Catching patient's attention at the right time to help them undergo behavioural change: Stress classification experiment from Blood Volume Pulse

Aneta Lisowska[1], Szymon Wilk[1], and Mor Peleg[2]

[1] Institute of Computing Science, Poznań University of Technology, Poznań, Poland
{aneta.lisowska, szymon.wilk}@put.poznan.pl
[2] Department of Information Systems, University of Haifa, Haifa, Israel
morpeleg@is.haifa.ac.il

Abstract. The CAPABLE project aims to improve the wellbeing of cancer patients managed at home via a coaching system recommending personalized evidence-based health behavioral change interventions and supporting patients compliance. Focusing on managing stress via deep breathing intervention, we hypothesise that the patients are more likely to perform suggested breathing exercises when they need calming down. To prompt them at the right time, we developed a machine-learning stress detector based on blood volume pulse that can be measured via consumer-grade smartwatches. We used a publicly available WESAD dataset to evaluate it. Simple 1D CNN achieves 0.837 average F1-score in binary *stress vs. non-stress* classification and 0.653 in *stress vs. amusement vs. neutral* classification reaching the state-of-art performance. Personalisation of the population model via fine-tuning on a small number of annotated patient-specific samples yields 12% improvement in *stress vs. amusement vs. neutral* classification. In future work we will include additional context information to further refine the timing of the prompt and adjust the exercise level.

Keywords: Blood volume pulse · Stress · Classification · Wearable · Fogg Behavioral Model

1 Introduction

Cancer patients frequently experience negative emotions such as stress, sadness and fear for the future, that hinder their emotional wellbeing [13], correlate with reduced treatment compliance [5], and increase risk of mortality [15]. Improving the emotional and physical wellbeing of patients is a goal of the Horizon 2020 CAncer PAtient Better Life Experience (CAPABLE) project. CAPABLE aims to develop and implement new persuasive computing methods that will provide cancer patients at home with continuous support for treatment adherence and the development of positive health habits that ultimately can improve their

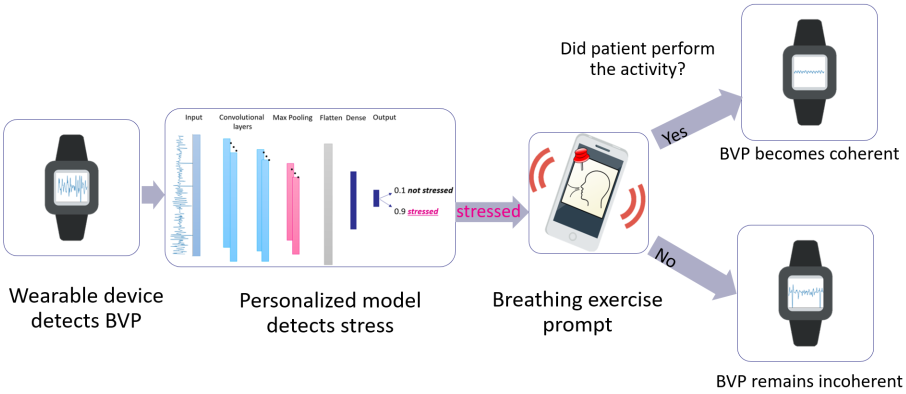

Fig. 1. Stress intervention system

wellbeing. Patients are equipped with a smartwatch, which monitors their pulse, sleep, and physical activity, as well as a coaching system (a virtual coach), accessed via a mobile app. Our goal is to design an app-based mobile intervention aims to help cancer patients form desirable health habits that will improve their quality of life and wellbeing that builds emotional resilience. The virtual coach can suggest to patients evidence-based activities from the domain of mindfulness and positive psychology, known to have positive effects on one's wellbeing. Each (tiny) activity is meant to become a habit aligned with patients physical and psychological wellbeing goals.

