Having a Bad Day? Detecting the Impact of Atypical Events Using Wearable Sensors

Keith Burghardt(✉), Nazgol Tavabi, Emilio Ferrara, Shrikanth Narayanan, and Kristina Lerman

USC Information Sciences Institute, Marina del Rey, CA 90292, USA
keithab@isi.edu

Abstract. Life events can dramatically affect our psychological state and work performance. Stress, for example, has been linked to professional dissatisfaction, increased anxiety, and workplace burnout. We therefore explore the impact of atypical positive and negative events on a number of psychological constructs through a longitudinal study of hospital and aerospace workers. We use causal analysis to demonstrate that positive life events increase positive affect, while negative events increase stress, anxiety and negative affect. While most events have a transient effect on psychological states, major negative events, like illness or attending a funeral, can reduce positive affect for multiple days. These findings provided motivation for us to train machine learning models that detect whether someone has a positive, negative, or generally atypical event. We show that wearable sensors paired with embedding-based learning models can be used "in the wild" to help detect atypical life events of workers across both datasets. Extensions of our results will offer opportunities to regulate the negative effects of life events through automated interventions based on physiological sensing.

Keywords: Psychological constructs · Atypical events · Causal modeling · Wearable sensors · Machine learning

1 Introduction

As organizations prepare their workforce for changing job demands, worker wellness has emerged as an important focus, especially since the COVID-19 pandemic. Worker wellness is central to organization missions to maintain optimal job performance by developing a healthy and productive workforce. This goal is especially important in high-stakes jobs, such as healthcare providers and other frontline workers, where job-related stress often leads to burnout and poor performance [11,15,22], and is one of the most costly modifiable health issues at the workplace [10]. An additional challenge faced by workers is balancing job demands with equally stressful events in their personal life. Adverse events—such as illness or death of a family member and the death of a pet—may amplify worker stress and

R. Thomson et al. (Eds.): SBP-BRiMS 2021, LNCS 12720, pp. 257–267, 2021.
https://doi.org/10.1007/978-3-030-80387-2_25

potentially harm job performance. On the other hand, positive life events—such as getting a raise, getting engaged, or taking a vacation—may decrease stress and improve well-being. The ability to detect such atypical life events in a workforce can help organizations better balance tasks to reduce stress, burnout, and absenteeism and improve job performance. Until recently, detecting such life events automatically, in real time and at scale, would have been unthinkable. Recent advances in accurate and relatively inexpensive wearable sensors, however, offer opportunities for unobtrusive and continuous acquisition of diverse physiological data. Sensor data allows for real-time, quantitative assessment of individual's health and psychological well-being [12,18] . It could also provide insights into atypical life events that individual workers experience and could affect their well-being and job performance, while keeping information about these personal events private from the organization. However, the connection between sensor data, atypical life events, and individual well-being, has not been demonstrated for such dynamic environments, especially in real-world scenarios.

In this paper, we ask what effects do atypical events have on workers, and can we detect these events with non-invasive wearable sensors? We study two large longitudinal studies of hospital and aerospace workers who wore sensors and reported ecological momentary assessments (EMAs) over the course of several months. Workers also reported whether they had experienced an atypical event. We apply difference-in-difference analysis, a causal inference method [20], to measure the effect of atypical events—either positive or negative events—on individual psychological states and well-being. We show atypical events have dramatic effects on psychological states, which motivates our event detection method. Namely, negative events increase self-reported stress, anxiety, and negative affect by 10–20% or more, while decreasing positive affect over multiple days. Positive life events, meanwhile, have little effect on stress, anxiety, and negative affect, but boost positive affect on the day of the event. Overall, negative atypical events have a greater impact on worker's psychological states than positive events, which is in line with previous findings [2]. Next, we show that it is possible to detect these events from a non-invasive wristband sensor. We propose a method that learns a representation of multi-modal physiological signals from sensors by embedding them in a lower-dimensional space. The embedding provides features for classifying when atypical events occur. Detection results are improved over baseline F1 scores by up to a factor of nine, and achieve ROC-AUC of between 0.60 and 0.66.

