

On Using Physiological Sensors and AI to Monitor Emotions in a Bug-Hunting Game

Natalia Silvis-Cividjian
Vrije Universiteit
Department of Computer Science
Amsterdam, The Netherlands
n.silvis-cividjian@vu.nl

Joshua Kenyon
Vrije Universiteit
Department of Computer Science
Amsterdam, The Netherlands

Elina Nazarian
Vrije Universiteit
Amsterdam, The Netherlands

Stijn Sluis
Vrije Universiteit
Amsterdam, The Netherlands

Martin Gevonden
Vrije Universiteit
Department of Biological Psychology
Amsterdam, The Netherlands

ABSTRACT

Although software testing is key to a safe society, the process itself is often perceived by students as boring and stressful. Therefore, only few consider a career in testing. The adverse effect is sub-optimally tested code, with dangerous bugs left undetected. A better understanding of what testers “feel” when learning the skill in class can remedy this situation, by means of personalized, motivating bio-feedback. In order to test our hypothesis, we propose an innovative approach that uses physiological wearable sensors (cardiac activity, respiration, and skin conductance) to monitor in real-time the affective state of testers engaged in a bug-hunting game. This is a work in progress. We present the envisioned methodology and the results of two feasibility experiments. The first experiment created a training dataset, by recording bio-signals and self-reports from eleven participants involved in a mood-induction session followed by a bug-hunting task. The second experiment showed that it is possible to use deep-learning to recognize emotions from a large set of labelled multimodal (ECG, EDA and ICG) physiological data. The classification accuracy using a binary (positive-negative) emotions model was 85%, higher than the accuracy obtained using a four-emotions (anxious, down, enthusiastic and relaxed) model (57%). Future work includes optimizing the sensory system, improving the accuracy of automated emotions recognition, increasing the validity of ground-truth emotions labelling, and investigating ways to provide individualized and formative (instead of summative) bio-feedback. The proposed approach can contribute to a more sentiment-aware education, and a more objective evaluation of the effect of teaching interventions.

CCS CONCEPTS

• **Software and its engineering** → **Software verification and validation**; • **Human-centered computing** → *Laboratory experiments*.

KEYWORDS

software testing education, bug-hunting gamification, automated emotion recognition, sentiment analysis, biometric ECG signals, deep-learning

ACM Reference Format:

Natalia Silvis-Cividjian, Joshua Kenyon, Elina Nazarian, Stijn Sluis, and Martin Gevonden. 2024. On Using Physiological Sensors and AI to Monitor Emotions in a Bug-Hunting Game. In *Proceedings of the 2024 Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2024)*, July 8–10, 2024, Milan, Italy. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3649217.3653611>

1 RATIONALE

Software testers are the last line of defense against threats that vary from train delays and private data leaks, to fatal plane crashes and malfunctioning medical devices. Unfortunately, checking that a software product satisfies its requirements for *all* possible scenarios is a rather boring activity, not attractive for computer science (CS) students. The adverse effect is that motivated testers are difficult to find on the job market [7, 39, 44]. A challenge for computing educators is to find ways to motivate students to enjoy testing and not give up learning too soon. Our response to this challenge is a mature Software Testing course, offered at the Vrije Universiteit in Amsterdam to approximately 100 CS graduates yearly. One of its characteristic elements is a rich exposure of students to bugs [36]. An example is DBugIT, an online interactive tool developed in our department to assess students’ testing skills by means of a bug-hunting game [38]. DBugIT is a product of the VU-BugZoo Comenius Teaching Fellow innovation project we initiated in 2019, with the aim to make software testing more exciting [37]. A game in DBugIT mimics a situation similar to the routine work of a tester; students are provided with the requirements specification of a small software module, like a BMI calculator, a discount calculator for a web-shop, or a control software for a smart home thermostat, and a fault-seeded executable implementation, also known as a mutant. The source code is visible for white-box testing tasks and invisible for black-box testing tasks. Students know from the beginning that their software under test contains at least one bug, but they do not know its type, nor the location. Examples of bugs we deliberately injected are commonly-made coding errors, such as: off-by-one, an excluded boundary of a domain, or an omitted check for division



This work is licensed under a Creative Commons Attribution International 4.0 License.

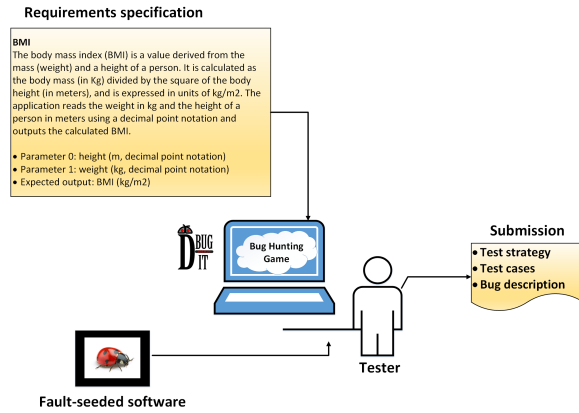


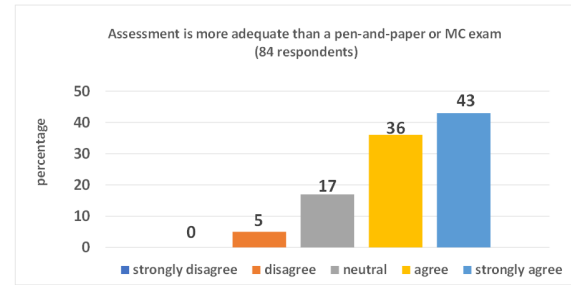
Figure 1: The principle of the bug-hunting game.

