In Search of Harmful Stress

Jaakko Tervonen
jaakko.tervonen@vtt.fi
VTT Technical Research Centre of Finland
Espoo, Finland

Johanna Närväinen
johanna.narvainen@vtt.fi
VTT Technical Research Centre of Finland
Kuopio, Finland

Jani Mäntyjärvi
jani.mantyjarvi@vtt.fi
VTT Technical Research Centre of Finland
Oulu, Finland

Kati Pettersson
kati.pettersson@vtt.fi
VTT Technical Research Centre of Finland
Espoo, Finland

ABSTRACT

Human body produces different physiological stress reaction when you hit a toe to a doorstep than when you panic at a job interview. The impact for body's homeostasis varies depending on the reaction type and some reactions are harmful to our health. Currently, stress estimation is focused on binary identification between stress and non-stress stages. More detailed separation of stress reaction types is needed for detecting harmful stress. In this study, the Extreme Gradient Boosting algorithm was used to classify a baseline condition and physiological and psychosocial stress, based on psychophysiological signals monitored using a wrist sensor device. Classification was robust in separating the two stress states from baseline and from each other. The results provide support for novel approaches utilizing fine-grained estimation of stress type from wearable sensor data.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing*; • **Applied computing** → *Health care information systems*; *Health informatics*.

KEYWORDS

stress detection; wearable sensing; psychophysiology

ACM Reference Format:
Jaakko Tervonen, Johanna Närväinen, Jani Mäntyjärvi, and Kati Pettersson. 2021. In Search of Harmful Stress. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (UbiComp-ISWC '21 Adjunct), September 21–26, 2021, Virtual, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3460418.3479335

1 INTRODUCTION

Recent development in stress monitoring has focused on improving the data processing pipeline and, on the other hand, finding the best signals and machine learning classifiers for binary (e.g. stress vs. non-stress [6]) state detection [8] or classifying the intensity of stress [3]. Even when a variety of stress stimuli was used [4], the induced stress state was considered a single state. However, there are different types of stressors and resulting stress, and their influence on health and well-being varies accordingly. It is obvious that a more fine-grained, yet robust and field-capable method for classification of stress types is much needed. In this paper we demonstrate our contribution to this challenge.

It is chronic stress that is detrimental to our well-being and mental health. Transient stress can help us perform better in challenging situations, but when repeated with no sufficient recovery, it can turn into chronic stress; slowly and insidiously since the human system adapts efficiently. Different types of stress initiate different physiological processes, with specific effects on the body's homeostasis and risk of chronic stress.

There are two stress-responsive axes: sympatho-adrenaline-medullary (SAM) and hypothalamus-pituitary-adrenalin (HPA). The SAM axis activation starts with the sympathetic nervous system, which increases arousal level via adrenaline [9, 10]. The HPA axis activity is commenced with the hypothalamus, which triggers a chain of events that causes the release of cortisol into the bloodstream [10]. It has been suggested that especially frequent HPA axis activation is harmful to our health and well-being and can cause depression, anxiety, and chronic stress [9]. The activation of HPA and SAM axes is often measured from saliva (via biomarkers such as cortisol and alpha-amylase) [10]. However, saliva samples are laborious and not optimal for monitoring stress. Robust and practical methods to detect not only the general stress but also the type of stress in everyday life are much needed.

Maastricht Acute Stress Test (MAST) is a stress-eliciting task, which has been developed to quickly and effectively activate the human stress response [10]. MAST consists of alternating trials of physical pain/discomfort (immersion of hand in ice water; cold pressor task CPT) and psycho-social stress (mental arithmetics (MA) task with time pressure and penalization). It has been suggested that these two stressors have significantly different impact on the stress response: the CPT induces strong SAM activation, while the MA stimulates the HPA axis [10].

Acute stress is reflected in physiology and detectable in different biosignals [8]. The type of stress makes a difference: a stress detector trained on an arithmetic task performed worse for other types of stress induction tasks [6]. In our recent study we used heart rate (HR, HRV) features derived from electrocardiogram to detect rest, CPT, and MA (i.e. MAST) in a laboratory setup and achieved 70.2% classification accuracy [7]. It has been suggested that electrodermal

UbiComp-ISWC '21 Adjunct, September 21–26, 2021, Virtual, USA

ACM ISBN 978-1-4503-8461-2/21/09.
https://doi.org/10.1145/3460418.3479335

activity (EDA) could complement HR and HRV measurements and provide more insight into the stress physiology [1]. In recent studies, unobtrusive, wearable devices with HR and EDA sensors have been used successfully to assess stress or cognitive load [2, 5, 6, 11] using machine learning methodology.