Physiological data from wearable sensors allows for studying individual response to atypical life events in the wild, creating opportunities for testing psychological theory about affect and experience. In addition, sensors data creates the possibility of passive monitoring to detect when individuals have stressful or negative experiences, while preserving the privacy of these events from the organization. Informed consent of workers, and strong data protections will still be necessary, but organizations can improve the health and well-being of their workforce and reduce their detrimental effects on vulnerable populations through detecting such experiences.

2 Related Work

In this paper, we explore the effect of positive and negative events on human behavior, and how to detect these events with wearable sensors. There exists extensive research on how sensors can be used to detect patterns and changes in human behavior , including psychological constructs such as stress, anxiety, and affect (cf. literature review of wearable sensors [1]). These papers most often explore detection of stress either induced exogenously (cf. [12]), or (as in our work) endogenously (cf. [3]). There is also literature on detecting bio-markers associated with other psychological constructs, such as anxiety [13], positive and negative affect [21], and depression [4].

Past research has often suffered from two limitations. First, research has focused on either short time intervals (up to two weeks) and very small sample sizes (on the order of tens of subjects) [3,12,18], or collected data sporadically (once every several months) [6]. Second, previous literature has typically detected very short-term stresses (e.g., stresses that affect people on minute level [3,12,18]) rather than individual stressful events that impact someone over the longer term, such as funerals. Our work differs from these previous studies through continuous evaluation over several weeks of hundreds of subjects, allowing us to robustly uncover effects in diverse populations. Moreover, we uncover patterns associated with unusually good or bad events that can affect multiple psychological constructs over multiple days.

3 Data

The data used in this paper comes from two studies aimed at understanding the relationship between individual variables, job performance, and wellness [17]. The studies took place in high-stress environments, but with diverse workforces. Conclusions and methods that generalize across these studies offer hope these results can generalize to a variety of workplaces. Both studies had similar longitudinal design and collected similar data, despite being conducted in different locations and recruited different populations. The *hospital* workforce data was collected during a ten-week long study that recruited 212 hospital workers. The *aerospace* workforce data was collected from 264 subjects, and most details of their data collection, including survey questions asked and the use of wearable devices, match [17]. The studies administered daily surveys to collect self-assessments of participant stress, sleep, job performance, organizational behavior, and other personality constructs. We focus on positive affect, negative affect, anxiety and stress, which we discuss in greater detail in the psychological construct section.

In this paper, we use data collected from *Fitbit* wristbands. We focus on this modality since it was common to both studies. The Fitbit wristband captures dynamic heart rate and step count, and offers daily summary data based on heartrate and step count that includes: daily minutes in bed, daily minutes asleep, daily sleep efficiency, sleep start and end time, and time spent in “fat burn,” “cardio,” or “out of range” heartrate zones.

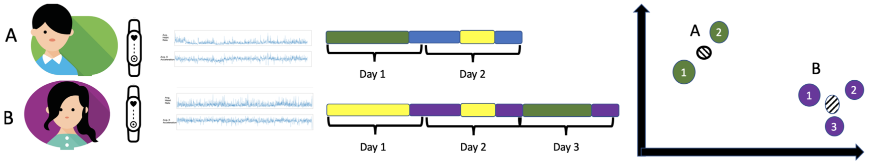

Fig. 1. Overview of the modeling framework. Sensor data collected from participants A and B (left two panels) is fed into a non-parametric HMM model that outputs state sequences (middle panel). Output from the HMM model is used to learn embeddings for each day of each participants (right panel). The daily embedding (lighter and darker-colored circles) and the average embedding for each participant (hashed circles) are used as features. These features, and daily atypical event labels, are then fed into an SVM classifier to predict whether any given day is atypical.