by zero. For each mutant, students have to design an adequate test strategy, specifying the test techniques and their associated test cases. A test case consists of a program input and the expected output. The tool is interactive; students can feed any test inputs to the mutant, and the system will execute the mutant and return the program's output. It is the task of the students to conclude if a test case passed or failed. When they think they found the bug, the students have to describe it as accurately as possible, together with its possible location, which is more than is normally asked from a tester. Students type their report in an embedded editor and submit it.

The bug-hunting task in our course was open for five days. After the deadline passed, a teacher read the submissions offline and graded them. In the grading scheme, the soundness of the test strategy had the highest weight (50%), followed by the quality of test cases (25%) and the description of the bug using a bug taxonomy (25%) (see Fig. 1).

The initiative was evaluated after each deployment, by means of student surveys. The surveys contained usual questions, such as whether a bug-hunting game is a more adequate assessment than the traditional pen-and-paper or multiple choice exams, but also more unusual questions, such as which emotions students have experienced during the bug hunting. The answers to these two questions, integrated over four years, are shown in Fig. 2.

The surveys results showed that the majority of students experienced excitement and joy when they found the bug, which was reassuring for us. However, some were anxious and stressed that they may not find the bug before the deadline, or frustrated when the requirements were not clear to them. Other students were irritated by the slow game execution. We could explain this by inspecting the DBugIT activity logs, which showed a high users' load shortly before the deadline, caused by the students who procrastinated to start working on the task until the last moment. Also, we had a few students who initially started with the task, but never submitted it. It would be interesting to know why; maybe this was the result of accumulated negative emotions, such as frustration, stress or fear of failure? A limitation of our survey was that the answers shown in Fig. 2. were given by students days after the game ended, and not *during* the game. It is therefore uncertain how



"I **laughed** out loud when I found the bug in black box testing. I had a very nice assignment for this and I **enjoyed** testing it. I was **eager** to find the bug(s). To be honest, it was **stressful** for me because I knew there was a bug but it wasn't obvious. **Excitement** when I found the bug. Slight **stress** with the black box testing that I would not find the bug. Feeling of **achievement** when found a bug. **Excitement** when I figured out how the bug could be resolved. However, I got a bit **frustrated** with some of the requirements, which were sometimes underspecified."

Figure 2: An excerpt from the students' survey related to the adequacy of the assessment and the emotions experienced during the bug-hunting game.

accurately students described their feelings. Maybe they forgot about some emotions? Also, in their self-reports students told us that they liked the game, but this is a very subjective answer. What if they just wanted to please us?

Summarizing, we learned from the evaluations that during a bug-hunting game, our students have experienced a mix of positive and negative emotions. However, we do not know exactly *what* they felt and *when* they felt it. Was this a missed opportunity? Existing scientific evidence showing that emotions substantially influence learning, motivation and performance [17, 27, 40] made us believe that indeed, this was the case. This led us to the following curiosity-driven question: Imagine that we can measure the emotions our students go through during a bug-hunting game, will this benefit our teaching?

In order to answer this question, we modelled a traditional classroom using a rather unusual control systems theory, a well-known engineering topic (see Fig. 3). In this model, the students are a controlled process and the teacher is the controller, with control actions such as conveying knowledge, assessing student's knowledge and skills, and generating feedback. The teacher's mental model regarding the students ("How well did the students learn?", "Are the students motivated?") is influenced by the feedback provided by the student, such as essays, exams or solutions to exercises, answers to surveys, etc. Based on this mental model, the teacher adapts his/her future behaviour.

In this model, our idea translates to enabling an additional feedback channel running from students towards the teacher, carrying their affective state information. English and English (1958) defines affect as:

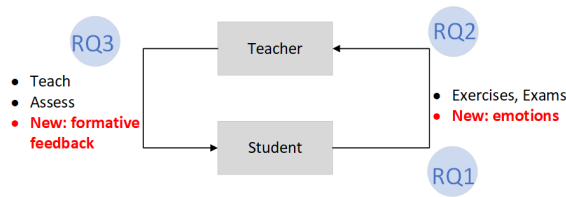


Figure 3: The envisioned emotion-aware teaching model.

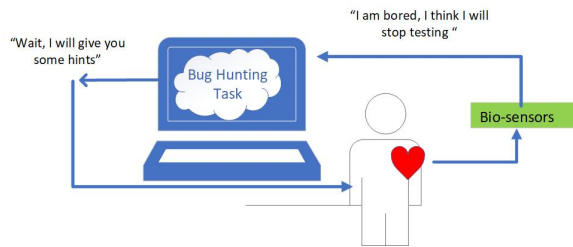


Figure 4: Illustration of a user scenario.

