Detection of Subtle Stress Episodes During UX Evaluation: Assessing the Performance of the WESAD Bio-Signals Dataset

Alexandros Liapis, Evanthia Faliagka, Christos Katsanos, Christos Antonopoulos, Nikolaos Voros

Detection of Subtle Stress Episodes During UX Evaluation: Assessing the Performance of the WESAD Bio-signals Dataset

Alexandros Liapis[1,2], Evanthia Faliagka[1], Christos Katsanos[3], Christos Antonopoulos[1], Nikolaos Voros[1]

[1]Department of Electrical and Computer Engineering, University of Peloponnese, Patras, Greece
{a.liapis, e.faliagka}@esda-lab.gr {ch.antonop, voros}@uop.gr

[2]Hellenic Open University, School of Science and Technology, Patras, Greece

[3] Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
ckatsanos@csd.auth.gr

Abstract. Stress is a highly subjective condition and may largely vary in different contexts. Bio-signals have been widely used by researchers and practitioners to monitor stress levels. Consequently, various bio-signals datasets for stress recognition have been recorded. The most of publicly available physiological datasets have been emotionally annotated in a context where users have been exposed to intense stressors, such as movie clips, songs, major hardware/software failures, image datasets, and gaming. However, it remains unexplored how effectively such datasets can be used in different contexts. This paper investigates the performance of the publicly available dataset named WESAD (Wearable Stress and Affect Detection) in the context of UX evaluation. More specifically, skin conductance signal from WESAD was used to train four machine learning classifiers. Regarding the binary classification problem (stress vs. no stress), models' accuracy was rather high (at least 91.1%). However, it was found that their effectiveness in assessing stress in the context of UX was rather poor when a new bio-signals dataset was used.

Keywords: Stress Detection, UX Evaluation, Bio-signals, Electrodermal Activity.

1 Introduction

The combined research efforts of various fields, such as Human-Computer Interaction, Ubiquitous Computing, Ambient Intelligence and Internet of Things, have substantially increased the interest on affective qualities of software products [1]. User eXperience (UX) emerged as a research field allowing HCI researchers and practitioners to better understand users' interaction experiences by using tools and techniques beyond traditional user interaction metrics [2]. As a term, UX encompasses aspects such as usability, usefulness, aesthetics and emotions [3]. UX design often begins before the product is

even in the user's hands. Designing and developing for UX requires a deep understanding of how users feel during their interaction with a system or a product [4].

Emotional aspects of UX can be measured by using a variety of approaches, such as post-questionnaires, interviews and observation. However, these methods have been criticized as time consuming and prone to subjectivity [5]. Alternatively, modalities such as facial expression [6], speech tone [7] and touchscreen patterns analysis [8, 9] have been proposed to minimize any subjectivity effect. Towards the same direction, bio-signals monitoring (e.g., heart rate, respiration, skin conductance) is also an approach that has been adopted by researchers in the context of UX evaluation [10].

In UX evaluation of interactive environments, one is mostly interested in the identification of system flaws [11]. A system or a product with flaws can cause undesirable activations of users' physiology widely referred as "*fight or flight*" event or stress [12].

Skin Conductivity (SC), also known as Galvanic Skin Response (GSR), is one of the most well-studied psychophysiological markers of the functioning of people's Autonomic Nervous System. The evolution of appealing wearables [13] has further transformed SC into a popular measurement allowing experiments to take place in more ecologically valid settings [14] at a relatively low cost [15]. SC signal characteristics, such as peak height and instantaneous peak rate, are reliable indicators of stress level of a user. In [16] an extensive summary of SC research in relation to stress is presented.

There is a large number of publicly available physiological datasets [17–19] for stress research that have been emotionally annotated in a context where users have been exposed to intense stressors (e.g., movie clips, songs, major hardware/software failures, image datasets, and gaming). Although such approaches are able to create stress prediction models with rather high classification accuracy, it remains questionable if they could be effectively used in capturing subtle stress responses, which are mostly expected in UX evaluation studies [20].