Fogg's Tiny Habits Behaviour Model [3] proposes that habits formation (i.e., performing a target activity) depends on three factors: motivation, ability to perform the task (which depends on the task's difficulty) and the presence of a trigger reminding the person to perform the target behaviour. For the purpose of this feasibility study, we assume that patients are motivated to achieve their wellbeing goals and that there are simple behaviours that match their abilities; therefore we concentrate on designing appropriately-timed prompts. To identify the simplest behaviour that all cancer patients have the ability to perform and that can impact their emotional wellbeing we draw inspiration from *Integrated Performance Model* [23]. According to Watkins, physiology impacts emotions, which in turn elicit feeling, thinking, and behaviour, leading to desired outcomes (results). Hence to achieve the best outcome, the starting point is changing one's physiology. Watkins points out that controlled rhythmic breathing can influence one's heart rate variability (HRV) and change one's physiological state, leading to a state of 'stable variability' called *coherence*[23]. This is in line with research that investigated the impact of breathing techniques on the autonomic and the central nervous systems [25]. Conscious slow breathing has been found to reduce negative emotions and stress [24] and increase emotional control [6]. Therefore, in CAPABLE, breathing exercise is the target behaviour that our cancer patients have the ability to perform and can benefit from [7].

The challenging part and a primary focus of this paper, is designing the trigger that will deliver the recommendation to perform the breathing exercise at a suitable time. We propose to use wearable sensors and machine learning for that purpose. We want to prompt the patient when they might most need a mindful breathing intervention; at the point their physiology signals their stress (Fig. 1).

There is a considerable volume of literature around automatic stress detection methods using measurements from electrocardiogram (ECG), respiratory rate, electrodermal activity (EDA), speech excitation, body posture or thermal infrared imaging [4]. In this paper, we pay particular attention to methods that can detect stress from blood volume pulse (BVP) (Section 2). We designed a personalised stress detector, leveraging only BVP (Section 3). BVP can be continuously captured with minimum inconvenience to the patient wearing a consumer-grade smartwatch. Additionally, BVP is altered by the breathing pattern [21], thus it can provide useful evaluation of effectiveness of the prompt and the breathing exercise intervention. In this paper we focus on experimental evaluation of the first part of the system, namely of our proposed stress detector (Section 4) that will dictate the appropriate timing for the breathing exercise prompt. Specifically, we intend to answer the following research questions:

- Is it possible to develop a stress detector using raw BVP data (without hand-crafted features)?
- Does personalisation of a stress detector improve its performance?

2 Related Work

Stress is characterised by high physiological arousal [9]. Regardless of the person's age, tense arousal is associated with higher heart rate [16]. However, discriminating stress from other emotions, such as excitement or flow (an intrinsically rewarding experience of being immersed in a task [1]), is not straightforward, as both result in elevated physiological arousal and modulation of the heart period [14]. According to Russel, emotional states can be mapped into a 2D space [17], where arousal can be assigned to the first dimension and valence to the second one [10]. Valence reflects how positive or negative the emotion is. From the perspective of this study, it is important to discriminate between positive arousal states and stress, as both fall high in the arousal spectrum compared to relaxation but differ in the enjoy-ability of the experience (valence).

McCraty and Rees argue that positive emotions are characterised by the coherent pattern of the heartbeat rather than the heart rate [12]. Therefore, to distinguish between emotions it is important to acquire the full pulse signal and inspect fluctuation in the heartbeats rather than look only at the cumulative measure of the beats per minute. The heart beat-to-beat variability over time can be captured through measurement of the BVP using Photoplethysmography (PPG) sensor. The PGG sensors can be attached anywhere on the body, however, most commonly they are placed on a finger [22] or wrist [19] and importantly they are integrated into consumer-grade smartwatches [18].

