

On the Pulse of Requirements Elicitation: Physiological Triggers and Explainability Needs

Hannah Deters^{1,*†}, Jakob Droste^{1†} and Kurt Schneider¹

¹Leibniz University Hannover, Software Engineering Group, Hannover, Germany

Abstract

In a time of increasingly complex software systems, explainability is an emerging software quality aspect that can support users by increasing understandability and providing guidance. If the explanations provided by a system are not appropriate, or if there are too many explanations, users are obstructed rather than supported. The elicitation of explainability requirements is subject to confirmation bias and hypothetical bias, and it often relies on the tacit knowledge of stakeholders. Furthermore, there is a need to identify explainability needs during system runtime, as different users of the same system may have vastly different explainability needs. To address these biases and enable the detection of explainability needs, we propose the observation and analysis of user biometrics during runtime. For instance, we assumed that the need for explanations might correlate with an increased stress level, which could be detected via biometric sensors. In this paper, we report an experiment in which we had nine participants wearing a biometric watch while they navigated a software system that purposefully induced explainability needs. The preliminary results of our experiment indicate that explainability needs may be detected via physiological triggers. In particular, we identified electrodermal activity as a notable indicator for the need for explanations.

Keywords

Explainability, Physiological Triggers, Biometric Sensors, Requirements Engineering

1. Introduction

Explainability is the capability of a system to be explained to its users [1]. In complex software systems, explainability can be used as a means to provide transparency and understandability, aiming to foster user trust [2, 3]. Although the general usefulness of explanations in software has been shown in previous work [2, 4], explainability needs for the same software system differ between stakeholder groups [2, 5]. Indeed, providing inappropriate explanations or too many explanations can confuse and frustrate users [1]. To avoid such a conflict, explainability needs have to be elicited carefully [5].


In: D. Mendez, A. Moreira, J. Horkoff, T. Weyer, M. Daneva, M. Unterkalmsteiner, S. Bühne, J. Hehn, B. Penzenstadler, N. Condori-Fernández, O. Dieste, R. Guizzardi, K. M. Habibullah, A. Perini, A. Susi, S. Abualhaija, C. Arora, D. Dell'Anna, A. Ferrari, S. Ghanavati, F. Dalpiaz, J. Steghöfer, A. Rachmann, J. Gulden, A. Müller, M. Beck, D. Birkmeier, A. Herrmann, P. Mennig, K. Schneider. Joint Proceedings of REFSQ-2024 Workshops, Doctoral Symposium, Posters & Tools Track, and Education and Training Track. Co-located with REFSQ 2024. Winterthur, Switzerland, April 8, 2024.

*Corresponding author.

†These authors contributed equally.

✉ hannah.deters@inf.uni-hannover.de (H. Deters); jakob.droste@inf.uni-hannover.de (J. Droste); kurt.schneider@inf.uni-hannover.de (K. Schneider)

ORCID 0000-0001-9077-7486 (H. Deters); 0000-0001-8746-6329 (J. Droste); 0000-0002-7456-8323 (K. Schneider)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The elicitation of explainability requirements involves a number of biases that are typically not in the focus of requirements engineering research. On one hand, asking stakeholders if they want specific explanations in a given scenario, leads to confirmation bias [5], i.e., they will tend to agree rather than disagree when asked if they want a certain explanation. On the other hand, asking stakeholders if they might want explanations for a system that does not exist yet relies on tacit knowledge of the respondent [6] and leads to hypothetical bias [7], i.e., stakeholders will make more false judgments in hypothetical scenarios. In order to provide users with appropriate explanations when - and only when - they need them, a method to monitor end-users' needs for explanations during system runtime is required.

In this paper, we present the preliminary results of an experiment with nine participants. In particular, we had our participants wear a biometric watch while they navigated a software system that purposefully induced needs for explanations. We recorded and analyzed their interactions with the system as well as their biometrical data for the duration of the experiment. Our preliminary results indicate varying correlations between physiological triggers and the need for explanations. We only found a weak correlation between participants' blood volume pulse and their explainability needs, and we found no correlations involving skin temperature and heart rate. However, electrodermal activity had a notable correlation to our participants' need for explanations. Overall, we find that the interdependence of physiological triggers and explanatory needs warrants further investigation. This includes the examination of further biometrics and more different contexts in which explainability needs may arise.