“A class name for feeling, emotion, mood, temperament...a single feeling-response to a particular object or idea....the general reaction toward something liked or disliked...the dynamic or essential quality of an emotion; the energy of an emotion” [9].

The main advantage of this additional channel will be the possibility to provide just-in-time, personalized formative feedback that will boost students’ motivation. We illustrate this advantage using the following user scenario (see Fig. 4).

“A student works on a bug-hunting task and the system continuously monitors their emotions. When the student has found a bug, the system rewards them with divine points in a BugJar. When the system detects that the student is bored or frustrated and risks to quit testing too soon, a motivating message will pop-up, encouraging them to keep on testing, also reminding that a reward is waiting for them. If the student finds the task too difficult and feels stuck, the system detects this and provides some tips on which test inputs to try, or which line of code to look at in order to find the bug, or it might decide to reduce the difficulty level of the task, and provide some theory or an easier assignment first”.

We also identified another, more long-term advantage of the proposed approach, namely the creation of a more reliable instrument to measure the effect of in fact any teaching initiative.

When we became confident that our idea has the potential to benefit teaching in a software testing class, we focused on its implementation, and formulated the following research questions (see Fig. 3).

RQ1. How to sense the affective state of a tester engaged in a bug-hunting game?

RQ2. How to classify a tester’s emotions based on the sensed biometric data ?

RQ3. How to generate motivating feedback based on a tester’s emotions?

In this position paper, we will describe the first steps we took towards implementing a novel emotion-carrying channel running from students towards the teacher, based on physiological sensors,

self-reports and machine learning. The aim of this study was to answer RQ1 and RQ2. Answering RQ3 and generating a personalized motivating feedback based on the gauged emotions is future work. This is, to the best of our knowledge, the first attempt at using sentiment analysis to create an engaging software testing learning environment.

2 HOW TO IMPLEMENT AN AUTOMATED EMOTION RECOGNITION SYSTEM?

Automated emotion recognition starts with sensing data from a human subject. The sensed data is then processed and presented as input to a classifier, an algorithm which will decide to which class the measured emotion belongs. Some machine learning algorithms need an additional training phase. Finally, the approach is evaluated, by calculating performance indicators such as classification accuracy, where one compares the predicted emotions with the actual ones, also called ground truth, as reported by the subject. A perfect classification means a 100% accuracy, which rarely happens in practice. In the next subsections, we will outline different technical solutions one could adopt to implement each of these steps.

2.1 Which sensors to use?

An easy way to detect emotions is based on human physical signals, such as facial expression, hand gestures or body posture, which can be sensed using a video camera. However, the reliability of this approach cannot be guaranteed, as it is relatively easy for people to hide or fake their real emotions. Another, more reliable way to detect emotions is to analyze speech or text [25]. An even more objective, yet challenging method to recognize emotions is using physiological signals, such as electrical brain activity, also called electro-encephalogram (EEG), electrical heart activity, or electrocardiogram (ECG), galvanic skin response (GSR), also known as galvanic skin conductivity (GSC), or electrodermal activity (EDA), blood pressure (BP), and electrical conductivity of the thorax, also called impedance cardiogram (ICG). Because it is not clear yet which signals are the most successful in predicting human emotions, it is very common to use a combination of many types of sensors (called sensor fusion) [15, 35, 45].

2.2 Which emotions to recognize?

Currently, there is no golden standard technique for emotions categorization. First, because it is still unclear how exactly different emotions are generated and what factors influence them. Second, because emotions are complex, dynamic processes and a human is in fact at any moment experiencing a mix of emotions. Some researchers prefer to talk about moods, like Fisher [11] and Khan et al. [23], because moods tend to be prolonged events that are easier to measure, while emotions are a more episodic process. A few models to categorize emotions exist though. The simplest way to classify emotions in sentiment analysis experiments is using two classes: negative and positive. Other models work with three factors that define human affective response to stimuli, such as pleasure, arousal, and dominance. Pleasure, or valence, represents how positively or negatively the emotion is experienced. Arousal or excitement, refers to the intensity of that emotion. Dominance

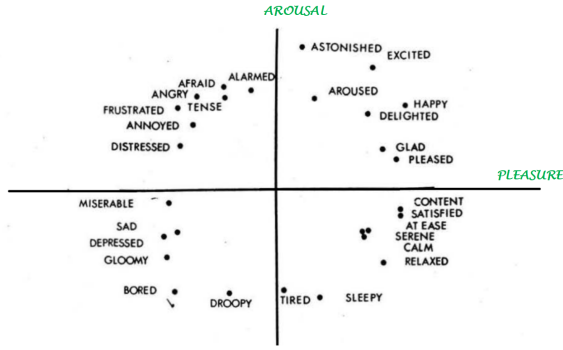


Figure 5: The Russell's circumplex model for emotions categorization. Adapted from [32].

is defined by the extent to which an individual feels in control of the situation. Russell's 'circumplex model of affect', developed in 1980, classifies emotions as a function of only two of these three factors, namely arousal and pleasure [32] (see Fig. 5). The model of Paul Ekman (1992) categorizes the core emotions in six classes: anger, disgust, fear, happiness, sadness, and surprise [8].