Schmit *et al.* utilise both wrist and chest-worn sensors to acquire a multimodal dataset for Wearable Stress and Affect Detection (WESAD)[19]. This

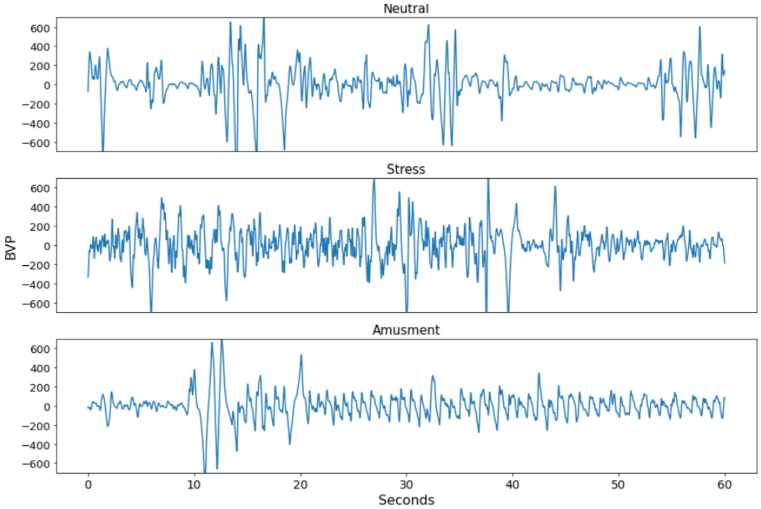

Fig. 2. An exemplar 1 minute snippets of BVP measurements from one subject in the WESAD dataset captured during 3 different emotions: neutral, stress, amusement.

data set is particularly interesting because it captures the positive arousal state of amusement alongside stress and neutral emotion from 15 subjects (Fig. 2). Constructively, Schmit et al. [19] provide baseline performance of multiple machine learning methods trained on features extracted from each of the sensors separately and in combination. When utilising only BVP, Linear Discriminant Analysis (LDA) achieved the best emotion classification performance in leave-one-out evaluation. The reported F1 score for the three-class problem is 0.547 from LDA, closely followed by Random Forest (0.538) and AdaBoost (0.533). In the reduced version of the problem limited to a binary classification of stress vs. not-stress, the F1 score from BVP reached 0.831. This is the state of the art performance of a generalised model on this dataset when leveraging only BVP. Nevertheless, given that people differ in their physiological response, personalised stress models might offer better performance [20]. Indikawati and Winiarti used all wrist sensor measurements from WESAD dataset to train personalised emotion classifier and obtained 88-99% classification accuracy using Random Forest [8]. However, they do not report the performance using BVP without the inclusion of EDA and temperature measurements.

3 Methods

3.1 Dataset

We use publicly available WESAD dataset [19] for development of our stress detector. The BVP measurements (64Hz) from 15 subjects are accompanied by the annotation of stress, amusement and baseline state, corresponding to the experimental conditions that were designed to evoke these emotions and can be used as the labels for training of the machine learning models. For each subject, the duration of the baseline condition is ∼ 20 minutes, of amusement - 6 minutes, and of stress - 10 minutes.

The WESAD dataset comes additionally with the Short Stress State Questionnaire (SSSQ) questionnaire, which captures subjective reports on how the participants felt during each experimental condition. Our evaluation considers self-reported worry that reflects the degree of negative emotion of a given subject.

3.2 Stress Detector

Unlike prior methods applied to this dataset [19,8],we do not use hand crafted features derived from the BVP, but directly input 60-second long snippets of the signal to a simple 1D convolution neural network. This type of model was used previously for classification of 1D signal gathered with a wearable device[11]. The model is constructed from two convolutional layers with 16, 8 filters respectively and kernel size of 3, followed by max pooling layer and fully connected layer with 30 nodes and output layer, which size corresponded to the number of the classes (emotions). The number of filters and nodes was chosen empirical. Each convolutional layer has a ReLu activation function and output layer with softmax. The model is implemented in Keras with a TensorFlow backend.