The rest of this paper is structured as follows: We provide background information and related work for this paper in Section 2. The study design is laid out in Section 3. We present and discuss the results of our experiment in Section 4 and Section 5. The conclusion of this paper and a discussion of future work are found in Section 6.

2. Background and Related Work

2.1. Explainability

Explainability is a non-functional requirement [1] that is typically related to artificial intelligence (AI) systems, within the context of explainable AI (XAI) [4]. In XAI, explainability is commonly understood as a means to provide interpretability [8] by explaining machine learning models, and by reasoning the models' decisions [9]. The goal of providing these explanations to end-users is usually to increase transparency and understandability [2], which can ultimately foster user trust [1]. However, recent research on explainability as a non-functional requirement has revealed that explainability may support goals other than just providing interpretability for AI systems [2]. In particular, explanations may be used to guide users [10], increasing the usability of a system. Another AI-independent example are privacy explanations, which can provide transparency and foster user trust [11].

Explainability requirements must be carefully elicited to meet the individual explanatory needs of the stakeholders [2]. Even within the same stakeholder group of the same software system, individual stakeholders may have vastly different explainability requirements [5]. If the presented explanations are not appropriate for the user, or if too many explanations are provided, users might end up confused and frustrated, rather than empowered [1]. Whether an explanation

is appropriate depends on its addressee [2], but also on the goals of the explainer [12] and on the context within which the system is used [2, 13]. Considering the diverse explanation needs of different addressees, there has been research into personalized explanations [14]. Furthermore, the consideration of explainer goals has been researched by Deters et al. [12]. However, while past works highlight the importance of context for providing explanations [2, 13], there is still a need to research how the need for explanations can be detected within any context, i.e., at system runtime.

2.2. Biases in the Elicitation of Explainability Requirements

In addition to commonly encountered biases of requirements engineering, such as confirmation bias [15], the elicitation of explainability requirements is subject to additional disruptive factors, namely hypothetical bias [7] and tacit knowledge [6]. Chazette et al. [16] researched methods for the elicitation of explainability requirements and found that the most common methods are interviews, personas and questionnaires. They also highlight the prevalence of explainability scenarios, in which participants put themselves into hypothetical software use cases in which explanations might be needed [16].

Explainability requirements elicited via interviews and questionnaires and personas that are based on them may be subject to confirmation bias [5]. If study participants are confronted with a software system (real or hypothetical), and asked if they want explanations for certain situation, they might lean towards asking for more explanations than they actually need [5], i.e., they are biased towards confirming the assumptions of the experimenter. This happens because the study participants are not consciously aware that too many explanations might also lead to negative effects [1], such as an increased cognitive load [17].

Explainability scenarios [16] lead to hypothetical bias [7]. For one, the participants of an explainability scenario might not be actual stakeholders of the software. On top of that, scenarios with a software system that does not exist yet ask their participants to state whether or not they require explanations in a yet unknown situation. Explainability needs typically arise when confusion or frustrations with the software are encountered and an explanation is needed to provide clarity and guidance [1]. Asking stakeholders of a future system where and when they might encounter problems that require explanations is unreasonable and relies heavily on tacit knowledge of the stakeholders [6].

2.3. Biometric Watch

Biometric watches are devices that can record various physiological data. The *empatica* watch “Embrace Plus” used in our experiment records electrodermal activity (EDA), blood volume pulse (BVP), heart rate (HR) and body temperature, among other things. Electrodermal activity generally refers to all electrical phenomena in the skin tissue [18]. The *empatica* watch records the skin conductance level in microSiemens (μS) by using a constant voltage. The EDA value provides indications of the body’s response to stress, temperature or exercise [19, 20, 21]. With these stimuli, the so-called “sudomotor innervation” is increased by the sympathetic nervous system (SNS), which causes the EDA value to increase [19]. The delay between a stimulus and an increase in the EDA value is approximately 2 seconds [18]. BVP describes the changes

in peripheral blood volume [21]. Some features extracted from the BVP are also capable of detecting stress, cognitive load and affect [21]. The HR describes the rate of heartbeats and can be estimated using the BVP [20].

