Emotions and Perceived Productivity of Software Developers at the Workplace

Daniela Girardi, Filippo Lanubile, Nicole Novielli, Alexander Serebrenik

Abstract—Emotions are known to impact cognitive skills, thus influencing job performance. This is also true for software development, which requires creativity and problem-solving abilities. In this paper, we report the results of a field study involving professional developers from five different companies. We provide empirical evidence that a link exists between emotions and perceived productivity at the workplace. Furthermore, we present a taxonomy of triggers for developers' positive and negative emotions, based on the qualitative analysis of participants' self-reported answers collected through daily experience sampling. Finally, we experiment with a minimal set of non-invasive biometric sensors that we use as input for emotion detection. We found that positive emotional valence, neutral arousal, and high dominance are prevalent. We also found a positive correlation between emotional valence and perceived productivity, with a stronger correlation in the afternoon. Both social and individual breaks emerge as useful for restoring a positive mood. Furthermore, we found that a minimum set of non-invasive biometric sensors can be used as a predictor for emotions, provided that training is performed on an individual basis. While promising, our classifier performance is not yet robust enough for practical usage. Further data collection is required to strengthen the classifier, by also implementing individual fine-tuning of emotion models.

Index Terms—Emotion awareness, emotion detection, biometric sensors, empirical software engineering, human factors

1 Introduction

Affective states such as personality traits, attitudes, moods, and emotions play a crucial role on people's everyday performance at work, especially for activities that require creativity and problem-solving skills [1], as software development. Programmers experience and express different emotions [2] during their daily work, which may have an impact on job performance.

According to Graziotin et al. [3], happy software developers achieve better performance. Conversely, unhappiness brings developers to lose motivation in completing tasks and to leave the company [4]. The relation between positive emotions and self-assessed productivity was also confirmed by recent lab studies [5], [6], which also investigate the triggers for emotions experienced by developers during programming tasks. Being stuck and working under time pressure emerged as the most frequent causes for negative emotions, as well as unexpected technical difficulties and unfulfilled information needs. Along the same line, a recent field study at Microsoft investigated what makes a working day a good day for software developers [7]. The authors found that good workdays increase job satisfaction, which is reported as associated to the perception of contributing value to a project.

In this study, we focus on the emotions experienced by software developers at the workplace. Consistently with previous research on developers' emotions during programming tasks [6], [8], [9], we operationalize emotions along continuous dimensions. Following Russel [10], we describe the emotion stimulus in terms of its (un)pleasantness, ranging from low to high *valence*, and level of activation, ranging from low to high *arousal*. Furthermore, we include consideration of *dominance*, that is a person's perception of being in control of a situation. A priori, one might have thought that developers, being human beings, should experience the entire range of emotions at the workplace. However, different professionals have been shown to experience and express different ranges of emotions while at work: e.g., Foster and Sayers [11] reported about physiotherapists not experiencing calmness and serenity, which in our terms would correspond to high valence and low arousal. As such, we formulate our first research question as follows:

RQ1 What is the range of developers' emotions at the workplace?

As a second goal, we aim at investigating the relationship between self-reported emotions and productivity at the workplace. Previous studies conducted in a laboratory setting [6], [8], [9] report a positive association between emotional valence and self-assessed productivity of software developers engaged in a programming task. We seek to confirm and extend these findings in an *in-vivo* setting, by also expanding the observation period to the entire workday. As such, we formulate our second research question as follows:

RQ2 To what extent are developers' emotions related to self-assessed productivity during the workday?

Being able to identify the most frequent triggers for positive and negative emotions of developers enables informed decisions about the organization of work, towards improving the general well-being as well as the productivity of the individuals and the teams [7]. As such, we formulate our third research question:

- *D. Girardi, F. Lanubile, and N. Novielli are with the University of Bari, Italy. E-mail: daniela.girardi@uniba.it, filippo.lanubile@uniba.it, nicole.novielli@uniba.it*
- *A.Serebrenik is with the Eindhoven University of Technology, The Netherlands. E-mail: a.serebrenik@tue.nl*

RQ3 What are the triggers for developers' positive and negative emotions at the workplace?