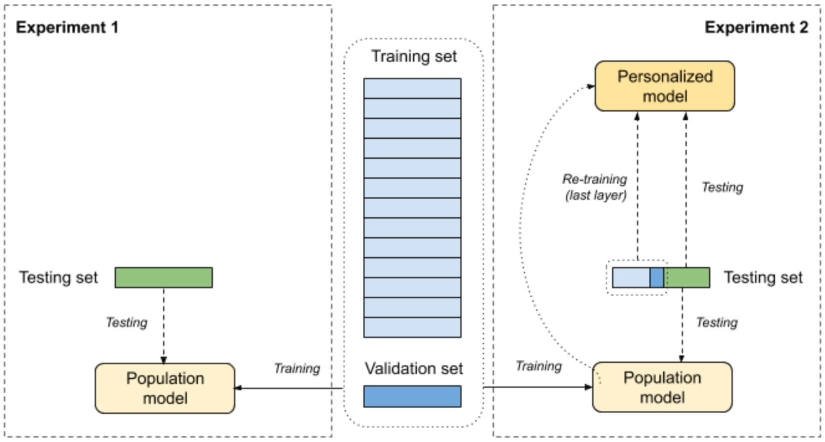

Fig. 3. Experimental set up.

3.3 Experiments

To answer our research questions we conduct two experiments (Fig. 3). The first experiment investigates if our stress detector trained on BVP signals reaches state-of-the-art classification performance. In the second experiment we examine if stress detector personalisation can provide further performance boost.

Experiment 1: When evaluating the generalised stress detector we follow Schmidt *et al.* [19] and adapt a leave-one-out (LOO) approach to directly benchmark against other previously reported approaches. The extracted examples from the test patient have a window size of 3840 samples and a step size of 1600 samples, yielding ∼88 test examples (Fig. 4). We use weighted F1-score as a performance metric recommended for unbalanced classification tasks.

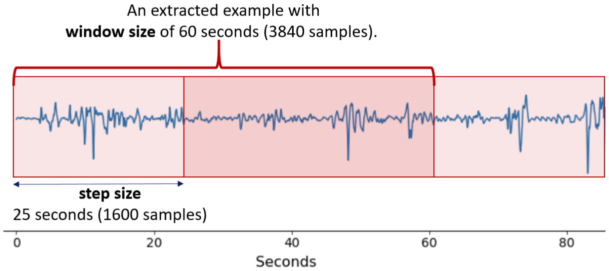
An extracted example with
window size of 60 seconds (3840 samples).
step size
25 seconds (1600 samples)
0
20
40
60
80
Seconds

Fig. 4. An example of test example extraction from patient signal.

For population model training we extract training examples with window size of 3840 samples. In two-class problem the step size is 18 samples for non-stress condition and 12 samples for stress condition, which yields ∼117000 training and 8600 validation examples. In three-class problem, we use a step size of 18, 10, 12 samples for baseline, stress and amusement conditions respectively, yielding ∼130000 training and 9300 validation examples. The step size was varied during training to reduce the imbalance between classes.

Experiment 2: The number of annotated examples that can be obtained from the patient using the mobile app is very small, therefore training the model from scratch on the single patient data is not feasible. We suggest that the personalised model can be trained by fine-tuning of the population model on a small number of annotated samples from the patient of interest. All the layers of the population model except the last one are frozen and the first half of the patients' BVP signal from each emotional condition is used for the model retraining. Similarly as in population model training the step size for non-stress events was of a size 18 samples and 12 for stress, resulting in ∼4000 retraining examples per patient. In three-class classification baseline examples were extracted with a step size 16, stress with 7 and amusement of 5 yielding ∼6000 retraining examples. In both personalised fine tuning conditions 80% of retraining examples were used for training and 20% for validation. To be able to directly compare population and personalised models we applied both the second half of the patient's BVP signals. We used the same windows and step size (3840 and 1600 samples, respectively) as in experiment 1 and obtained ∼43 test examples for each patient. Obtained performance is reported using micro F1.

In all experimental conditions we run model training 3 times with different random seeds used for model initialisation and batch ordering. We trained the model with categorical cross-entropy using Adam optimiser and batch size 256. To avoid overfitting we use early stopping on validation data.