The aim of this work is to detect harmful stress. Psychophysiology during rest and two stressors (activation of the SAM and HPA stress-responsive axis) was recorded using a wearable wristband, the stress types (rest, CPT, and MA) were classified, and a reduced set of physiological parameters were found with a feature elimination process. This research is the first attempt to detect SAM and HPA stress reactions with a wrist device. The objective is to explore the possibilities for more detailed acute stress classification in daily-life monitoring context and further in the prevention of chronic stress.

2 METHODS

Study participants were healthy volunteers (n=17, 7 male, age mean and standard deviation 26.2 ± 6.2 years). The experimental protocol consisted of 2 min baselines (BL) in the beginning and in the end of the protocol and a stress induction. The MAST stress protocol consisted of 10 minutes of alternating phases of physical (CPT) and psycho-social (MA) stress, with each lasting from 45 s to 90 s [10].

Psychophysiology was monitored using the Empatica E4 wristband (Empatica Inc, MA, USA) providing EDA and blood volume pulse (BVP), sampled at 4 Hz and 64 Hz, respectively. After MAST, the experienced stress in CPT and MA phases and the maximum pain experienced during CPT was rated from 1 to 9.

Data processing consisted of preprocessing, feature extraction, and classification. All EDA and BVP data were analyzed in segments of 45 s with 15 s window slide, selected according to alternating phase lengths in the MAST. The EDA signal was preprocessed by filtering with a sliding mean filter, decomposing into phasic and tonic component, and detecting the skin conductance responses (SCR). The BVP signal was first filtered with a third order Butterworth bandpass filter with cutoff frequencies 0.5 Hz and 8 Hz, and then heartbeats were detected from the filtered signal. Interbeat intervals (IBIs) were extracted by computing the time between heartbeats and heart rate (HR) was computed from IBIs. Feature extraction followed [7, 11] and all extracted features are listed in Table 1.

To account for subjective physiological responses, the features were normalized by person-specific z-score standardization which was shown to perform better than other normalization strategies in [2]. The model was validated with leave-one-subject-out (LOSO) cross-validation as recommended in [8]. The data were classified with the Extreme Gradient Boosting (XGB) classifier since it has shown good performance earlier in similar contexts [2, 7].

The classification task was to classify between BL, CPT and MA. To eliminate subjectivity in the assessment of current state, the task labels were used as ground truth instead of the subjective ratings. The BL class data consisted of the two baseline periods. Classification performance was estimated with accuracy, the percentage of correct classifications, and F1-score, the harmonic mean of precision and recall, weighted according to class distribution.

Features were selected with an iterative procedure where the least important feature was eliminated until just one feature was left.

Table 1: Extracted features.

Type	Feature	Specification
EDA	Statistics of the signal, the tonic (ton) and the phasic (phas) component and the first derivative (d1)	mean, median (med), standard deviation (std), minimum (min), maximum (max), upper and lower quartile (uq, lq) and coefficient of variation (cv)
	Phasic: scr npeaks, scr height mean, scr amplitude mean, scr rise time mean, scr recovery time mean, power	Number of SCR peaks, their mean height and amplitude and mean rising and recovery time, signal power in five bands between 0.1Hz - 0.6Hz
	Tonic: cortim	Correlation with time
HR	Statistics of the signal and the first derivative (d1)	same as above for EDA along with range and slope
HRV	mean nni, median nni, range nni	normal-to-normal interval mean, median, range
	SDNN, SDSD	Std of IBIs and successive differences
	(p)NN20, (p)NN50	Percentage and number of IBIs differing more than 20ms/50ms
	RMSSD	Root mean square of successive differences
	CVNNI, CVSD	Ratio of SDNN and mean IBI, and RMSSD and mean IBI
	VLF, LF, HF, TotPow	Power in very low, low, high frequency bands, and total power
	LF/HF	Ratio of LF and HF
	LFNU, HFNU	Normalised LF and HF
	HRVTI	HRV triangular index
	CVI, CSI, modified CSI	(modified) cardiac sympathetic index, cardiac vagal index
	SD1, SD2, SD2/SD1	Poincaré plot std perpendicular and along the identity line, their ratio

Feature naming later in text and figures: type_feature_statistic.
Abbreviations used later are in parenthesis.