Psychological Constructs. The data used for this study includes daily self-assessments of psychological states provided by subjects over the course of the study. These constructs include self-assessments of stress, anxiety, positive affect, and negative affect, which were found to significantly change during an atypical event. In contrast, job performance, personality, alcohol and tobacco use, and sleep quality did not significantly change. Stress and anxiety were measured by responses to questions that read, "Overall, how would you rate your current level of stress?" and "Please select the response that shows how anxious you feel at the moment" respectively and have a range of 1–5. Positive and negative affect were measured based on 10 questions from [16] (five questions for measuring positive affect and five for measuring negative affect) and have a range of 5–25. **Atypical Event** In addition to these constructs, subjects were also asked, "Have any atypical events happened today or are expected to happen?". If subjects replied yes, they had the option of add free-form text describing the atypical event. In the hospital dataset, there are 8,155 days of recorded data, of which 958 days have atypical events (11.7%). The aerospace dataset has 10,057 days of data, of which 1,503 are considered atypical (14.9%). We have access to the free-form text in the hospital data, which was filled out by participants in 87% of all atypical events, and is categorized as positive, negative, or very negative events [17]. Surprisingly, the severity of the event could not be easily gleaned from sentiment analysis, such as VADER [14], as these tools gave neutral sentiment to text samples that were clearly negative (e.g., "at a funeral"). Of all categorized atypical events, 210 (24%) were positive, 626 (71%) were minor negative events, and 39 (4.5%) were major negative events.

4 Methods

4.1 Causal Inference Method

Atypical events are often described in free-form text as exogenous, e.g., an injured family member or unusually heavy traffic. We can therefore conjecture that atypical events create an as-if random assignment of any given subject over time (these events typically happen at random). This is not always true, as in the case of subjects who report being on vacation multiple days, or are at different stages of burying a loved one. These are, however, relatively rare instances, with sequential events occurring in less than 15% of atypical events in either dataset and exclusion of this data does not significantly affect results. To determine the effect of atypical events on subjects, we use a difference-in-difference approach to causal inference. Specifically, we look at all subjects who report an atypical event and then look at a subset who report stress, anxiety, negative affect, or positive affect the prior day (83% of all events). We finally take the difference in their self-reported constructs from the day before the event. If subjects report construct values after the event we report the difference between these values and the day prior to an atypical event. We contrast these measurements with a *null model*, in which we find subjects who did not report an atypical event on the same days that other subjects reported an atypical event, and find the change in their construct values from the prior day. The difference between construct values associated with the event and the null model is the *average treatment effect* (ATE).

4.2 Representation Learning

We detect atypical events by embedding individuals' heartrate and step count time series data into a vector space, using the framework proposed in [19]. We then train models to identify where in this space atypical events occur. Based on [19], each subject's physiological data is interpreted as a multivariate time series, as described in Fig. 1, left panels. The time series are transformed into sequences of hidden Markov states using a Beta Process Auto Regressive HMM (BP-AR-HMM) [7] (Fig. 1, center panel). Unlike classical hidden Markov models, BP-AR-HMM is flexible by allowing the number of hidden states to be inferred from the data. Based on these datasets the model found 73 states in the hospital data, and 130 states in the aerospace dataset, i.e., we find 73–130 general states/activities that subjects perform, although a subset of these activities are observed in a day. These states are shared among all subjects, rather than being subject specific. This makes it feasible to embed data across different subjects and across different days. After the states are learned, we calculate the stationary distribution of time spent in each state to embed the daily data into the activity space (Fig. 1, right panel). This can be calculated from the HMM transition matrix by finding the eigenvector corresponding to the largest eigenvalue of the matrix.

5 Results

Causal Effect of Atypical Events. How do atypical events affect individual's psychological states? We apply a difference-in-difference approach to measure the impact of atypical events on self-reported psychological constructs. We first look at the effect of atypical events across all our datasets, as shown in Fig. 2.

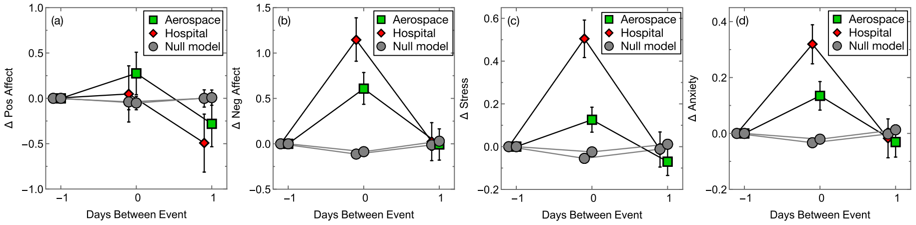

Fig. 2. Effect of atypical events among the datasets studied. (a) Positive affect, (b) negative affect, (c) stress, and (d) anxiety. Green squares and red diamonds show aerospace and hospital datasets, respectively, and gray circles are null models. (Color figure online)

