail.

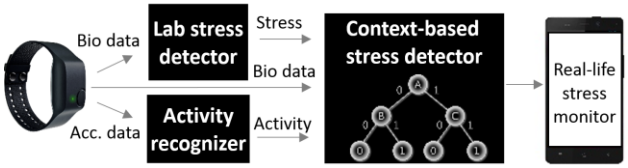

Figure 8. Context-based method for stress detection in unconstrained environments (real life).

4.2.1 Laboratory stress detector

The details about the laboratory-stress detector are provided in Section 3.3, where we explain the complete approach for stress detection in laboratory conditions (constrained environments). The laboratory-stress detector is built on the laboratory data and uses a data window of 5 minutes with 50% overlap (thus it provides a prediction every 2.5 minutes). The output of the laboratory-stress detector is provided as an input to the context-based stress detector.

4.2.2 Activity-recognition classifier

It is important for a stress-detection system to be aware of the user's physical activity, since physical activity elicits physiological arousal similar to the physiological arousal elicited by psychological stress. For this purpose, the Empatica device provides acceleration data, which has proven to be successful for recognizing activities [28][46]. A detailed description of the presented ML approach to activity recognition can be found in one of our previous papers [47]. The method is based on the award-wining approach from the EvAAL activity recognition competition [28].

The activity-recognition classifier outputs an activity every 2 seconds. The outputs are aggregated over the data window of 5 minutes, by changing each to "an activity level" (lying = 1, sitting = 2, standing = 3, walking = 4, running/cycling = 5) and averaging over the window. The average activity level is passed as a feature to the context-based stress detector.

Even though there are a lot of other activities in real life, the six activities that the activity recognizer provides are enough to represent the user's activity level because the predictions are averaged over a period of 5 minutes. In more active periods the predictions with higher activity level will predominate, and in less active periods the activities with lower activity level will predominate.

4.2.3 Context-based stress detector

The context-based stress detector was developed to distinguish between psychological stress in real life and many situations which induce a similar physiological arousal (e.g., exercise, eating, hot weather, etc.). It classifies every 10 minutes as stressful or non-stressful.

As features, it uses statistical functions (mean, max and average) over the 4 outputs of the laboratory stress detector, the previous output of the context-based detector and the 20th percentile of each sensor data (HR, BVP, IBI, ST and EDA signal). The 20th percentile was used to provide some information from each

sensor to the context-based classifier. In addition we added the features mostly used in the related work for stress detection – standard deviation of the IBI samples and features based on the EDA peak analysis.

The context features are: whether there was any high-intensity activity in the last 30 minutes, whether there was any medium-intensity activity in the last 20 minutes, the hour of the day, and the type of the day – workday/weekend.

4.3 Real-life experimental results

On the real-life data we performed two types of experiments, aggregation experiments and context vs. no-context experiments. The aggregation experiments provide information about the influence of the size of the aggregating window used for extracting features (contexts) on the performance of the context-based classifier. The context vs. no-context experiments provide a comparison between the context-based and a no-context approach, where the no-context method is the laboratory stress detector applied directly on the real-life data. For both experiments we used LOSO evaluation and the evaluation metrics described in Section 3.3.8.

4.3.1 Aggregation experiments

For the aggregation experiments we varied the size of the aggregating window from 10 minutes to 30 minutes, and monitored the performance of several machine-learning algorithms. The results are presented in Table 7. The rows represent the size of the aggregating window, and the columns represent the mean F-score (mean value of F-score for "no stress" and F-score for "stress") for each of the algorithms. In general, the algorithms perform better for a smaller aggregation window (10 to 17.5 minutes). This may be because the context (e.g., the activity of the user) is changing in a time-span of 10-15 minutes. The best performing algorithm is the Decision Tree which is interesting since it is also one of the simplest algorithms in the experiments.

4.3.2 Context vs. No-context

In these experiments we took the best-performing model from the aggregating experiments (DT with an aggregating window of 10 minutes) and compared its performance to a no-context classifier.. The results are presented in Table 8. It can be seen that the context-based classifier performs significantly better than the no-context classifier. For example, the context-based classifier achieves a mean F-score of 0.9 (mean value of 0.99 and 0.81) and the no-context classifier achieves a mean F-score of 0.47.