Girardi et al. [22] used the *empatica* watch to find a link between emotions and perceived productivity in development teams. Their work also focused on the values of EDA, BVP and HR. While the study does not deal with the elicitation of requirements, it shows that it is possible to detect emotions with the *empatica* watch. Schmidt et al. [23] conducted a literature review on affect recognition with wearable devices. Affect recognition is the recognition of a person's affective state (e.g. stress) in order to investigate decision-making, psychological well-being or similar [23]. Schmidt et al. [23] analyzed 46 papers to determine which sensors are frequently used for affect detection. They found that about 74% use EDA, 33% use skin temperature and 9% use HR. Many of the reviewed studies focused on analyzing stress levels. To the best of our knowledge, there is no previous research on whether biometric data can be used to assess requirements.

3. Research Design

3.1. Research Questions

Our first research question addresses which of the biometric data types recorded by the *empatica* watch is most appropriate for identifying the need for explanation (RQ1). Our second research question assesses whether this data can realistically be used to identify the need for explanation among users (RQ2).

RQ1 Which biometric parameter is most appropriate for identifying the need for explanations?

RQ2 Is the parameter found in RQ1 capable of supporting the identification of explanation requirements?

3.2. Study Design

We conducted a controlled experiment to test whether certain biometric data is a suitable means to elicit explainability needs. Participants were confronted with tasks designed to trigger a need for explanation. Throughout the study, participants wore the *empatica* watch, which recorded the EDA and BVP. Furthermore, we recorded the screen to verify whether there was a need for explanation at the intended tasks.

3.2.1. Tasks

We used *Microsoft Excel* as a test object. Using *Excel*, participants were asked to complete five tasks, with each five to seven subtasks. Three subtasks intended to trigger the need for explanation. We included enough filler tasks before each stimulus to ensure that the last stimulus was distant enough to avoid past tasks from influencing the current tasks. All tasks that are supposed to trigger a need for explanation are described in Table 1. The first task triggering the need for explanation (E1) was a task that was difficult to perform. This was intended to trigger the need for interaction explanations which explain how to use a system correctly. The second

Table 1

Tasks that aimed at inducing a need for explanation

Task	Triggered type of explanation need
E1	The user knows the task (the goal) but does not know how to perform it with the software. <i>The participant was asked to rotate the text in a field at an angle of 45 degrees.</i>
E2	The user expects a different output from the system than he receives. <i>The participant was asked to create a graph according to instructions. The final result looked different in the instructions than in reality.</i>
E3	The user does not know the terminology used. <i>The participant was asked to create a so-called pivot table. We gave instructions on how to do this but did not explain what a pivot table is.</i>

task triggering explanation need (E2) was designed in such a way that the result provided by *Excel* was different from the expectation we created. E2 aimed at simulating unexpected system behavior, causing the users to need an explanation of how the system arrived at this result. The third task triggering explanation need (E3) contained terms that the participants were unlikely to know. It was therefore intended to trigger the need for terminology explanations.

3.2.2. Data Collection

We collected two types of data. Firstly, biometric data using the *empatica* watch. These data are numerical values as time series, with one value per second. Secondly, we captured screen recordings showing the interaction of the participants while completing the tasks. These recordings are videos between 15 and 45 minutes. The data collected in this study includes personal medical data of our participants and can therefore not be disclosed. Before the actual study, the participants were also asked to complete a questionnaire about their age, gender and previous experience with *Excel*.

3.2.3. Demographics

The participants were acquired through convenience sampling from our personal and professional network. Seven of the nine participants were between 20 and 30 years old and two participants were over 60 years old. There were five female and four male participants. 22% of the participants stated that they had a lot of experience with *Excel*, 67% stated that they had medium experience and 11% stated that they had little experience.