2.3 Which classifiers to choose?

A classifier is an algorithm that is able to categorize a new-comer as belonging to one of the known classes. A simple way to conduct sentiment analysis is using a rule-based classifier, like VADER (Valence Aware Dictionary and sEntiment Reasoner) [21], used for opinion mining on social media texts. A next level classifier is template matching, that needs a large database of known fingerprints. A new comer's fingerprint is matched with all known fingerprints until the right and the most similar class wins. More advanced classifiers use probabilistic models such as Bayes Networks [34] or machine learning approaches, such as neural networks that accept features as inputs, or deep-learning algorithms such as convolutional neural networks (CNN), and recurrent neural networks (RNN), where no feature extraction is needed [28, 45].

2.4 How to evaluate classification?

The most difficult part in evaluating a sentiment analysis experiment is to know the ground truth, or in other words, what a subject really feels. Traditionally, in psychology the ground truth in emotion recognition is obtained with pen-and-paper self-reports, in which the subject just answers a questionnaire during or immediately after the experiment. Modern applications that make use of a graphical user interface, present the user with a pop-up window, asking to estimate the intensity of a certain emotion (using a number, or a slider).

An interesting alternative to these is the Self-Assessment Manikin (SAM) [3, 26, 30], a non-verbal, picture-based instrument to directly measure the pleasure, arousal, and dominance associated with a person's affective reaction to a wide variety of stimuli (see Fig. 6). The SAM has the advantage that it is not language-dependent and can reduce the ambiguity inherent to such a complex process like reporting one's feelings.

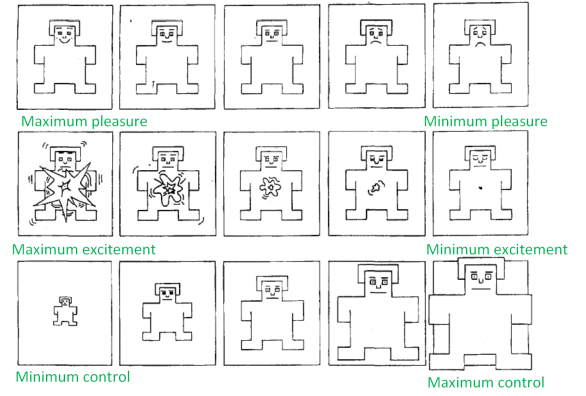


Figure 6: The picture-based dialog window of SAM. Adapted from [3].

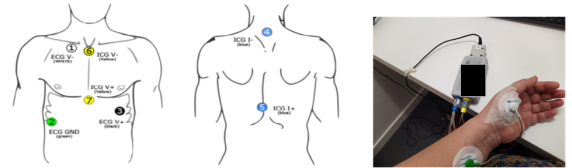
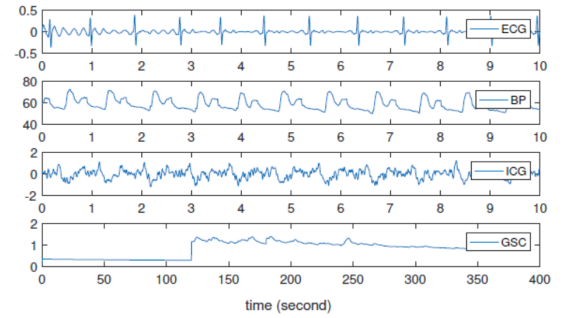


Figure 7: Samples of typical ECG-EDA-ICG bio-signals, adapted from [22] and the placement of electrodes for ECG-EDA-ICG recordings using the VU-AMS device [43].

3 FIRST RESULTS

As preliminary work, we performed two feasibility experiments, both using the Vrije Universiteit Ambulatory Monitoring Systems (VU-AMS) wearable ECG-EDA-ICG cardiac activity recording device [43], with seven electrodes attached on the subject's chest, back and hand, as shown in Fig. 7.

Experiment#1. Monitoring emotions in real-time

The goal of this experiment was to address RQ1 and gain experience in (ECG-EDA-ICG) biometric data acquisition. For this purpose, we recruited eleven CS students who volunteered to play the role of a tester. The experiment was prepared with care. Participants received an information letter and signed a written informed consent

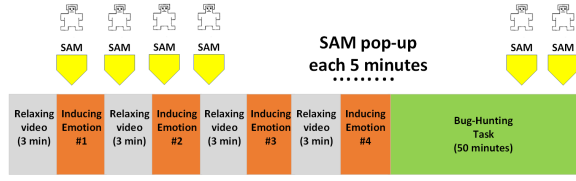


Figure 8: The emotions induction and monitoring protocol.

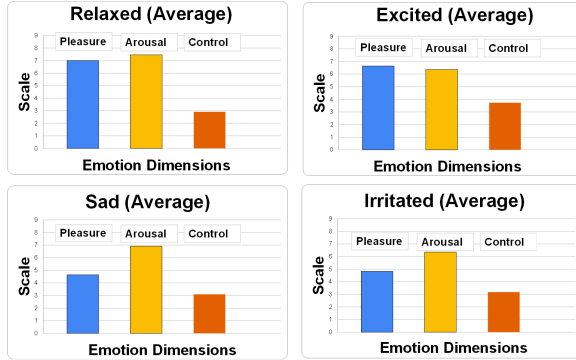


Figure 9: Self-reported SAM answers for each of the four induced emotions, plotted as an average over eleven subjects.

before the start of the experiment. Also, the study was in advance discussed with the faculty ethical committee, who recommended us to design an adequate data management plan. Participants were compensated with a gift voucher.