Additionally, the confusion matrix "with context" in shows that the Precision (95%) of the model is higher than the Recall (70%) by 25 percentage points. This means that the model detects (recalls) 70% of the stress events with a precision of 95%.

Table 8. Confusion matrices and performance measures for Context vs no-context approach.

	No-Context		With Context	
	No Stress	**Stress**	**No Stress**	**Stress**
No Stress	3308	1630	4932	6
Stress	34	125	47	112
Precision	0.99	0.07	0.99	0.95
Recall	0.67	0.79	1.00	0.70
F1 score	0.80	0.13	0.99	0.81
Mean F1 score	0.47		0.9	

To explore the link between the precision and recall we present the precision-recall curve in Figure 9. A precision-recall curve shows model's performance for a varying decision threshold. A low decision threshold means that the model would classify each instance as "stress", thus it would have a high recall, but also a low precision because all the "no stress" instances would be classified as "stress". On the contrary, a high decision threshold leads to a conservative model, meaning it would classify an instance as "stress" only when it is completely sure.

The precision-recall curve in Figure 9 shows that a recall higher than 70% is achieved only when the model's precision is lower than 60%. The highest achieved performance by the model is: precision 98% and recall 70%

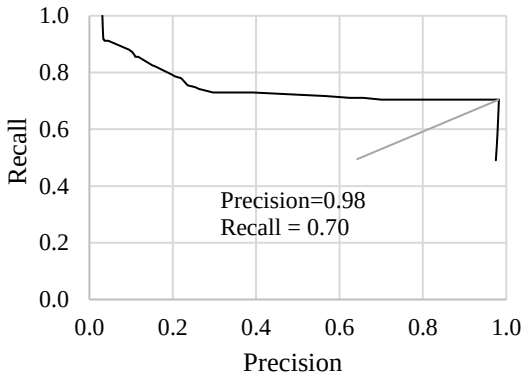

Figure 9. Precision-recall curve for the best performing model.

Table 7. Mean F-score on the real-life data for varying aggregation-window size and varying ML algorithms using LOSO evaluation.

Aggregation-window size in seconds	Decision Tree	Ensemble Selection	Random Forest	Bagging	SVM	KNN	Naïve Bayes	Boosting	Majority
10	0.90	0.80	0.74	0.78	0.69	0.69	0.58	0.51	0.49
12.5	0.85	0.75	0.73	0.76	0.71	0.71	0.60	0.49	0.49
15	0.84	0.74	0.75	0.75	0.67	0.68	0.60	0.49	0.49
17.5	0.86	0.74	0.78	0.77	0.70	0.70	0.58	0.48	0.48
20	0.73	0.69	0.71	0.69	0.68	0.68	0.59	0.51	0.48
22.5	0.72	0.62	0.65	0.61	0.57	0.64	0.57	0.51	0.48
25	0.68	0.63	0.62	0.59	0.61	0.65	0.60	0.48	0.48
27.5	0.68	0.62	0.66	0.60	0.55	0.63	0.59	0.54	0.48

5. Visualization of stress events

In the previous sections we overviewed, proposed and evaluated machine learning methods for stress detection in laboratory and in real life. However, in order for the methods to be useful, they should be integrated in a larger system that helps people to overcome stressful situations. Such a system can help users to overview stressful events, provide health advice and suggest stress-relieving exercises upon detected stress. In the next section we will describe how our context-based method for stress detection can be utilized in an e-health system.

Figure 10 presents a visualization of the output of the context-based stress detection method. The data is from the subject that had the biggest amount of data, Subject 2. On the x-axis is the day on which the data is collected, on the y-axis is the hour of the day, and the color corresponds to the intensity of stress. Thus, each square represents the stress intensity of the corresponding hour in one day. The decision whether there is stress is provided by the context-based model, and the intensity is calculated using the predictions of the laboratory model (since the context-based model is a binary classifier). This type of visualization provides information about stressful patterns throughout the day.