This paper investigates the performance of such a dataset in the context of UX evaluation. To the best of our knowledge, this is the first paper to do so. More specifically SC signals of 15 users from the publicly available physiological dataset named WESAD (Wearable Stress and Affect Detection [21]) were used in order to train four machine learning classifiers (L-SVM, C, SCM, Q-SVM and sTree). Next, a publicly[1] available emotionally annotated bio-signals dataset, made available by Liapis et al. [22], was considered as the ground truth dataset in order to evaluate the performance of the aforementioned classifiers in stress detection in the context of UX evaluation. This ground truth dataset consists of SC segments that have been emotionally annotated by users' valence-arousal ratings. Such self-reported periods indicate usability issues confronted while they were interacting with a web-based platform during a UX evaluation study. Using users' self-reporting as ground truth is a common practice in UX research [23, 24].

[1] https://www.researchgate.net/publication/350855591_EDA_Dataset_in_UX_Context.rar

2 Description of Datasets

2.1 WESAD Dataset

WESAD [21] is a publicly available multimodal physiological dataset for wearable stress and affect detection. It includes the following physiological signals: blood volume pulse (BVP), electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG), respiration (RESP), body temperature (TEMP), and three axis acceleration (ACC). The bio-signals were recorded during a lab study in which 15 participants with a mean age of 27.5 years (SD=2.4) were exposed in three different affective states: neutral, stress, and amusement. The Trier Social Stress Test (TSST) was employed by the researchers in order to elicit stress. Regarding the binary classification problem (stress vs. non-stress), a classification accuracy of up to 93% was reported when all physiological signals participated in the training process. Classification was also conducted by using only EDA data. In this case, the accuracy was 80%.

Effective identification of an emotional state requires the recording of adequate bio-signals, under well-organized experimental conditions (field or lab) by using the appropriate sensing equipment. However, the number of recorded signals affects the number of sensors that are required. On the contrary, Liu et al. [25] used a single bio-signal in order to create a more practical, unobtrusive and comfortable wearable system for stress detection. In particular, the SC signal along with Linear Discriminant Analysis (LDA) were used in order to discriminate three stress levels: low, medium and high. A classification accuracy of 81.82% was achieved. In addition, Jussilla et al. [26] proposed an effective stress management bio-sensor named "smart ring". Smart ring measures EDA from the palmar side of the wearers. Such approaches might be a better tradeoff between recognition performance and computational load, which in turn could be a promising line of research for the development of practical personal stress monitors. These studies are our rationale to use only SC in the present study.

Despite the high classification accuracy in both approaches (all signals vs. EDA), WESAD authors indicate that "*results should be interpreted with caution due to the limitations of WESAD, regarding the number of subjects and the lack of age and gender diversity*". Furthermore, the authors invite the research community to consider their dataset for algorithm development and benchmarking, which is an objective of this paper.

2.2 Ground Truth Dataset

In the section, we present the ground truth dataset that was used to assess the stress detection mechanisms created from the WESAD dataset. More specifically, the ground-truth dataset consists of SC segments that have been emotionally annotated by users' self-reported valence-arousal ratings. Such periods indicate usability issues confronted while users were interacting with a platform during a UX evaluation study [22].

More specifically the aforementioned study involved 30 participants (13 female), aged between 18 and 45 (Mean=32.1, SD=7.1) who were asked to complete a set of interaction tasks in a web-based service while their SC signal was recorded. At the end

of each user testing session each participant was involved in a retrospective think aloud (RTA) protocol in order to report any usability issues (UIs) that she/he confronted while performing the interaction tasks. For each one of his/her retrospectively reported UIs, the participant was asked to provide: a) the duration of the confronted UI and b) an emotional rating, using the emotional scale of Valence (from 1 to 9)–Arousal (from 1 to 9). Overall, a number of 113 emotionally annotated UIs were reported. For each annotated UI there is an associated segment of SC signal that constitutes the ground truth bio-signals dataset that are used in the present study to test the classifiers created from the WESAD dataset.