The subjects were wearing a VU-AMS device and were working on a computer, guided by a script we developed for this purpose. During the whole experiment, their physiological (ECG, EDA, ECG) signals were continuously recorded and saved in data files.

The experiment design and protocol are illustrated in Fig. 8. First, a mood-induction session took place, where we sequentially induced four emotions (irritation, sadness, excitement, relaxed) which we expected to be typical to a bug-hunting task. In order to induce the four emotions, we used images from the International Affective Picture System (IAPS) dataset [4]. We saved the recorded biometric data of each participant in a training dataset. After that, the subjects were directed to a bug-hunting task running in DBugIT that took 50 minutes. Their bio-metric data were recorded again, and saved in another, testing dataset.

In order to know the ground-truth, during the whole experiment, each participant was presented every five minutes with a SAM pop-up window on the screen, on which they had to rate their emotional state. Based on the following two articles [5, 33], we created a mapping between the four induced emotions and the three possible SAM scale answers (pleasure, arousal, and control). The recorded biometric data was annotated with time stamps and the SAM values reported by the subjects.

Processing the data in order to recognize emotions using AI is future work. Up to now, we only processed the answers self-reported by the participants in SAM. We show the results in Fig. 9 and Fig. 10.

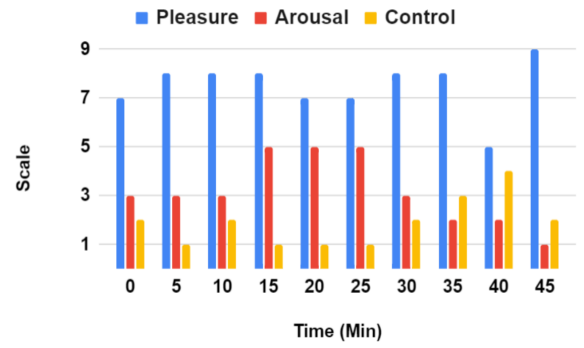


Figure 10: Self-reported SAM answers during the bug-hunting task, plotted for one subject.

From Fig. 9 we can see that the self-reported emotions fairly match the emotions we aimed to induce. However, we also see that it is difficult to distinguish between emotions only based on the level of arousal and pleasure. For example, a level of arousal much higher than pleasure was measured during both "sad" and "irritated" episodes. This might happen because static images like the ones we used are not such a strong stimulus. Next time, we will investigate whether using videos as stimuli can improve the situation. Also, it seems that the level of dominance (control) does not vary so strongly between different emotions. An explanation can be that the SAM pictures for "control" taken from the referenced articles were counter-intuitively mirrored, showing on the left side the minimum control level, instead of the expected maximum, which might have confused the subjects. Next time, we will mirror the labels for dominance in the SAM pop-up window.

From Fig. 10 we observe a variability in the levels of pleasure and arousal during the bug-hunting task for one person, which is positive news. Of course, it is too soon to draw conclusions, so it might be just wishful thinking to believe that between $t = 20$ min and $t = 25$ min, the subject started to feel bored and the pleasure level dropped, after which a peak in the excitement occurred at $t = 30$ min, signaling that the subject probably found the bug and started to enjoy again the task.

These preliminary results show how difficult it is to induce emotions in human subjects and gauge what they really feel. The limitations of this experiment are the very small number of participants, not knowing how to effectively induce and elicit emotions, and the poor scalability. Although the experiment was lightweight, it helped us to get new insights and ideas on how to tackle the second round of experiments.

Experiment#2. Training a deep-learning emotions classifier

This experiment aimed to answer RQ2, by investigating whether we can apply machine learning algorithms to classify emotions from recorded cardiac activity signals. To reduce the complexity of the task, we used for training an existing, validated and labeled dataset, created in the past for the purpose of stress measurement [41]. The dataset contained ECG-EDA-ICG recordings of 112 healthy adults. Participants were administered a range of mental and physical

stressors, including a cross that flares in a corner of a computer screen, stairs climbing, and vacuum cleaning. During these tasks, their physiological ECG-EDA-ICG responses were monitored and recorded. The ground-truth participants' emotions were labeled by filling out a pen-and-paper questionnaire, directly following each of a series of tasks. Subjects were asked to rate their emotional states on a 7-point Likert scales ranging from 1 (not at all) to 7 (very). The categories consisted of four positive emotions, namely 'relaxed', 'cheerful', 'enthusiastic' and 'content', and five negative emotions, namely 'insecure', 'lonely', 'anxious', 'irritated' and 'down'. These emotions, commonly used in Ecological Momentary Assessment (EMA) studies by the University of Maastricht and the KU Leuven, were chosen given their coverage of the whole emotional spectrum described in the Russel's model and their maximum within-person time-lagged variability [18]. Because the data was recorded before we started this study, the emotions categories were not particularly related to software testing.