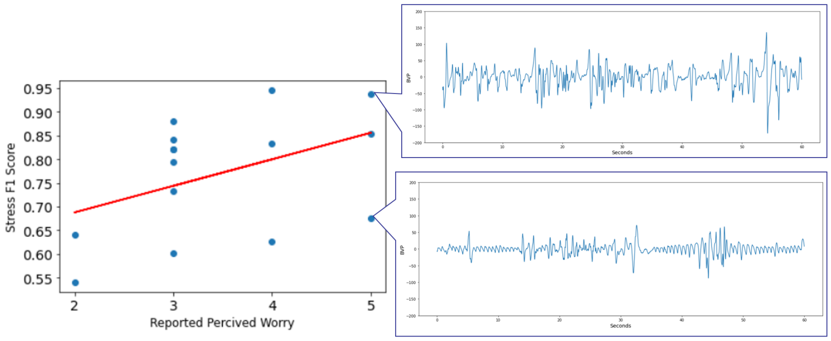

Fig. 5. Left: Population model performance vs. subjects perceived worry. Right: An exemplar BVP signal from 1 minute of the stressed condition.

4 Results

Experiment 1: In discriminating *stress vs. non-stress* our model reached average across all subjects of 0.837 F1 score and in three-class condition 0.653. These results are slightly higher than the results obtained by LDA, the best generalised approach using BVP previously reported [19]. Nevertheless, the variation in performance between subjects is very high at 0.71-0.98 in a two-class problem. We hypothesised that the generalised model works well for subjects who are more stressed and is less accurate for moderately stressed subjects. To investigate whether the model performance reflects the degree of negative emotion of the subjects we plot the stress F1 score from the 3 class model against self-reported worry (Fig. 5). Note that the weak positive trend between model performance and subjects perceived worry is not significant ($r_s = 0.477, p = 0.07$). The stress detector performed better for the subjects, whose stress response manifested in very erratic BVP.

Experiment 2: Table 1 shows the F1 score for each patient using the small patient-specific testing set. In *stress vs. non-stress* classification, personalised model achieved better classification results for 10 subjects out of 15, there was no difference in performance for 1 subject and for 4 subjects personalised model yielded slightly worse results. The decrease in performance might be due to an insufficient number of samples retraining examples leading to overfitting of already well performing population models. On average personalised model achieved a higher score of 0.822 compared to the population model 0.813 with similar standard deviation (0.08). In a classification of *baseline (neutral) emotion vs. stress vs amusement*, the personalisation of the model leads to better performance for 13 subjects, for 1 there was no difference and for one the performance decreased. The average F1 score for the personalised model is 0.705 (0.119) and population model 0.585 (0.176). Note the variance of the population models in this condition is higher than of the personalised models. In three-class problem the advantage of the model personalisation is more apparent than in binary classification task.

	stress vs. non-stress		***baseline vs. stress vs. amusement***	
subj.	Population Model Mean [std]	Personalised Model Mean [std]	Population Model Mean [std]	Personalised Model Mean [std]
2	0.894 [0.023]	0.813 [0.050]	0.715 [0.046]	0.732 [0.040]
3	0.821 [0.064]	0.829 [0.060]	0.520 [0.064]	0.585 [0.087]
4	0.870 [0.083]	0.911 [0.023]	0.870 [0.064]	0.870 [0.064]
5	0.937 [0.022]	0.937 [0.030]	0.468 [0.045]	0.754 [0.011]
6	0.762 [0.019]	0.778 [0.011]	0.651 [0.022]	0.683 [0.030]
7	0.889 [0.011]	0.944 [0.011]	0.548 [0.085]	0.635 [0.129]
8	0.849 [0.062]	0.794 [0.059]	0.595 [0.070]	0.500 [0.118]
9	0.675 [0.011]	0.714 [0.019]	0.373 [0.096]	0.619 [0.101]
10	0.742 [0.057]	0.735 [0.054]	0.462 [0.070]	0.636 [0.067]
11	0.891 [0.011]	0.915 [0.011]	0.535 [0.083]	0.814 [0.076]
13	0.845 [0.022]	0.791 [0.083]	0.636 [0.029]	0.659 [0.090]
14	0.752 [0.044]	0.760 [0.040]	0.535 [0.087]	0.783 [0.194]
15	0.620 [0.044]	0.674 [0.033]	0.194 [0.029]	0.543 [0.105]
16	0.814 [0.068]	0.891 [0.077]	0.845 [0.040]	0.868 [0.011]
17	0.833 [0.077]	0.841 [0.064]	0.818 [0.037]	0.894 [0.021]
Mean	0.813	**0.822**	0.584	**0.705**
Std	0.084	0.081	0.176	0.119