3 Classifiers Creation: Training Process and Results

Non-Specific Skin Conductance Responses (NS-SCRs) from the SC signals included in the WESAD dataset were used as the training dataset for the development of the stress classifiers. The use of intensive sub-periods that might appear within an emotional period can probably contribute to the final assessment of the experienced emotion (i.e., feeling stressed, happy, angry etc.). In terms of stress detection, intensive sub-periods could be interpreted as NS-SCRs within a stress session. A validated software named PhysiOBS [22, 27], freely available, was used to detect and extract NS-SCRs.

For each SC signal in the WESAD dataset, we used only the signal part that is associated with the stress sessions (TSST) as already mentioned in section 2.1. The detected significant NS-SCRs segments within each TSST session were used as the stress class and the rest parts of the TSST session were used as the non-stress class (see Fig. 1). This specific dataset creation approach has been also applied in [28].

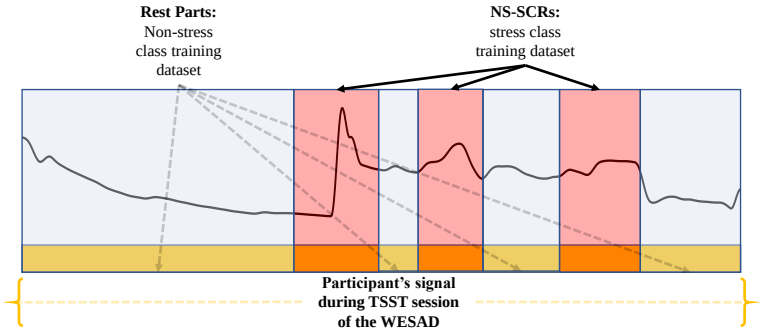

Fig. 1. An example of the dataset creation process. From each SC signal in the WESAD dataset we extract NS-SCRs and use them as the stress class. The rest parts of the session constitute the non-stress class.

More specifically, the NS-SCRs extraction process consists of the following steps. Initially, the SC signals were smoothed using Hann function and then normalized as proposed in [29]. Subsequently, we used the signal amplitude from which we extracted 21 features as proposed in [30]. Next, the extracted features were provided as input to

the selected machine learning algorithms aiming to differentiate the two emotional states (stress vs non-stress). A 5-fold cross-validation training was applied in all classification methods. Overall, the training dataset consisted of 380 cases; 165 in the class stress (NS-SCRs) and 215 in the class non-stress (signal rest parts). As proposed in [16], NS-SCR's segments with duration larger or equal to 4 seconds from NS-SCR's initial deflection to peak were considered; a rule also applied in [31].

Regarding the binary problem (stress vs non-stress), Table 1 presents the obtained performance metric for each trained classifier. All classifiers achieved high accuracies (at least 91%). The best classification result was achieved by the sTree classifier (95.8%). These results indicate that our applied training method improved the classification results compared to the 80% accuracy reported by [21] when using only the SC signal. Furthermore, the confusion matrix (see Fig. 2) presents details about the correctly classified cases per trained model.

Table 1. Performance for each classifier (F1 score was also calculated). The F1-score is an important metric when there are imbalanced classes as in our case.

	Precision	**Recall**	**Accuracy**	**F1-Score**
C-SVM	89,7%	89,7%	**91,1%**	89,7%
L-SVM	92,6%	91,5%	**93,2%**	92,1%
Q-SVM	92,4%	88,5%	**91,8%**	90,4%
sTree	96,3%	93,9%	**95,8%**	95,1%

		Predicted Classes / model							
		C-SVM		*L-SVM*		*Q-SVM*		*sTree*	
		Stress	*No Stress*	*Stress*	*No Stress*	*Stress*	*No Stress*	*Stress*	*No Stress*
True Class	*Stress*	148	17	151	14	146	19	155	10
	No Stress	17	198	12	203	12	203	6	209

Fig. 2. Confusion matrix. Overall, the training dataset consisted of 380 cases; 165 in the class stress and 215 in the class no-stress. Green parts show the correctly classified cases for each classifier.