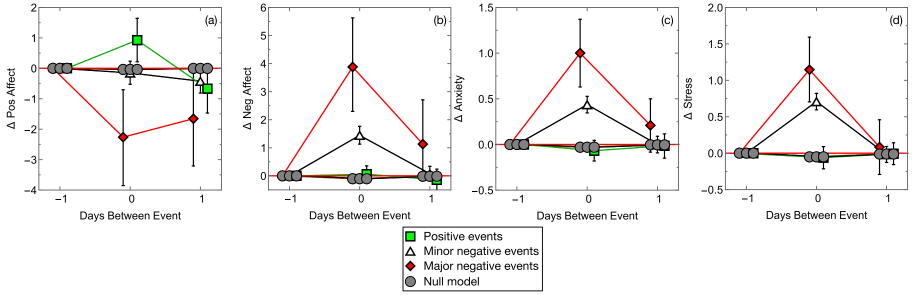

Fig. 3. Effect of different types of atypical events. (a) Positive affect, (b) negative affect, (c) stress, (d) anxiety. Green squares are positive events, white triangles are minor negative events, red diamonds are major negative events, and gray circles are null models. (Color figure online)

Atypical events, on average, have a relatively small effect on positive affect the day of the event (difference from null $= 0.09, 0.33$; p-value$= 0.6, 0.009$, for hospital and aerospace data, respectively). We notice a decrease in positive affect from the day of the event to the day after the event (difference $= 0.54, 0.55$; p-values $= 0.0015, 0.017$ for aerospace and hospital data, respectively). On the other hand, there is a substantial increase in negative affect, stress, and anxiety (p-values < 0.001), although changes are smaller in the aerospace dataset.

For the hospital data, we categorized atypical events as positive, minor negative, or major negative events using free text descriptions. Namely, negative life events (e.g., funerals) were classified as major negative events; sickness, daily hassles (e.g., traffic), and unpleasant work events were classified as minor negative events, and positive life and work events were classified as positive events. Each category (e.g., positive or negative life events) is defined in [17]. Most atypical events are negative, such as a fight with the spouse, traffic, or deaths. Examples of reported positive events include passing a test or a promotion. The effect of different types of atypical events is shown in Fig. 3. We find that positive events increase positive affect (p-value$= 0.009$), but, surprisingly, have no statistically significant effect on negative affect, stress, or anxiety (p-value ≥ 0.3). Minor negative events do not substantially change positive affect on the day of the event (difference from null $= -0.15$, p-value$= 0.57$), and have a small effect on positive affect the day after the event (difference from null $= -0.42$, p-value$= 0.04$). On the other hand, they significantly increase negative affect, anxiety, and stress (p-value < 0.001). Finally, major negative events both decrease positive affect the day of the event and the day after the event (p-value $= 0.005, 0.03$ respectively). These results point to a strong diversity in atypical events, and support the idea that "bad is stronger than good" [2]: adverse, or negative, events have a stronger effect on people than positive events, and are reported as atypical events more often.

Detecting Atypical Events. We evaluate performance of three classification tasks using sensor data: (1) detecting whether an atypical event occurred on that day; (2) detecting whether subjects experienced a good day; or (3) detecting whether subjects experienced a bad day. For (2) and (3) the classification task was "1" if subjects experienced a good or bad day, respectively, and "0" otherwise. Hence we simplify all tasks into a binary detection task. We emphasize that these last two tasks are only available for the hospital data.

We use ten-fold cross-validation. Performance metrics are averaged across all held-out folds. Two type of experiments are presented in the paper. In the first set (Table 1) datapoints are split at random, In the second set (Table 2) subjects are split into training and testing sets to approximate a cold-start scenario, where model is trained on one cohort of subjects and tested on another cohort. The challenge of the latter detection task is that we need to classify if a subject has a good or bad day despite not being trained on any previous data from that subject. We use three performance metrics for evaluation. Area under the receiver operating characteristic (ROC-AUC), F1 score, and precision. These metrics are chosen because the data is highly imbalanced.

Table 1. Performance of atypical event detection from sensors in both datasets with randomly sampled cross-validation.