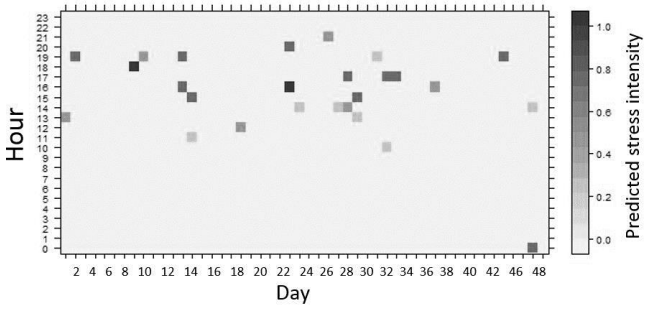

Figure 10. Daily level of stress for the subject with biggest amount of data, Subject 2. Each square represents the stress intensity on a scale from 0 to 1 for the corresponding hour in one day.

Figure 11 presents another visualization of the output of the context-based stress detection. On the x-axis is the hour of the day and on the y-axis is the stress level. Each line represents one subjects marked as S1–S5. For example for Subject 2 (S2), it can be seen that the subject's stressful events are between 13:00 and 20:00. The subject commented that the figure helped him to recall that the "stressful" hours are usually his late-working hours and over-time working hours. To reduce the stressful events he may start going to work earlier and stop working overtime.

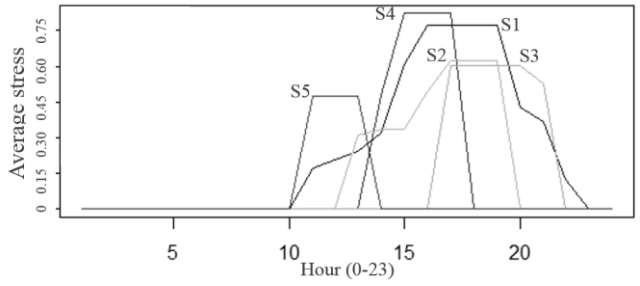

Figure 11. Average level of stress per hour for each of the five subjects.

6. Conclusion

The objective of this study was to develop a method for stress detection which can accurately, continuously and unobtrusively monitor psychological stress in real life.

At the beginning, three challenges were identified: (i) subjectivity, (ii) fuzzy ground truth and (iii) no methods for direct monitoring of stress. For addressing the subjectivity and the fuzzy ground truth we used standardized scenarios for inducing stress, standardized questionnaire (STAI), and in the real-life experiments we used EMA in combination with a stress log. In addition, we proposed a segmentation method for splitting the real-life data into stress / no stress events. The third challenge (monitoring the other components of the stress response) will be addressed in future work.

Having these challenges in mind, the problem of stress detection was first analyzed in laboratory conditions using off-the-shelf wrist device equipped with bio-sensors, and the extracted laboratory knowledge was applied on real-life data. In real life, the laboratory-stress detector achieved a mean F-score of 0.47 and a precision of 7% for detecting stress events, which is not acceptable in reality. However, when additional context information was added, the context-based method achieved a mean F-score of 0.9 and a precision of 95%. These results are significantly better and indicate possible use in real life. The context information was required to distinguish between psychological stress in real life and the many situations which induce a similar physiological arousal (e.g., exercise, eating, hot weather, etc.). Adding a context information to the stress detection system is a novel idea which significantly improved the performance of the system on the real-life data.

Our system consist of a wrist device and a processing unit which may be a smartphone, tablet, PC, etc. The Empatica wrist device provides BVP, EDA, ST, ACC, HR and IBI data, and costs around 1000€. However, at the time of performing the experiments, it was the only wrist device that fulfilled the requirements. Nowadays, there are cheaper devices that provide similar data, e.g., Microsoft Band. We are currently running experiments with the Microsoft Band in order to re-evaluate the method. Due to the fast progress of the electronics, one might expect additional devices that would be capable of providing the sensory data required by the method.