The plot of sensitivity versus 1-Specifity is called Receiver Operating Characteristic (ROC) curve and the area under this ROC curve is called Area Under the Curve (AUC) (see Fig. 3). Both ROC and AUC are effective measures of accuracy. This curve plays a central role in evaluating diagnostic ability of tests to discriminate the true state of subjects. In our case, the AUC can be interpreted as the probability that a randomly chosen stress signal is rated or ranked as more likely to be stress than a randomly chosen non-stress signal. All classifiers achieved high AUC (at least 94%). The best AUC result was achieved by the L-SVM classifier (98%).

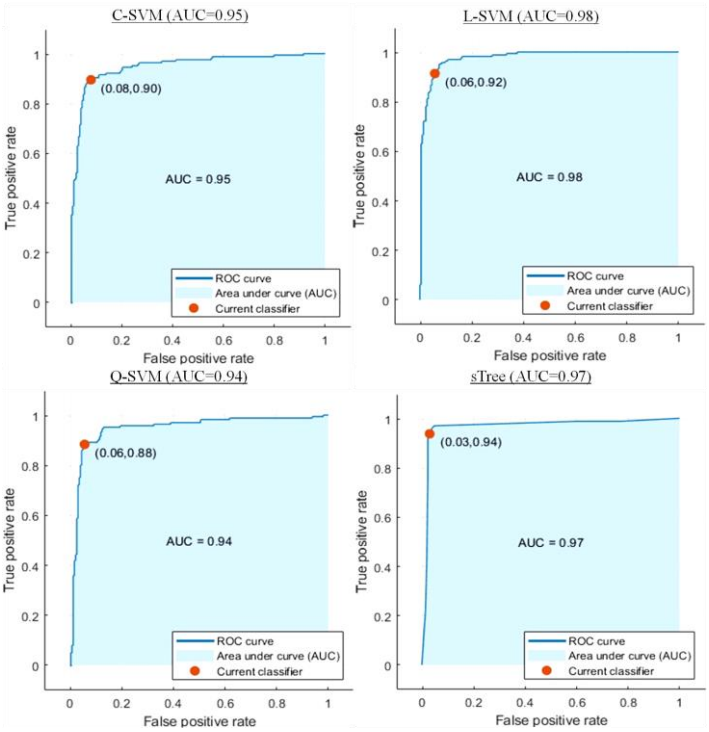

Fig. 3. Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) performance for each trained model. Red dot in each plot shows classifiers' optimal performance between True Positive Rate (TPR) and False Positive Rate (FPR).

4 Evaluation of Classifiers

In section 3 we presented the training process and the results of four classifiers. To this end, the SC signals from the publicly available dataset named WESAD were used. In order to measure the performance of the created models we used an existing UX biosignals dataset (see section 2.2). More specifically, the test dataset consists of 113 emotionally annotated (according to VA ratings) user-reported SC segments. Kappa coefficient [32] metric was used to quantify the agreement between users' emotional ratings and created classifiers. Agreement among raters ranges from -1 to 1. Values near or below zero suggest that the agreement is probably attributable to chance. On the contrary, the higher the positive value of Kappa is, the higher the agreement is.

Any SC segment with Valence lower than 5 and Arousal greater than 5 was assigned as stress and the rest SC segments as non-stress [29, 33]. Next, the 113 SC segments, were used as an input to the trained classifiers. For each segment each classifier returned the classification result (1= stress, 2=non-stress). The returned values of the stress models were compared with participants' self-reported stress ratings that constitutes our ground truth dataset. Table 2 presents the interrater reliability for each classifier.