In order to categorize emotions, we used two models. First we used a simple two (positive-negative) classes model. The person's positive mood level was calculated as the average of the four self-reported positive emotions levels. The level of the negative mood was calculated as the average of the five self-reported negative emotions levels. Averaging is a very common approach in the emotion-related literature and specifically in EMA studies. Next, we also used a four-emotions model, consisting of the following emotions: 'anxious', 'down', 'enthusiastic' and 'relaxed'). We have chosen these particular emotions because they are the closest to the ones experienced by a software tester, and because they sufficiently cover the emotions spectrum defined in the Russel's model.

We trained a deep-learning classifier based on a type of Recurrent Neural Network (RNN) called Long Short-Term Memory (LSTM) Network [19, 20], programmed in MATLAB. The results show that the simple binary model returned the best classification accuracy of 84.4%, whereas the four-emotions model obtained a lower accuracy, of approximately 57%.

The limitations of this experiment are the relatively low size of training data, the fact that only one deep-learning algorithm was investigated, whereas there are so many algorithms suitable for automated emotion recognition, and the fact that the training data was old, and not particularly targeted at software testers. However, these findings confirmed that it is possible to train a machine learning classifier on multimodal (ECG, EDA and ICG) biometric data to automatically recognize emotions.

4 RELATED WORK

We were encouraged in our work by CS education researchers who conducted experiments that demonstrated that gamification in a software testing class increases motivation, engagement and performance, by exploiting the competitive nature of humans [2, 12]. In the last two decades, researchers have been urging educators to better listen to the learner's feelings and link emotions with motivation and cognition [1, 10, 27]. A literature review on using sentiment analysis in education in general, but not particularly in computer science can be found in [46]. A literature review on the state-of-the-art of automated emotion recognition systems can be found in [24]. Over the years, researchers have conducted many

studies using different physiological sensors, such as EEG, GSR, ECG, BP and hormone levels. Consistently, their conclusion was that the heart, and not the brain activity is the most reflective indicator of one's emotional states [6, 29].

Outside the classroom, various researchers experimented with automated emotion recognition used to analyze and improve software engineer's well-being at work. In particular, Muller et al. and Girardi used EEG signals to get insights into the behaviour of programmers [14, 31]. Grassi et al. used EDA signals to assist Agile teams in their retrospective meetings [16]. Fritz [13] used a combination of eye trackers, EEG, EDA and ECG sensors to assess task difficulty in coding. Vrzakova [42] used a multimodal sensing combination (eye trackers, GSR and pressure sensors) and machine learning to monitor the emotions of participants in a code review session in a large company. All authors reported promising results, with an accuracy of around 85%.

In our proposal, we were inspired by the experiences from both worlds, that of affective education and of software engineering teaching. So far, we are not aware of any similar biometric studies that gauge learner's emotions in order to create an engaging environment in a software testing classroom.

5 CONCLUSION AND FUTURE WORK

This paper elaborated on an innovative idea to boost students' motivation in a software testing classroom by adding a new interaction channel between teacher and student, based on automated emotion detection. Two lightweight feasibility experiments confirmed that it is possible to monitor the emotional state of testers working on a bug-hunting task using a multimodal (ECG, EDA and ICG) combination of physiological sensors, and that deep-learning algorithms can be trained to make sense of the raw biometric data. However, we also realized that the road will be long and full of opportunities and challenges. A multidisciplinary approach involving computer science, biological psychology, and education expertise will be needed to turn this idea into success. Future work plans include optimizing the emotion recognition process, involving a larger number of participants, increasing the validity of self-reported ground-truth emotion labels, adding less-intrusive sensors to increase scalability, and exploring ways to generate emotion-aware motivating feedback tailored for novice software testers. On the long term, the approach can become interesting for other educators who experiment with sentiment analysis or capture-the-flag (CTF) style learning initiatives and gamification.

ACKNOWLEDGMENTS

The EMOTT research project was partially funded by the Network Institute, Vrije Universiteit Amsterdam through the Academy Assistants program. DBugIT is a product of the VU-BugZoo project, funded by the NRO, The Netherlands Initiative for Education Research, as part of a Comenius Teaching Fellow grant. The authors would like to thank Chris Corsello for his work on the preliminary machine-learning experiments; Mircea Țăulescu for advising on the first steps of our journey; the developers of DBugIT and Marketa Ciharova, Denise van der Mee and Cor Stoof for their help in setting up the physiological experiments; all students who participated as volunteers in our experiments.