6.1 Limitations

The three main limitations of the method are:

- **Sample size.** Even though the proposed context-based method for stress detection was tested on 55 days of real-life data, this data belongs to only to 5 subjects. To confirm the obtained results we need a bigger population.
- **Age structure.** The proposed stress detection method is highly dependent on physiological signals that depend on age, sex and physical fitness. However, the experimental data in our study, both the laboratory and real-life data, belong to healthy male subjects with mean age 28 and standard deviation 4. To check the robustness of the method it needs to be tested on a bigger population with a higher variety in terms of health, sex and age.
- **Devices.** The overall data in the study is collected using the Empatica device, thus the proposed context-based method is biased towards that device.

The three limitations of the method are considered in the Fit4Work project [3]. Data is being gathered from a bigger

population and a higher variety in terms health, age and gender. Additionally, the Microsoft Band will be used as a wrist device for collecting the data. Finally, the proposed context-based method will be tested using the data gathered in the project trials.

6.2 Future work

The proposed method can be integrated in a system that helps people to overcome stressful situations. Such a system may overview stressful events, provide health advice and suggest stress-relieving exercises upon detected stress. For example, the method will be integrated in the Fit4Work system [3] as a part of a project that aims to develop an easy-to-use and unobtrusive system to support older workers in reducing and managing physical and mental stress resulting from their occupation.

In addition, in the future we plan to implement:

- **Dynamic and richer contexts**. We proved that context is crucial for improving the performance of the context-based stress detection method. For now, it uses the activity of the user and other date-time contexts (e.g., the hour of the day, the day of the week, etc.). The context-based models perform better with contexts extracted over smaller data windows (10–17.5 minutes) compared to larger data windows (20–30 minutes). This may be because the context is changing (e.g., the activity of the user) in a time-span of 10–15 minutes. Instead of using a fixed data window for extracting the context, a dynamic window can be utilized which changes with respect to the change in the context.

- **Personalization.** Stress is subjective, i.e., personal. This is confirmed both by the definition of stress and by the related work. Personalization can be achieved by using person-specific normalized data for extracting features. In addition to the personalized features, personalized ML models can be utilized either by adaptation of existing ML models (e.g., by using transfer learning [48]) or by building completely new models on a labeled data provided by the users.

- **Monitoring the other components of the stress response.** To monitor the other two components of the stress response (behavioral and affective component) we plan to incorporate methods for monitoring subjects' emotional state [49], and stress related behavioral changes [25] [26].

- **Internet of things and smart cities.** Can a stress-detection module be a part of an Internet of things network in a smart-city system? For example, by monitoring the stress level of public bus drivers, Rodrigues et al. [50] managed to construct a stress map of a city which can lead to better management of public transportation. Similarly, a stress map can be constructed based on different objectives, e.g., users' occupation, which can lead to recognizing the most stressful occupations that may require medical attention.

Finally, the dataset used in the study will be online in our AmI repository [51], which is another contribution in the field of stress detection in laboratory and real-life environments, considering the amount of data available.

7. Acknowledgments

The study presented in this paper was partially funded by the Fit4Work project, which is part of the AAL Joint Program. The authors would like to thank Vito Janko and Bozidara Cvetkovic for developing the smartphone application for collecting real-life data.

8. REFERENCES

[1] S.C. Segerstrom, G.E. Miller, "Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry," Psychological Bulletin 130, 601. 2004.

[2] Calculating the cost of work-related stress and psychosocial risks, [Online]. Available: https://osha.europa.eu/en/tools-and -publications/publications/literature_reviews/calculating-the-cost-of-work related-stress-and-psychosocial-risks, [Accessed 13.07.2016].

[3] Fit4Work Project, [Online]. Available: http://www.fit4work-aal.eu/ [Accessed 15.7.2016].

[4] H. G. Ice, G. D. James," Measuring Stress in Humans: A practical Guide for the field," Cambridge university press, 2007.

[5] Zephyr chest strap bio-sensor, [Online]. Available: https://www.zephyranywhere.com/products/bioharness-3

[6] M. Garbarino, M. Lai, D. Bender, R. W. Picard, S. Tognett , "Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition," 4th International Conference on Wireless Mobile Communication and Healthcare, pp. 3-6, 2014.

[7] R. Trobec, A. Rashkovska, V.Avbelj, "Two Proximal Skin Electrodes — A Respiration Rate Body Sensor," Sensors, 2012.