Dataset	Construct	Model	ROC-AUC	F1	Precision
Hospital workforce	Atypical event	Random	0.50	0.12	0.12
		Aggregated	0.57	0.24	0.15
		Embedding	**0.66**	**0.37**	**0.32**
	Positive event	Random	0.50	0.03	0.03
		Aggregated	**0.63**	0.08	0.04
		Embedding	0.62	**0.27**	**0.30**
	Negative event	Random	0.50	0.08	0.08
		Aggregated	0.57	0.17	0.10
		Embedding	**0.61**	**0.27**	**0.24**
Aerospace workforce	Atypical event	Random	0.50	0.15	0.15
		Aggregated	**0.59**	0.31	0.21
		Embedding	**0.60**	**0.32**	**0.36**

We compare detection quality for two types of models: models using features from statistics of aggregated data, and models using features based on time series embeddings. For the aggregated model, we create several features based on aggregated statistics of signals and static modalities. These statistics included the sum, mean, median, variance, kurtosis, and skewness of signals the day before, the day of, and the day after each day, as well as all FitBit daily summary data. Missing data is substituted with mean statistic value within the training or testing set. Statistics before and after each day were created because some physiological features, such as mean heart rate, might change before an atypical event, and some may change after, such as sleep duration. We use Minimum Redundancy Maximum Relevance [5] to select the best features (23 and 26 for the aerospace and hospital data respectively). Important features in the hospital data relate to sleep (the top feature was tomorrow's minutes in bed). Important features in the aerospace dataset tend to relate to heart rate (the top feature was the number of minutes in the "fat burn" heart rate zone in the past day).

Representations from HMM embeddings were learned for the day of, and the day after each day (no summary features are used). We also include the centroid of embeddings for each person in the training data as features, to control for subject-specific differences in behavior. We did not use any additional feature selection because embedding naturally reduces the feature dimensions. Imputation is also not needed because the HMM learns states based on the amount of data available for that day.

We train several candidate classification methods using aggregate features: logistic regression, random forest, support vector machines, extra trees [9], AdaBoost [8], and multi-layered perceptrons. The majority class (no atypical event) is downsampled such that the number of datapoints in each class are equal, which improves the models over using the raw data or upsampling the minority class Based on ten-fold cross-validated F1, ROC-AUC, and precision, we find atypical events in the hospital dataset are best modeled with random forests, while the aerospace workforce dataset is best modeled with logistic regression. In comparison, positive events are best modeled with random forests but negative events are best modeled with extra trees. Finally for embedding features, we used the SVM classifier with no down-sampling, which follows the original work [19].

We demonstrate the first set of results in Table 1. We find that the HMM embedding-based model outperforms other models. The ROC-AUC for the HMM-based model is 0.60 for the aerospace workforce and 0.66 for the hospital workforce. Positive and negative events similarly have an ROC-AUC of 0.61–0.63. F1 and precision exceed random baselines by factors of two to nine. The seemingly low F1 and precision are due to the rarity of atypical events, especially for positive events, which only happen on 3% of days, and negative events which only happen in 8% of all days. A detection therefore represents a "warning sign" that a worker may have had an negative event that day. Overall, detecting atypical events shows promise. Alternatively, a model may be trained on one dataset and tested on another (cold-start scenario). These results are presented in Table 2. Atypical events can be detected 91–220% above baselines based on

F1 score, but results are more modest than in the Table 1, with a reduction in ROC-AUC from 0.66 to 0.58 for hospital atypical events. These results are alike to other recent papers, which split subjects into training and testing and found relatively poor model performance (cf. [18]). These results suggest that models will perform best when personalized to subjects or transfer learning methods are developed for these data.

Table 2. Performance of atypical event detection from sensors in both datasets with subject held-out detection.