REFERENCES

- [1] Elizabeth Acosta-Gonzaga and Aldo Ramirez-Arellano. 2021. The Influence of Motivation, Emotions, Cognition, and Metacognition on Students' Learning Performance: A Comparative Study in Higher Education in Blended and Traditional Contexts. *SAGE Open* 11, 2 (2021), 21582440211027561. <https://doi.org/10.1177/21582440211027561>
- [2] Raquel Blanco, Manuel Trinidad, María José Suárez-Cabal, Alejandro Calderón, Mercedes Ruiz, and Javier Tuya. 2023. Can gamification help in software testing education? Findings from an empirical study. *Journal of Systems and Software* 200 (2023), 111647. <https://doi.org/10.1016/j.jss.2023.111647>
- [3] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59.
- [4] M. M. Bradley and P. J. Lang. 2007. The International Affective Picture System (IAPS) in the study of emotion and attention. *Handbook of emotion elicitation and assessment* 18 (2007).
- [5] Oana Bălan, Gabriela Moise, Livia Petrescu, Alin Moldoveanu, Marius Leordeanu, and Florica Moldoveanu. 2020. Emotion Classification Based on Biophysical Signals and Machine Learning Techniques. *Symmetry* 12, 1 (2020).
- [6] Diego Candia-Rivera, Vincenzo Catrambone, Julian F. Thayer, Claudio Gentili, and Gaetano Valenza. 2022. Cardiac sympathetic-vagal activity initiates a functional brain–body response to emotional arousal. *Proceedings of the National Academy of Sciences* 119, 21 (2022), e2119599119. <https://doi.org/10.1073/pnas.2119599119>
- [7] Anca Deak, Tor Stålhane, and Daniela Cruzes. 2013. Factors Influencing the Choice of a Career in Software Testing among Norwegian Students. *IASTED Multiconferences - Proceedings of the IASTED International Conference on Software Engineering, SE 2013* (03 2013). <https://doi.org/10.2316/P.2013.796-032>
- [8] P. Ekman, R. W. Levenson, and W. V. Friesen. 1983. Autonomic nervous system activity distinguishes among emotions. *Science* 221, 4616 (1983), 1208–1210.
- [9] Horace B English and Ava Champney English. 1958. A comprehensive dictionary of psychological and psychoanalytical terms: A guide to usage. (1958).
- [10] Peter Op't Eynde and Jeannine E Turner. 2006. Focusing on the complexity of emotion issues in academic learning: A dynamical component systems approach. *Educational Psychology Review* 18 (2006), 361–376.
- [11] Cynthia D. Fisher. 2000. Mood and emotions while working: Missing pieces of job satisfaction? *Journal of Organizational Behavior* 21, 2 (2000), 185–202.
- [12] Gordon Fraser. 2017. Gamification of Software Testing. In *2017 IEEE/ACM 12th International Workshop on Automation of Software Testing (AST)*. 2–7. <https://doi.org/10.1109/AST.2017.20>
- [13] Thomas Fritz, Andrew Begel, Sebastian C. Müller, Serap Yigit-Elliott, and Manuela Züger. 2014. Using Psycho-Physiological Measures to Assess Task Difficulty in Software Development. In *Proceedings of the 36th International Conference on Software Engineering*. 402–413.
- [14] Daniela Girardi, Nicole Novielli, Davide Fucci, and Filippo Lanubile. 2020. Recognizing developers' emotions while programming. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*.
- [15] A. Goshvarpour, Abbasi, and Goshvarpour A. 2017. An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomedical Journal* 40 (2017), 355 e368.
- [16] Daniela Grassi, Filippo Lanubile, Nicole Novielli, and Alexander Serebrenik. 2023. Towards Supporting Emotion Awareness in Retrospective Meetings. *2023 IEEE/ACM 45th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)* (2023), 101–105.
- [17] Daniel Graziotin, Fabian Fagerholm, Xiaofeng Wang, and Pekka Abrahamsson. 2018. What happens when software developers are (un)happy. *Journal of Systems and Software* 140 (2018), 32–47.
- [18] Laila Hasmi, Marjan Drukker, Sinan Guloksuz, Claudia Menne-Lothmann, Jeroen Decoster, Ruud van Winkel, Dina Collip, Philippe Delespaul, Marc De Hert, Catherine Derom, Evert Thiery, Nele Jacobs, Bart P. F. Rutten, Marieke Wichers, and Jim van Os. 2017. Network Approach to Understanding Emotion Dynamics in Relation to Childhood Trauma and Genetic Liability to Psychopathology: Replication of a Prospective Experience Sampling Analysis. *Frontiers in Psychology* 8 (2 Nov. 2017). <https://doi.org/10.3389/fpsyg.2017.01908>
- [19] Sepp Hochreiter. 1998. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 6 (1998), 107–116. <https://api.semanticscholar.org/CorpusID:18452318>
- [20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [21] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* (2014).
- [22] Aya Khalaf, Mohsen Nabian, Miaolin Fan, Yu Yin, Jolie Wormwood, Erika Siegel, Karen S. Quigley, Lisa Feldman Barrett, Murat Akcakaya, Chun-An Chou, and Sarah Ostadabbas. 2020. Analysis of multimodal physiological signals within and between individuals to predict psychological challenge vs. threat. *Expert Systems with Applications* 140 (2020), 112890. <https://doi.org/10.1016/j.eswa.2019.112890>