[8] K. Hovsepian, M. Absi, T. Kamarck, and M. Nakajima, "cStress : Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment," ACM Conf. on Ubiquitous Computing, 2015.

[9] H. Sarker et al. "Finding Significant Stress Episodes in a Discontinuous Time Series of Rapidly Varying Mobile Sensor Data," Human-Computer Interaction, 2016.

[10] J. A. Healey and R. W. Picard. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. IEEE Trans. Intell. Transp. Syst., Vol. 6, no. 2, pp. 156–166, 2005.

[11] A. Muaremi, A. Bexheti, F. Gravenhorst, B. Arnrich, and G. Tröster. Monitoring the Impact of Stress on the Sleep Patterns of Pilgrims using Wearable Sensors. IEEE-EMBS Int. Conf. Biomed. Heal. Informatics, pp. 3–6, 2014.

[12] A. D. S. Sierra and C. Ávila. Real-Time Stress Detection by Means of Physiological Signals. Gr. Biometrics, 2013.

[13] J. Ramos, J. Hong, and A. K. Dey, "Stress Recognition - A Step Outside the Lab," Proc. Int. Conf. Physiol. Comput. Syst., pp. 107–118, 2014.

[14] A. Sano and R. W. Picard. Stress Recognition Using Wearable Sensors and Mobile Phones. Hum. Assoc. Conf. Affect. Comput. Intell. Interact., pp. 671–676, 2013.

[15] J. Hernandez, R. R. Morris, and R. W. Picard. Call Center Stress Recognition with Person-Specific Models with Person-Specific Models. Affective Computing and Intelligent Interaction. Vol. 6974 pp. 125–134, 2011.

[16] J. Zhai, A. Barreto, "Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables," International Conference of the IEEE Engineering in Medicine and Biology Society, 2006.

[17] P. Melillo, M. Bracale, L. Pecchia, "Nonlinear Heart Rate Variability features for real-life stress detection. Case study: students under stress due to university examination," BioMedical Engineering, OnLine, 2011.

[18] W. Handouzi, C. Maaoui, A. Pruski, and A. Moussaoui. Short-term anxiety recognition from blood volume pulse signal. IEEE 11th Int. Multi-Conference Syst. Signals Devices, SSD 2014, 2014.

[19] J. Wijsman, B. Grundlehner. Wearable physiological sensors reflect mental stress state in office-like situations. Affective Computing and Intelligent Interaction (ACII), Humaine Conference on pp. 600–605, Sep. 2013.

[20] Z. Dharmawan, "Analysis of computer games player stress level using EEG data," Master of Science Thesis Report, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Netherlands, 2007.

[21] D. McDuff, J. Hernandez, S. Gontarek, R. Picard. COGCAM: Contact-free Measurement of Cognitive Stress During Computer Tasks with a Digital Camera. In Proceedings for the Computer and Human Interaction Conference (CHI), 2016.

[22] I. Mohino-Herranz, R. Gil-Pita, J. Ferreira, M. Rosa-Zurera, F. Seoane. "Assessment of Mental, Emotional and Physical Stress through Analysis of Physiological Signals Using Smartphones," Sensors, pp. 25607-27, 2015.

[23] H. Lu et al., "StressSense: Detecting Stress in Unconstrained Acoustic Environments using Smartphones," ACM Conf. on Ubiquitous Computing, pp. 351-360, 2012.

[24] P. Adams, M. Rabbi, T. Rahman, and M. Matthews. Towards Personal Stress Informatics: Comparing Minimally Invasive Techniques for Measuring Daily Stress in the Wild. pac.cs.cornell.edu, 2011.

[25] R. Wang et al., "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," ACM Conf. on Ubiquitous Computing, pp. 3-14, 2014.

[26] G. Bauer, P. Lukowicz, "Can Smartphones Detect Stress-Related Changes in the Behaviour of Individuals?" Pervasive Comp and Comm Workshops, 2012.

[27] M. Gjoreski, H. Gjoreski, M. Luštrek, M. Gams, "Automatic detection of perceived stress in campus students using smartphones," 11th International Conference on Intelligent Environments, Prague, 2015.

[28] H. Gjoreski et al. "Competitive Live Evaluation of Activity-recognition Systems," IEEE Pervasive Computing, Vol. 14, pp. 70–77, 2015.