Dataset	Construct	Model	ROC-AUC	F1	Precision
Hospital workforce	Atypical event	Random	0.50	0.12	0.12
		Aggregated	0.55	0.22	0.14
		Embedding	**0.56**	**0.23**	**0.16**
	Positive event	Random	0.50	0.03	0.03
		Aggregated	0.57	0.065	0.035
		Embedding	**0.58**	**0.08**	**0.05**
	Negative event	Random	0.50	0.08	0.08
		Aggregated	**0.57**	0.15	0.09
		Embedding	0.56	**0.16**	**0.10**
Aerospace workforce	Atypical event	Random	0.50	0.15	0.15
		Aggregated	**0.58**	**0.30**	**0.20**
		Embedding	0.54	0.25	0.17

6 Conclusion

We discover that atypical events and negative events substantially increase stress, anxiety, and negative affect. Major negative events are found to reduce positive affect over multiple days, while positive events improve positive affect that day. We also demonstrate that wearable sensors can provide important clues about whether someone is experiencing a positive or negative event. We find atypical events can be predicted with ROC-AUC of up to 0.66 with relatively little hyperparameter tuning. More improvements are therefore possible to predict atypical events. These results point to the importance and detectability of atypical events, which offer hope for remote sensing and automated interventions in the future.

Acknowledgements. The research was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800005.

References

1. Banaee, H., Ahmed, M.U., Loutfi, A.: Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. Sensors **13**(12), 17472–17500 (2013)
2. Baumeister, R., Bratslavsky, E., Finkenauer, C., Vohs, K.: Bad is stronger than good. Rev. Gene. Psychol. **5**, 323–370 (2001)
3. Can, Y.S., Chalabianloo, N., Ekiz, D., Ersoy, C.: Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study. Sensors **19**(8), 1849 (2019)
4. Canzian, L., Musolesi, M.: Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: UbiComp (2015), pp. 1293–1304. ACM, New York, USA (2015)
5. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. **3**(02), 185–205 (2005)
6. Edwards, D., Burnard, P., Bennett, K., Hebden, U.: A longitudinal study of stress and self-esteem in student nurses. Nurse Educ. Today **30**(1), 78–84 (2010)
7. Fox, E.B., et al.: Joint modeling of multiple time series via the beta process with application to motion capture segmentation. Ann. Appl. Stat. **8**(3), 1281–1313 (2014)
8. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
9. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)
10. Goetzel, R.Z., et al.: Ten modifiable health risk factors are linked to more than one-fifth of employer-employee health care spending. Health Aff. **31**(11), 2474–2484 (2012)
11. Gray-Toft, P., Anderson, J.G.: Stress among hospital nursing staff: its causes and effects. Soc. Sci. Med. A **15**(5), 639–647 (1981)
12. Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans. Intell. Transp. Syst. **6**(2), 156–166 (June 2005)
13. Huang, Y., et al..: Discovery of behavioral markers of social anxiety from smartphone sensor data. In: DigitalBiomarkers (2017), pp. 9–14 (2017)
14. Hutto, C.J., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)
15. Kyriakou, K.: Detecting moments of stress from measurements of wearable physiological sensors. Sensors **19**(17),(2019)
16. Mackinnon, A., Jorm, A.F., Christensen, H., Korten, A.E., Jacomb, P.A., Rodgers, B.: A short form of the positive and negative affect schedule: evaluation of factorial validity and invariance across demographic variables in a community sample. Person. Individ. Diff. **27**(3), 405–416 (1999)
17. Mundnich, K.: Tiles-2018: a longitudinal physiologic and behavioral data set of hospital workers. Sci. Data **7**(354) (2020)
18. Smets, E.: Large-scale wearable data reveal digital phenotypes for daily-life stress detection. npj Digit. Med. **1**(67) (2018)
19. Tavabi, N.: Learning behavioral representations from wearable sensors. arXiv preprint arXiv:1911.06959 (2019)
20. Varian, H.R.: Causal inference in economics and marketing. PNAS **113**(27), 7310–7315 (2016). https://doi.org/10.1073/pnas.1510479113

21. Yan, S., Hosseinmardi, H., Kao, H., Narayanan, S., Lerman, K., Ferrara, E.: Estimating individualized daily self-reported affect with wearable sensors. ICHI **2019**, 1–9 (2019). https://doi.org/10.1109/ICHI.2019.8904691
22. Zamkah, A., Hui, T., Andrews, S., Dey, N., Shi, F., Sherratt, R.S.: Identification of suitable biomarkers for stress and emotion detection for future personal affective wearable sensors. Biosensors **10**(4), 40 (2020)