According to the levels of agreement presented in [34], the Q-SVM achieved a non-significant slight agreement; Kappa = 0.17, p>0.05, 95% CI [-0.01, 0.35]. Considering the 95% confidence interval (CI), it was found that the agreement ranged between poor and fair. The rest three classifiers (C-SVM, L-SVM, sTree) had Kappa values very close to zero, which means that there was no agreement at all.

Table 2. Interrater reliability (IRR) values and confidence interval (CI) 95% for each classifier.

Trained Model	Kappa Value	95% CI
C-SVM	-0.02	[-0.20, 0.16]
L-SVM	0.02	[-0,16, 0,20]
Q-SVM	0.17	[-0.01, 0.35]
sTree	-0.06	[-0.24, 0.13]

5 Conclusions, limitations and future work

This paper investigates the performance of the multimodal WESAD dataset in the context of UX evaluation. We used only the SC signal from the WESAD dataset because it is a reliable indicator of stress. Regarding the binary classification problem (stress vs non-stress) accuracy of up to 95.8% was reached.

An existing bio-signals dataset, which consists of SC segments, was used as the ground truth dataset in order to assess the performance of the stress classifiers in the context of UX evaluation. In contrast with the training dataset (WESAD), the ground truth dataset was recorded during a UX evaluation study, a context that could cause subtle emotional reaction. More specifically, the ground truth dataset represents users' self-reported periods of usability issues confronted while they were interacting with a web-based platform. We assessed the performance of the stress classifiers by conducting an interrater reliability analysis using the Kappa coefficient. The higher interrater reliability was found for Q-SVM but it was still non-statistically significant and poor-to-fair; Kappa = 0.17, p>0.05. The aforementioned level of agreement is quite lower than the one presented in [22]. In the latter the same ground truth dataset was used. The reported interrater reliability was found to be statistically significant and fair-to-moderate; Kappa = 0.35, p<0.001. This is probably explained by the fact that in [22] the stress classifier was assessed against the ground truth dataset which was trained with bio-signals that had been recorded while users performed typical HCI tasks [35].

In this study we assessed the classifiers performance by using only skin conductance. Such an approach aims to maximize practicality by reducing the number of sensors while maintaining accuracy in high levels. Although classification results of the trained models were high, future efforts could consider more bio-signals and different combinations of them in the context of optimal balance vs efficiency. Furthermore, this study serves as a first proof of concept by investigating if a dataset emotionally annotated in a context where users have been exposed to intense stressors can indeed be used effectively in a different context (i.e., UX evaluation). In the next steps of our work more participants and additional datasets will be included to further increase objectivity and

accuracy of presented results. Additional approaches such as subject dependent training along with more in-depth analysis techniques such as deep learning should also be investigated. Creating more flexible stress assessment mechanism analysis combining bio-signals dataset from various contexts could be also a challenge for future work.

Overall, the preliminary results presented in this paper reveal that performance of the available bio-signal datasets in various contexts should be carefully taken into consideration. Although the one size fits all approach is not suggested, this study provides some interesting insights on the generalizability of the bio-signals datasets.

Acknowledgments. This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH CREATE INNOVATE (project code: T2EΔK-02159 PLan-V).