- [23] Iftikhar Ahmed Khan, Willem-Paul Brinkman, and Robert M Hierons. 2011. Do moods affect programmers' debug performance? *Cognition, Technology & Work* 13 (2011), 245–258.
- [24] Smith K. Khare, Victoria Blanes-Vidal, Esmaeil S. Nadimi, and U. Rajendra Acharya. 2024. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion* 102 (2024), 102019. <https://doi.org/10.1016/j.inffus.2023.102019>
- [25] Bennett Kleinberg. 2020. Manipulating emotions for ground truth emotion analysis. arXiv:2006.08952 [cs.CL]
- [26] Peter Lang. 1980. Behavioral treatment and bio-behavioral assessment: Computer applications. *Technology in mental health care delivery systems* (1980), 119–137.
- [27] Elizabeth Linnenbrink. 2006. Emotion Research in Education: Theoretical and Methodological Perspectives on the Integration of Affect, Motivation, and Cognition. *Educational Psychology Review* 18 (01 2006), 307–314. <https://doi.org/10.1007/s10648-006-9028-x>
- [28] Kristina Machová, Martina Szabóová, Ján Paralič, and Ján Mičko. 2023. Detection of emotion by text analysis using machine learning. *Frontiers in Psychology* 14 (2023). <https://doi.org/10.3389/fpsyg.2023.1190326>
- [29] Rollin McCraty. 2016. Science of the Heart, Volume 2 Exploring the Role of the Heart in Human Performance An Overview of Research Conducted by the HeartMath Institute. (02 2016). <https://doi.org/10.13140/RG.2.1.3873.5128>
- [30] Jon D. Morris. 1995. Observations: SAM: The Self-Assessment Manikin An Efficient Cross-Cultural Measurement Of Emotional Response 1. *Journal of Advertising Research* (1995).
- [31] Sebastian C. Müller and Thomas Fritz. 2015. Stuck and Frustrated or in Flow and Happy: Sensing Developers' Emotions and Progress. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. 688–699.
- [32] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39 (1980), 1161–1178.
- [33] James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality* 11, 3 (1977), 273–294.
- [34] Jerritta Selvaraj, Prof Murugappan, Wan Khairunizam, and Szazali Yaacob. 2013. Classification of emotional states from electrocardiogram signals: A non-linear approach based on hurst. *Biomedical engineering online* 12 (05 2013), 44. <https://doi.org/10.1186/1475-925X-12-44>
- [35] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A Review of Emotion Recognition Using Physiological Signals. *Sensors* 18, 7 (2018).
- [36] Natalia Silvis-Cividjian. 2021. Awesome Bug Manifesto: Teaching an Engaging and Inspiring Course on Software Testing. In *2021 Third International Workshop on Software Engineering Education for the Next Generation (SEENG)*. 16–20.
- [37] Natalia Silvis-Cividjian, Rob Limburg, Niels Althuisius, Emil Apostolov, Viktor Bonev, Robert Jansma, Glenn Visser, and Marc Went. 2020. VU-BugZoo: A Persuasive Platform for Teaching Software Testing. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education (Trondheim, Norway) (ITiCSE '20)*. 553. <https://doi.org/10.1145/3341525.3393975>
- [38] Natalia Silvis-Cividjian, Marc Went, Robert Jansma, Viktor Bonev, and Emil Apostolov. 2021. Good Bug Hunting: Inspiring and Motivating Software Testing Novices. In *ITiCSE '21: Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V.1, Virtual Event, Germany, June 26 - July 1, 2021*. ACM, 171–177. <https://doi.org/10.1145/3430665.3456330>
- [39] Rodrigo E. C. Souza, Ronnie E. de Souza Santos, Luiz Fernando Capretz, Marlon A. S. de Sousa, and Cleyton V. C. de Magalhães. 2022. Roadblocks to Attracting Students to Software Testing Careers: Comparisons of Replicated Studies. In *Quality of Information and Communications Technology*, Antonio Vallecillo, Joost Visser, and Ricardo Pérez-Castillo (Eds.). Springer International Publishing, Cham, 127–139.
- [40] Chai M. Tyng, Hafeez U. Amin, Mohamad N. M. Saad, and Aamir S. Malik. 2017. The Influences of Emotion on Learning and Memory. *Frontiers in Psychology* 8 (2017).
- [41] D.J. van der Mee, M.J. Gevonden, J.H.D.M. Westerink, and E.J.C. de Geus. 2021. Validity of electrodermal activity-based measures of sympathetic nervous system activity from a wrist-worn device. *International Journal of Psychophysiology* 168 (2021), 52–64. <https://doi.org/10.1016/j.ijpsycho.2021.08.003>
- [42] Hana Vrzakova, Andrew Begel, Lauri Mehtätalo, and Roman Bednarik. 2020. Affect Recognition in Code Review: An In-situ Biometric Study of Reviewer's Affect. *Journal of Systems and Software* 159 (2020), 110434.
- [43] VU-AMS. [n. d.]. The VU Ambulatory Monitoring System. <https://vu-ams.nl/ams/>.
- [44] P. Waychal and L. F. Capretz. 2016. Why a Testing Career Is Not the First Choice of Engineers. *ArXiv abs/1612.00734* (2016).
- [45] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 2020. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Inf. Fusion* 59 (2020), 103–126. <https://api.semanticscholar.org/CorpusID:214058636>
- [46] Jin Zhou and Jun min Ye. 2023. Sentiment analysis in education research: a review of journal publications. *Interactive Learning Environments* 31, 3 (2023), 1252–1264. <https://doi.org/10.1080/10494820.2020.1826985>