3.2.4. Data Analysis

The first step of the data analysis consisted of checking whether the planned tasks triggered a need for explanation. To do this, we analyzed the videos manually to check for anomalies such as long pauses, unusual mouse movements or incorrect inputs. If a moment with a need for explanation was found, the time of this moment was noted. The second step was to find conspicuous features in the EDA, BVP, HR and temperature data. We processed the biometric

data independently of the previously found points in time. The peaks of the EDA values were identified manually by inspecting the data. There is no direct way to identify the characteristic shape of the EDA peaks with the data the *empatica* watch provides yet. We note that there are toolkits for the assessment of EDA values. For example, Aqajari et al. [24] developed a python toolkit to evaluate EDA values according to certain features. However, they evaluate EDA values in the form of electrodermal resistance measurement in kOhms, while the *empatica* watch outputs the skin conductance level in microSiemens. Since this study is only intended as an initial evaluation of whether the detection of explainability requirements is at all feasible using biometric data, the initial evaluation by manual detection of EDA peaks is sufficient. The HR and skin temperature are also evaluated manually for the time being. To analyze peaks in the BVP we used the generalized ESD test [25]. The ESD test can be used to detect outliers in time series data. The ESD test is more suitable for the BVP values, as it focuses on global outliers rather than specific peak shapes. However, it should still be noted that the ESD is not adjusted to BVP outliers and therefore does not provide an optimal evaluation.

4. Results

The screen recordings revealed that tasks E1 and E2 triggered a need for explanation in over 50% of participants. More precisely, E1 triggered a need for explanation in five out of nine participants and E2 in seven out of nine participants (see Table 2). For the task E3, the recordings did not reveal a need for explanation for any of the participants.

The data points for temperature and heart rate did not allow any insights into possible needs for explanation. The body temperature values that we collected via the *empatica* watch increase constantly without peaks. The HR values, which are calculated from the BVP, jump irregularly without any peaks or other outliers being recognizable. We will therefore only discuss the EDA and BVP values in more detail.

4.1. Recall

Two main findings are presented in Table 2. Firstly, the number of participants who had shown a need for explanation in the respective tasks E1 - E3 are listed. Secondly, Table 2 displays how many of these needs for explanation were recognized for both EDA and BVP values. Five participants had a need for explanation for task E1. For one of these participants, at least one

Table 2

Number of participants with need for explanation and the identified explanation needs by EDA and BVP

	Need for Expl.		BVP		EDA	
	#	%	#	%	#	%
E1	5	56%	1	20%	1	20%
E2	7	78%	6	85%	4	57%
E3	0	0%	-	-	-	-

EDA peak was detected when the need for explanation occurred. This means that for 20% of the participants, the EDA value peaked when an explanation was needed for task E1. The same applies to the outliers of the BVP value. For task E2, seven out of nine participants had a need for explanations. 57% of the participants had at least one EDA peak at the time of need. For the BVP values, an outlier was detected for 85% of participants at the time of the explanation need.

4.2. Precision

Table 3 shows the proportion of correct and incorrect conspicuous features in the EDA and BVP values. “Correct” means that the peak of the EDA value occurred at a point where the participant needed an explanation. Accordingly, “incorrect” means that a peak occurred at a point where the participant did not have any need for explanations. The points at which a need for explanation occurs were previously determined by analyzing the screen recordings. 73% of the detected outliers of the BVP value were false positives. If we keep in mind that there are significantly more points in time where no explanation is required than points in time where an explanation is required, these statistics indicate that the outliers of the BVP values are not completely random, but also not reliable. When analyzing the EDA peaks, only 32% of the values were false positives, indicating that the EDA value is more precise than the BVP value.

Table 3

Proportion of correctly and incorrectly identified need for explanation per participant

Participant	BVP			EDA		
	# Outlier	Correct	Wrong	# Peaks	Correct	Wrong
1	3	0%	100%	0	-	-
2	9	33%	67%	0	-	-
3	20	50%	50%	4	100%	0%
4	18	33%	67%	0	-	-
5	28	14%	86%	4	50%	50%
6	13	23%	77%	4	50%	50%
7	2	50%	50%	2	50%	50%
8	10	0%	100%	0	-	-
9	9	33%	67%	5	80%	20%
Total	112	27%	73%	19	68%	32%

5. Discussion

5.1. Answering the Research Questions

RQ1: Which biometric value is most appropriate for identifying the need for explanations? The body temperature and HR did not allow any conclusions to be drawn about the need for explanations. The BVP data allowed the identification of several outliers. These outliers had a high recall of 0.85 for task E2, but the precision was low at 0.27. This means that many explanatory needs were recognized by the BVP value, but frequent false positives occurred. Although the peaks in the EDA values had a lower recall of 0.57 for E2, the precision

was considerably higher at 0.68. The EDA therefore recognizes slightly less need for explanation than the BVP, but the number of false positives is significantly lower. In a requirements engineering context, the EDA value is therefore the most suitable indicator of the four biometric data analyzed for identifying the need for explanation.