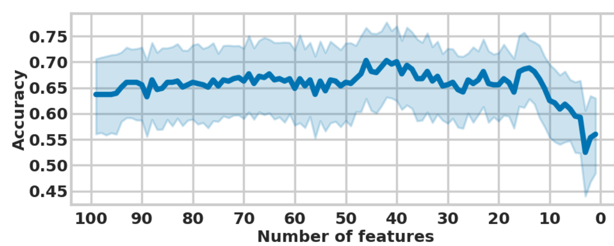

Figure 1: Detected accuracy and 95% bootstrap confidence interval during the feature elimination experiment.

Feature importance was estimated as relative impurity reduction: the importance of a feature was the normalized total reduction of Gini impurity brought by that feature in the XGB model. We report the most important features affecting classification, but a more thorough analysis of each feature's relation to SAM and HPA activation is left for future work.

3 RESULTS AND DISCUSSION

Fig. 1 shows the accuracies obtained during feature elimination. Starting at 99 features, the accuracy was rather stable until around twenty features were left, after which it started to decline. The maximum accuracy of 0.703 ± 0.152 was reached with 46 features, with F1-score of 0.705 ± 0.144 (values mean ± std).

The confusion matrix obtained during LOSO validation is shown in Table 2. Baseline was confused with CPT more often than MA. This is in line with subjective reports of stress: 5.6 ± 2.0 for MA and 3.7 ± 1.8 for CPT (mean ± std, scale 1-9). The CPT task is also

Table 2: Confusion matrix with the best feature set.

		Predicted		
		BL	CPT	MA
True	BL	135	23	14
	CPT	25	97	31
	MA	5	29	68

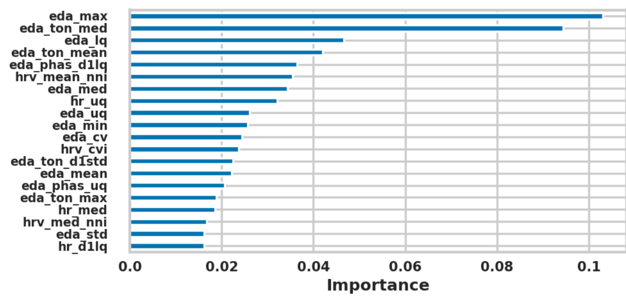

Figure 2: Relative importances of top-20 features with the best performing feature set.

more dynamic than MA, as the biosignal response to the initial cold sensation settles during the first 10 seconds (data not shown) and the pain sensation builds up gradually. Averaging over the window duration is likely to dilute some of the effect.

Subjective preferences and skills influence the intensity and even the type of stress induced by MAST. The physical pain felt during cold immersion varies, and people familiar with mental arithmetics and/or presenting in public tend to experience less anxiety while counting. According to subjective ratings, stress was experienced more in the arithmetic task, but the cold immersion clearly succeeded in inducing pain (maximum pain 5.1 ± 1.7; mean ± std, scale 1-9). Also the inter-individual psychophysiological responsivity varies greatly which has likely decreased the accuracy in the leave-one-out validation. The small number of participants contributes to the issue, but as persons are inherently different, also in larger data sets, the predictive power of models tends to remain low. One solution to this problem in long-term monitoring is personal optimization of the underlying model, built during a period when the users annotate their cognitive state.

Fig. 2 displays the top-20 feature importances of the best performing combination. The statistical features of the EDA signal and its tonic component were predominant in each task but some HR and HRV were also among the most important. This observation is contrary to [11], where HR and HRV features were more important than EDA features in binary cognitive load detection. However, the wearable device and the tasks employed were different which has probably affected the importance scores. Moreover, the tasks in the current study probably induced stronger stress response than in [11], which is reflected as higher activation of EDA.