Table 1. Mean F1 micro for each subject from 3 runs with different random seeds in personalised evaluation set up for both 2 class and 3 class classification problem. The better performing model is highlighted in grey.

5 Discussion

The CAPABLE project aims to support cancer patients with achieving their wellbeing goals. In this work, we take emotional health as a case study and propose a system for stress intervention. As a first step, we focus on empirical evaluation of the stress detector that we intend to utilize for prompting the patient to perform simple breathing exercise. Our generalized stress detector requires fewer preprocessing steps than previously used methods (as it is applied directly on the raw BVP signal) and achieves the state of the art performance on WESAD dataset. Nevertheless, there is a large variation in the appearance of the BVP signal in stress conditions between patients suggesting that personalization of the models could be beneficial. We find that personalisation of the generalised stress detector simply via fine-tuning on the small number of annotated samples shows an encouraging improvement in classification performance.

In practice, further improvements to the personalised models could be obtained with methods as active learning, where the patient is occasionally asked to report their emotion via annotation of their current emotional state. Similar techniques could drive improvements of the population model where annotated samples are gathered from multiple users and used to retrain the model in federated learning fashion. The best stress detectors might rely on a combination of the personalized models or assignment of the patient to the subpopulation model based on patients' similarity to that subpopulation.

However, from the perspective of the full stress intervention solution, the performance of the stress detector is not the only important factor; while others intended to simply detect stressful events, we try to determine when it is the best time to prompt a patient to perform a selected breathing activity. We hypothesize that prompting patients when they need the intervention the most (i.e., when they are stressed) will increase their compliance. We plan to evaluate this hypothesis by comparing the effectiveness against that of prompts sent randomly. Effectiveness could be concluded if coherent BVP would be measured soon after the patient clicked on the prompt reminding him to perform the breathing.

In future work, we will incorporate additional information such as GPS location, or time of the day, to try to predict the best timing of the prompt further. We also plan to follow *Fogg's 8 steps persuasive design process*[2], and figure out what is preventing users from performing target behaviour. Therefore, in case the activity was not performed, we will request the patient to specify whether the suggested exercise was adequate and whether the timing was good; the patient might have been stressed but the timing of the prompt could have been poor because the patient was performing some other activity (e.g. driving, shopping) preventing them from engaging in the suggested exercise.

Acknowledgments

The CAPABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875052.

References

1. Csikszentmihalyi, M.: Flow: The psychology of optimal experience, vol. 1990. Harper & Row New York (1990)
2. Fogg, B.J.: Creating persuasive technologies: an eight-step design process. In: Proceedings of the 4th international conference on persuasive technology. pp. 1–6 (2009)
3. Fogg, B.J.: Tiny Habits: The Small Changes That Change Everything. Houghton Mifflin Harcourt (2019)
4. Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., Tsiknakis, M.: Review on psychological stress detection using biosignals. IEEE Transactions on Affective Computing (2019)
5. Greer, J.A., Pirl, W.F., Park, E.R., Lynch, T.J., Temel, J.S.: Behavioral and psychological predictors of chemotherapy adherence in patients with advanced non-small cell lung cancer. Journal of psychosomatic research **65**(6), 549–552 (2008)
6. Gross, M.J., Shearer, D.A., Bringer, J.D., Hall, R., Cook, C.J., Kilduff, L.P.: Abbreviated resonant frequency training to augment heart rate variability and enhance on-demand emotional regulation in elite sport support staff. Applied psychophysiology and biofeedback **41**(3), 263–274 (2016)
7. Hayama, Y., Inoue, T.: The effects of deep breathing on 'tension–anxiety'and fatigue in cancer patients undergoing adjuvant chemotherapy. Complementary therapies in clinical practice **18**(2), 94–98 (2012)
8. Indikawati, F.I., Winiarti, S.: Stress detection from multimodal wearable sensor data. In: IOP Conference Series: Materials Science and Engineering. vol. 771, p. 012028. IOP Publishing (2020)