RQ2: Is the parameter found in RQ1 capable of supporting the identification of explanation requirements? At this point, the low precision of EDA means that a high number of users is needed to reliably predict the need for explanations. If the EDA values of several users peak at the same point of use, it is unlikely a false positive. For example, in a setting where users wear biometric watches on a daily basis and provide data for improvement purposes, the EDA value would be a suitable indicator to collect explanatory needs. It is also likely that the precision will increase if a suitable method to evaluate the EDA value automatically is found, as inaccuracies due to manual evaluation are eliminated.

5.2. Threats to Validity

We discuss the threats to the validity of this work in accordance with Wohlin et al. [26]:

The construct validity of our experiment is threatened by *mono-operation bias*, as all participants performed the same tasks with the same software. In the future, we plan to conduct experiments on physiological triggers for explainability in other contexts and settings. Furthermore, our experiment is subject to *mono-method bias*, as our only source of data are the biometric sensor data and the screen recording. In future experiments, we plan to supplement the biometric detection of explainability needs with questionnaires that can validate whether or not a need for explanation was truly present.

The internal validity of our experiment is not threatened by *selection bias*, as the demographic distribution of our sample was fairly spread out. That means we are confident that our results are not influenced by our selection of study participants. However, it is questionable if this applies at a sample size of nine participants. *Experimenter bias* only applies to the first author of this paper, i.e., they are invested in the results of this work. The other two authors, while working on explainability, are not invested in the explicit research of physiological triggers and therefore not influenced by experimenter bias. Our research approach is novel and the intention of the study was only revealed to participants after they finished the study. Therefore, it is unlikely that our work is threatened by *hypothesis guessing*, as our study participants were most likely unable to guess the goal of the study while they were participating.

The conclusion validity of our work is threatened, as we only applied descriptive statistics, rather than hypothesis testing. This is in part caused by the small sample size of nine participants. Therefore, the conclusions of this work should be seen as preliminary only. The *reliability of measures* is limited by the precision of the biometric sensors of the *empatica* watch. As the *empatica* watch is currently used in professional medical contexts, we are confident that the sensors are reliable enough for the purposes of our experiment. The manual evaluation of the EDA data also poses a threat to the conclusion validity, which could be mitigated by an automated method for evaluating the EDA value in future experiments.

The external validity of our work is threatened by the small sample size of nine participants. As such, our results are only preliminary and cannot be generalized to a larger population.

6. Conclusion and Future Work

Detecting requirements for explanations by wearing a biometric watch is rather unusual, but we consider it a creative elicitation method that can avoid biases and reliance on tacit knowledge. Our study suggests that EDA values may be an interesting indicator for explanatory needs that should be investigated further. Due to the small sample size in our study, we cannot make any definitive statements about whether EDA can actually predict the need for explanations, but our preliminary results motivate further research in this direction.

We will conduct more studies to investigate the effectiveness of EDA as an indicator for explanation needs. In particular, we want to investigate other methods for the evaluation of EDA, with the goal of achieving higher precision. We are confident that at a high level of precision, EDA could enable the elicitation of explanatory needs during system runtime and serve as a trigger for the automatic display of explanations.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No.: 470146331, project softXplain (2022-2025).