References

1. Sarsenbayeva, Z., Marini, G., van Berkel, N., Luo, C., Jiang, W., Yang, K., Wadley, G., Dingler, T., Kostakos, V., Goncalves, J.: Does Smartphone Use Drive our Emotions or vice versa? A Causal Analysis. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–15. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3313831.3376163.
2. Remy, C., Bates, O., Dix, A., Thomas, V., Hazas, M., Friday, A., Huang, E.M.: Evaluation Beyond Usability: Validating Sustainable HCI Research. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. p. 216:1-216:14. ACM, New York, NY, USA (2018). https://doi.org/10.1145/3173574.3173790.
3. Silvennoinen, J.M., Jokinen, J.P.P.: Aesthetic Appeal and Visual Usability in Four Icon Design Eras. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 4390–4400. Association for Computing Machinery, San Jose, California, USA (2016). https://doi.org/10.1145/2858036.2858462.
4. Díaz-Oreiro, I., López, G., Quesada, L., Guerrero, L.A.: Standardized Questionnaires for User Experience Evaluation: A Systematic Literature Review. Proceedings. 31, 14 (2019). https://doi.org/10.3390/proceedings2019031014.
5. Marshall, C., Rossman, G.B.: Designing qualitative research. Sage publications (2014).
6. Tarnowski, P., Kołodziej, M., Majkowski, A., Rak, R.J.: Emotion recognition using facial expressions. Procedia Computer Science. 108, 1175–1184 (2017).
7. Mao, Q., Dong, M., Huang, Z., Zhan, Y.: Learning salient features for speech emotion recognition using convolutional neural networks. IEEE transactions on multimedia. 16, 2203–2213 (2014).
8. Tikadar, S., Bhattacharya, S.: A Novel Method to Build and Validate an Affective State Prediction Model from Touch-Typing. In: Lamas, D., Loizides, F., Nacke, L., Petrie, H., Winckler, M., and Zaphiris, P. (eds.) Human-Computer Interaction – INTERACT 2019. pp. 99–119. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-29390-1_6.
9. Tikadar, S., Kazipeta, S., Ganji, C., Bhattacharya, S.: A Minimalist Approach for Identifying Affective States for Mobile Interaction Design. In: Human-Computer Interaction - INTERACT 2017. pp. 3–12. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67744-6_1.

10. Maier, M., Marouane, C., Elsner, D.: DeepFlow: Detecting Optimal User Experience From Physiological Data Using Deep Neural Networks. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. pp. 2108–2110. International Foundation for Autonomous Agents and Multiagent Systems, Montreal QC, Canada (2019).
11. Lazar, J., Feng, J.H., Hochheiser, H.: Research Methods in Human-Computer Interaction. John Wiley & Sons (2010).
12. Hernandez, J., Paredes, P., Roseway, A., Czerwinski, M.: Under Pressure: Sensing Stress of Computer Users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 51–60. ACM, New York, NY, USA (2014). https://doi.org/10.1145/2556288.2557165.
13. Lee, H., Kleinsmith, A.: Public Speaking Anxiety in a Real Classroom: Towards Developing a Reflection System. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–6. Association for Computing Machinery, Glasgow, Scotland Uk (2019). https://doi.org/10.1145/3290607.3312875.
14. Betella, A., Zucca, R., Cetnarski, R., Greco, A., Lanatà, A., Mazzei, D., Tognetti, A., Arsiwalla, X.D., Omedas, P., De Rossi, D.: Inference of human affective states from psychophysiological measurements extracted under ecologically valid conditions. Frontiers in neuroscience. 8, 286 (2014).
15. Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., Barral, O., Kosunen, I., Valtonen, T., Huotilainen, M.: The psychophysiology primer: a guide to methods and a broad review with a focus on human–computer interaction. Foundations and Trends® in Human–Computer Interaction. 9, 151–308 (2016).
16. Boucsein, W.: Electrodermal Activity. Springer US (2012).
17. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: DEAP: A Database for Emotion Analysis ;Using Physiological Signals. IEEE Transactions on Affective Computing. 3, 18–31 (2012). https://doi.org/10.1109/T-AFFC.2011.15.
18. Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M.A., Kraaij, W.: The SWELL Knowledge Work Dataset for Stress and User Modeling Research. In: Proceedings of the 16th International Conference on Multimodal Interaction. pp. 291–298. ACM, New York, NY, USA (2014). https://doi.org/10.1145/2663204.2663257.
19. Subramanian, R., Wache, J., Abadi, M.K., Vieriu, R.L., Winkler, S., Sebe, N.: ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. IEEE Transactions on Affective Computing. 9, 147–160 (2018). https://doi.org/10.1109/TAFFC.2016.2625250.
20. Alberdi, A., Aztiria, A., Basarab, A.: Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. Journal of biomedical informatics. 59, 49–75 (2016).
21. 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. ACM, New York, NY, USA (2018). https://doi.org/10.1145/3242969.3242985.
22. Liapis, A., Katsanos, C., Karousos, N., Xenos, M., Orphanoudakis, T.: User Experience Evaluation: A Validation Study of a Tool-based Approach for Automatic Stress Detection Using Physiological Signals. International Journal of Human–Computer Interaction. 37, 470–483 (2021). https://doi.org/10.1080/10447318.2020.1825205.
23. Pakarinen, T., Pietilä, J., Nieminen, H.: Prediction of Self-Perceived Stress and Arousal Based on Electrodermal Activity*. In: 2019 41st Annual International Conference of the