[29] M. Gjoreski, H. Gjoreski, M. Luštrek, M. Gams, "Continuous stress detection using a wrist device: in laboratory and real life," ACM Conf. on Ubiquitous Computing, Workshop on mentalhealth, pp. 1185-1193, 2016.

[30] M. Gjoreski, H. Gjoreski, M. Luštrek, M. Gams, "Continuous Live Stress Monitoring with a Wristband," ECAI, PAIS, pp. 1803-1804.

[31] K. Dedovic, R. Renwick, N. K. Mahani, and V. Engert, "The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain," Vol. 30, no. 5, pp. 319–325, 2005.

[32] S.J. Lupien, F. Maheu, M. Tu, A. Fiocco, T.E. Schramek, "The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition," Brain and Cognition, Vol. 65, pp. 209–237.

[33] T. Eftimov, P. Korošec, B. Koroušic Seljak, "Disadvantages of Statistical Comparison Of Stochastic Optimization Algorithms," Bioinspired Optimization Methods and their Applications, BIOMA, 2016.

[34] M. Wu. "Trimmed and Winsorized Eestimators," PhD thesis, Michigan State University, 2006.

[35] G.K. Palshikar, "Simple Algorithms for Peak Detection in Time-Series," International Conference on Advanced Data Analysis, Business Analytics and Intelligence, 2009.

[36] K. A. Herborn. "Skin temperature reveals the intensity of acute stress," Journal of Physiology & Behavior. Vol. 152, pp. 225–230, 2015.

[37] H. Deng, G. Runger, E. Tuv, "Bias of importance measures for multi-valued attributes and solutions," 21st International Conference on Artificial Neural Networks, 2011.

[38] J.R. Quinlan, "Improved use of continuous attributes in c4.5," J. Artif. Intell. Res., Vol. 4, pp. 77–90, 1996.

[39] R. Stuart, N. Peter, "Artificial Intelligence: A Modern Approach," 2nd Ed., Prentice Hall: New York, USA, 2003.

[40] D. Aha, D. Kibler, "Instance-based learning algorithms," Machine Learning Vol. 6, pp. 37–66, 1991.

[41] N. Cristianini, J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods," Cambridge University Press: Cambridge, UK, 2000.

[42] L. Breiman, "Technical Report No. 421," 1994.

[43] K. Michael, "Thoughts on Hypothesis Boosting," Unpublished manuscript, December 1988.

[44] H. Tin Kam, "Random Decision Forests," International Conference on Document Analysis and Recognition, Montreal, QC, Canada, pp. 278–282, 1995.

[45] R. Caruana, A. Niculescu, G. Crew, A. Ksikes, "Ensemble Selection from Libraries of Models," 21st International Conference on Machine Learning, 2004.

[46] H. Gjoreski, M. Luštrek, M. Gams, "Accelerometer Placement for Posture Recognition and Fall Detection". 7th International Conference on Intelligent Environments, Nottingham, UK, pp. 47–54, 2011.

[47] M. Gjoreski, H. Gjoreski, M. Luštrek, M. Gams, "How Accurately Can Your Wrist Device Recognize Daily Activities and Detect Falls?," Sensors, 2016.

[48] A. Maxhunia, P. Hernandez-Lealc, L. E. Sucarc, V. Osmanib, E. F. Moralesc, O. Mayorab, "Stress modelling and prediction in presence of scarce data," Journal of Biomedical Informatics, Vol. 63, pp. 344–356, 2016.

[49] M. Gjoreski, H. Gjoreski, A. Kulakov, "Machine Learning Approach for Emotion Recognition in Speech," Informatica Vol. 38, pp: 377–384, 2014

[50] J. G. P. Rodrigues, J. P. S.Cunha. "A Mobile Sensing Approach to Stress Detection and Memory Activation for Public Bus Drivers," IEEE Transactions on Intelligent Transportation Systems, 2015.

[51] Josef Stefan Institute. Ambient Intelligence Repository (AmI Repository). Available online: http://dis.ijs.si/ami-repository/ (accessed on 10 December 2016).