9. Johnson, A.K., Anderson, E.A.: Stress and arousal. Principles of psychophysiology: Physical, social and inferential elements. Cambridge University Press (1990)
10. Lang, P.J.: The emotion probe: studies of motivation and attention. American psychologist **50**(5), 372 (1995)
11. Lisowska, A., O'Neil, A., Poole, I.: Cross-cohort evaluation of machine learning approaches to fall detection from accelerometer data. In: HEALTHINF. pp. 77–82 (2018)
12. McCraty, R., Rees, R.A.: The central role of the heart in generating and sustaining positive emotions. Oxford handbook of positive psychology pp. 527–536 (2009)
13. Page, A.E., Adler, N.E., et al.: Cancer care for the whole patient: Meeting psychosocial health needs. National Academies Press (2008)
14. Peifer, C., Schulz, A., Schächinger, H., Baumann, N., Antoni, C.H.: The relation of flow-experience and physiological arousal under stress—can u shape it? Journal of Experimental Social Psychology **53**, 62–69 (2014)
15. Pinquart, M., Duberstein, P.: Depression and cancer mortality: a meta-analysis. Psychological medicine **40**(11), 1797–1810 (2010)
16. Riediger, M., Wrzus, C., Klipker, K., Müller, V., Schmiedek, F., Wagner, G.G.: Outside of the laboratory: Associations of working-memory performance with psychological and physiological arousal vary with age. Psychology and Aging **29**(1), 103 (2014)
17. Russell, J.A.: Affective space is bipolar. Journal of Personality and Social Psychology **37**(3), 345–356 (1979)
18. Saganowski, S., Kazienko, P., Dzieżyc, M., Jakimów, P., Komoszyńska, J., Michalska, W., Dutkowiak, A., Polak, A., Dziadek, A., Ujma, M.: Review of consumer wearables in emotion, stress, meditation, sleep, and activity detection and analysis. arXiv preprint arXiv:2005.00093 (2020)
19. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing wesad, a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 400–408 (2018)
20. Shi, Y., Nguyen, M.H., Blitz, P., French, B., Fisk, S., De la Torre, F., Smailagic, A., Siewiorek, D.P., al'Absi, M., Ertin, E., et al.: Personalized stress detection from physiological measurements. In: International symposium on quality of life technology. pp. 28–29 (2010)
21. Steffen, P.R., Austin, T., DeBarros, A., Brown, T.: The impact of resonance frequency breathing on measures of heart rate variability, blood pressure, and mood. Frontiers in public health **5**, 222 (2017)
22. Udovičić, G., erek, J., Russo, M., Sikora, M.: Wearable emotion recognition system based on gsr and ppg signals. In: Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care. pp. 53–59 (2017)
23. Watkins, A.: Coherence: The secret science of brilliant leadership. Kogan Page Publishers (2013)
24. Yu, X., Fumoto, M., Nakatani, Y., Sekiyama, T., Kikuchi, H., Seki, Y., Sato-Suzuki, I., Arita, H.: Activation of the anterior prefrontal cortex and serotonergic system is associated with improvements in mood and eeg changes induced by zen meditation practice in novices. International Journal of Psychophysiology **80**(2), 103–111 (2011)
25. Zaccaro, A., Piarulli, A., Laurino, M., Garbella, E., Menicucci, D., Neri, B., Gemignani, A.: How breath-control can change your life: a systematic review on psycho-physiological correlates of slow breathing. Frontiers in human neuroscience **12**, 353 (2018)