IEEE Engineering in Medicine and Biology Society (EMBC). pp. 2191–2195 (2019). https://doi.org/10.1109/EMBC.2019.8857621.
24. Bruun, A.: It's not complicated: a study of non-specialists analyzing GSR sensor data to detect UX related events. In: Proceedings of the 10th Nordic Conference on Human-Computer Interaction. pp. 170–183. ACM, Oslo Norway (2018). https://doi.org/10.1145/3240167.3240183.
25. Liu, Y., Du, S.: Psychological stress level detection based on electrodermal activity. Behavioural Brain Research. 341, 50–53 (2018). https://doi.org/10.1016/j.bbr.2017.12.021.
26. Jussila, J., Venho, N., Salonius, H., Moilanen, J., Liukkonen, J., Rinnetmäki, M.: Towards ecosystem for research and development of electrodermal activity applications. In: Proceedings of the 22nd International Academic Mindtrek Conference. pp. 79–87. Association for Computing Machinery, Tampere, Finland (2018). https://doi.org/10.1145/3275116.3275141.
27. Liapis, A., Karousos, N., Katsanos, C., Xenos, M.: Evaluating user's emotional experience in HCI: the physiOBS approach. In: Kurosu, M. (ed.) Human-Computer Interaction. Advanced Interaction Modalities and Techniques. pp. 758–767. Springer International Publishing (2014).
28. Liapis, A., Katsanos, C., Karousos, N., Xenos, M., Orphanoudakis, T.: UDSP+: Stress Detection Based on User-reported Emotional Ratings and Wearable Skin Conductance Sensor. In: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers. pp. 125–128. ACM, New York, NY, USA (2019). https://doi.org/10.1145/3341162.3343831.
29. Mandryk, R.L., Atkins, M.S.: A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human-Computer Studies. 65, 329–347 (2007). https://doi.org/10.1016/j.ijhcs.2006.11.011.
30. Healey, J., Picard, R.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems. 6, 156–166 (2005). https://doi.org/10.1109/TITS.2005.848368.
31. Bruun, A., Law, E.L.-C., Heintz, M., Alkly, L.H.A.: Understanding the Relationship Between Frustration and the Severity of Usability Problems: What Can Psychophysiological Data (Not) Tell Us? In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 3975–3987. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858511.
32. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 20, 37–46 (1960). https://doi.org/10.1177/001316446002000104.
33. Liapis, A., Katsanos, C., Sotiropoulos, D.G., Karousos, N., Xenos, M.: Stress in interactive applications: analysis of the valence-arousal space based on physiological signals and self-reported data. Multimed Tools Appl. 76, 5051–5071 (2017). https://doi.org/10.1007/s11042-016-3637-2.
34. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics. 159–174 (1977).
35. Liapis, A., Katsanos, C., Sotiropoulos, D., Xenos, M., Karousos, N.: Recognizing emotions in human computer interaction: studying stress using skin conductance. In: Abascal, J., Barbosa, S., Fetter, M., Gross, T., Palanque, P., and Winckler, M. (eds.) Human-Computer Interaction – INTERACT 2015. pp. 255–262. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-22701-6_18.