References

- [1] L. Chazette, K. Schneider, Explainability as a non-functional requirement: challenges and recommendations, *Requirements Engineering* 25 (2020) 493–514.
- [2] L. Chazette, W. Brunotte, T. Speith, Exploring explainability: a definition, a model, and a knowledge catalogue, in: *2021 IEEE 29th international requirements engineering conference (RE)*, IEEE, 2021, pp. 197–208.
- [3] N. Tintarev, J. Masthoff, Designing and evaluating explanations for recommender systems, in: *Recommender systems handbook*, Springer, 2011, pp. 479–510.
- [4] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), *IEEE access* 6 (2018) 52138–52160.
- [5] J. Droste, H. Deters, J. Puglisi, J. Klünder, Designing end-user personas for explainability requirements using mixed methods research, in: *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, IEEE, 2023, pp. 129–135.
- [6] A. Ferrari, P. Spoletini, S. Gnesi, Ambiguity and tacit knowledge in requirements elicitation interviews, *Requirements Engineering* 21 (2016) 333–355.
- [7] G. W. Harrison, E. E. Rutström, Chapter 81 experimental evidence on the existence of hypothetical bias in value elicitation methods, in: C. R. Plott, V. L. Smith (Eds.), *Handbook of Experimental Economics Results*, volume 1, Elsevier, 2008, pp. 752–767.
- [8] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, IEEE, 2018, pp. 80–89.
- [9] F.-L. Fan, J. Xiong, M. Li, G. Wang, On interpretability of artificial neural networks: A survey, *IEEE Transactions on Radiation and Plasma Medical Sciences* 5 (2021) 741–760.

- [10] H. Deters, J. Droste, M. Fechner, J. Klünder, Explanations on demand—a technique for eliciting the actual need for explanations, in: 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), IEEE, 2023, pp. 345–351.
- [11] W. Brunotte, A. Specht, L. Chazette, K. Schneider, Privacy explanations—a means to end-user trust, *Journal of Systems and Software* 195 (2023) 111545.
- [12] H. Deters, J. Droste, K. Schneider, A means to what end? evaluating the explainability of software systems using goal-oriented heuristics, in: *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, 2023, pp. 329–338.
- [13] W. Brunotte, J. Droste, K. Schneider, Context content consent—how to design user-centered privacy explanations, in: *The 35th International Conference on Software Engineering & Knowledge Engineering*, 2023.
- [14] N. Tintarev, J. Masthoff, Effective explanations of recommendations: user-centered design, in: *Proceedings of the 2007 ACM conference on Recommender systems*, 2007, pp. 153–156.
- [15] A. Zalewski, K. Borowa, D. Kowalski, On cognitive biases in requirements elicitation, *Integrating research and practice in software engineering (2020)* 111–123.
- [16] L. Chazette, J. Klünder, M. Balci, K. Schneider, How can we develop explainable systems? insights from a literature review and an interview study, in: *Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering*, 2022, pp. 1–12.
- [17] I. Nunes, D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, *User Modeling and User-Adapted Interaction* 27 (2017) 393–444.
- [18] W. Boucsein, *Electrodermal activity*, Springer Science & Business Media, 2012.
- [19] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, R. Picard, Automatic identification of artifacts in electrodermal activity data, in: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 1934–1937.
- [20] A. Kushki, J. Fairley, S. Merja, G. King, T. Chau, Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites, *Physiological measurement* 32 (2011) 1529.
- [21] E. Piciuccio, E. Di Lascio, E. Maiorana, S. Santini, P. Campisi, Biometric recognition using wearable devices in real-life settings, *Pattern Recognition Letters* 146 (2021) 260–266.
- [22] D. Girardi, F. Lanubile, N. Novielli, A. Serebrenik, Emotions and perceived productivity of software developers at the workplace, *IEEE Transactions on Software Engineering* 48 (2022) 3326–3341. doi:10.1109/TSE.2021.3087906.
- [23] P. Schmidt, A. Reiss, R. Dürichen, K. Van Laerhoven, Wearable-based affect recognition—a review, *Sensors* 19 (2019) 4079.
- [24] S. A. H. Aqajari, E. K. Naeini, M. A. Mehrabadi, S. Labbaf, N. Dutt, A. M. Rahmani, pyeda: An open-source python toolkit for pre-processing and feature extraction of electrodermal activity, *Procedia Computer Science* 184 (2021) 99–106.
- [25] B. Rosner, Percentage points for a generalized esd many-outlier procedure, *Technometrics* 25 (1983) 165–172.
- [26] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén, *Experimentation in software engineering*, Springer Science & Business Media, 2012.