The classification performance in this study was comparable to earlier multi-level stress detection studies with mobile wearable devices, e.g. 0.73 accuracy in classifying the intensity of stress [3] and class-wise F1-score between 0.50 and 0.82 in classifying three stress types but excluding the rest state from classification [5]. However, the performance was lower than achieved using a more involved sensing solution, e.g. up to 0.86 accuracy in [7]. Our next step is to try and improve the performance by leveraging transfer learning to utilize data collected in other similar studies.

4 CONCLUSIONS

Hit your toe or panic at a work? Here, we demonstrate a method for separating potentially harmful acute stress from benign stress, based on data from single wrist-worn device with reasonable accuracy. This capability enables a variety of new options for monitoring human mental state in everyday life. Applications e.g. in workplace well-being (detection of bad stress and recovery), private wellness monitoring (insights into daily activities and annoyances) and human technology interfaces (adapting to user's mental mode) are potential adapters of this approach.

ACKNOWLEDGMENTS

We thank Anita Honka and Emmi Antikainen for conducting the study visits, and Sari Järvinen and Juha M. Kortelainen for their help with study design and data collection. The work was funded by the Academy of Finland under GrantNo.: 334092 and VTT.

REFERENCES

[1] Shadi Ghiasi, Alberto Greco, Riccardo Barbieri, Enzo Pasquale Scilingo, and Gaetano Valenza. 2020. Assessing autonomic function from electrodermal activity and heart rate variability during cold-pressor test and emotional challenge. *Scientific reports* 10, 1 (2020), 1–13.

[2] Martin Gjoreski, Tine Kolenik, Timotej Knez, Mitja Luštrek, Matjaž Gams, Hristijan Gjoreski, and Veljko Pejović. 2020. Datasets for cognitive load inference using wearable sensors and psychological traits. *Applied Sciences* 10, 11 (2020), 3843. https://doi.org/10.3390/app10113843

[3] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. 2017. Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics* 73 (2017), 159–170. https://doi.org/10.1016/j.jbi.2017.08.006

[4] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2018. The Case for a Commodity Hardware Solution for Stress Detection. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 1717–1728. https://doi.org/10.1145/3267305.3267538

[5] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2020. Continuous Detection of Physiological Stress with Commodity Hardware. *ACM Trans. Comput. Healthcare* 1, 2, Article 8 (April 2020), 30 pages. https://doi.org/10.1145/3361562

[6] Varun Mishra, Sougata Sen, Grace Chen, Tian Hao, Jeffrey Rogers, Ching-Hua Chen, and David Kotz. 2020. Evaluating the Reproducibility of Physiological Stress Detection Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 147 (Dec. 2020), 29 pages. https://doi.org/10.1145/3432220

[7] Kati Pettersson, Jaakko Tervonen, Johanna Närväinen, Pentti Henttonen, Ilmari Määttänen, and Jani Mäntyjärvi. 2020. Selecting Feature Sets and Comparing Classification Methods for Cognitive State Estimation. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, New York, NY, USA, 683–690. https://doi.org/10.1109/BIBE50027.2020.00115

[8] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. 2019. Wearable-based affect recognition—a review. *Sensors (Switzerland)* 19, 19 (2019), 1–42. https://doi.org/10.3390/s19194079

[9] Christopher F. Sharpley. 2009. Neurobiological Pathways between Chronic Stress and Depression: Dysregulated Adaptive Mechanisms? *Clinical Medicine Insights: Psychiatry* 2 (2009), CMPsy.S3658. https://doi.org/10.4137/CMPsy.S3658

[10] Tom Smeets, Sandra Cornelisse, Conny WEM Quaedflieg, Thomas Meyer, Marko Jelicic, and Harald Merckelbach. 2012. Introducing the Maastricht Acute Stress Test (MAST): A quick and non-invasive approach to elicit robust autonomic and glucocorticoid stress responses. *Psychoneuroendocrinology* 37, 12 (2012), 1998–2008.

[11] Jaakko Tervonen, Kati Pettersson, and Jani Mäntyjärvi. 2021. Ultra-Short Window Length and Feature Importance Analysis for Cognitive Load Detection from Wearable Sensors. *Electronics* 10, 5 (2021), 19 pages. https://doi.org/10.3390/electronics10050613