Finally, we believe that enabling early detection of developers' emotions while at work might be useful to support their productivity and well-being, e.g. by suggesting just-in-time corrective actions thus preventing unhappiness and burnout, which might eventually lead to undesired turnover [12], [13] or by implementing strategies to support emotional awareness [14]. We envision the emergence of tools supporting the developers' well-being, leveraging non-invasive biometric sensors for timely and effective identification of negative emotions. Towards this goal, we aim at assessing the performance of a sensor-based classifier for emotional valence. As such, we formulate our fourth research questions as follows:

RQ4 To what extent we can predict the emotions of software developers at the workplace using lightweight biometric sensors?

To address our research questions, we performed a field study with 21 participants from 5 companies monitored for a minimum of two weeks during their daily activities. We asked participants to periodically self-report their emotional state, the performed activity, and their perceived productivity. Furthermore, we asked them to explain the causes for the reported emotions.

To answer RQ1 and RQ2, we analyze the range of emotions reported and their relation with perceived progress by fitting a linear-mixed model, as in previous work [3], [5], [6]. We collect self-reports about developers' emotions using a 5-point pictorial scale for each emotional dimension. To account for individual differences in self-reporting, we standardize scores before investigating the correlation with productivity, in line with previous research [5], [6], [8].

To answer our next research question (RQ3), we perform a qualitative analysis of the data collected through self-report. Specifically, we execute a coding study aimed at discovering the causes of positive and negative emotions experienced by software developers at work.

As for sensor-based emotion recognition (RQ4) we use supervised machine learning to train a classifier for developers' emotions based on biometric features. We rely on a minimal set of biometrics including the galvanic skin response and heart-related measurements, collected using a wristband, in line with previous findings that such sensor configuration is reliable in a lab environment [6].

The contributions of this work are as follows:

- We develop a taxonomy of emotional triggers related to software development at workplace;
- We build and assess a supervised classifier for developer's emotions at the workplace based on developers' biometrics collected using non-invasive sensors;
- We confirm and extend previous findings from lab studies by investigating the relation between emotions and perceived productivity during workdays;
- We build and distribute a lab package to verify, replicate, and build upon the present study[1].

1. Available at: https://github.com/collab-uniba/biometrics

The reminder of the paper is structured as follows. In Section 2 we present the background and related work. In Section 3 the data collection protocol of the field study. Then, we provide empirical answers to our RQs inSections 4, 5, and 6. Finally, we discuss the implications and threats to validity in Section 7, and conclude in Section 8.

2 Background and Related Work

2.1 Emotion model

We ground our study on the categorization of emotions by Russel [15], also known as the Circumplex Model of Affect. The model represents emotions according to *valence*, i.e. the pleasantness vs. unpleasantness of the emotion stimuli, and *arousal*, i.e. its level of activation vs. deactivation. Pleasant emotional states, such as happiness, are associated with *positive* valence, while unpleasant ones, such as sadness, are associated with *negative* valence. Arousal describes the level of activation of the emotional state ranging from inactive or *low*, as in calmness or depression, to active or *high*, as in excitement or tension. Beyond valence and arousal, and in line with previous studies [3], [16], we measure affective states according to a third dimension known as *dominance* (or control), that is the extent to which an individual feels in control of the situation.

2.2 Emotion Awareness in Software Development

Why are developers (un)happy? Ford and Parnin [17] surveyed 45 software developers to identify the causes of frustration while programming. They provide a list of 11 categories, which include issues with program comprehension or poor tooling, personal issues, and fear of failure. Graziotin et al. [18] further contribute to the identification of causes for developers' emotions through a survey involving ∼2K developers. Among the top 10 frequent causes of unhappiness, they include being stuck in problem solving, time pressure, dealing with bad code quality or underperforming colleagues, feeling inadequate or suffering from personal issues not related to work, dealing with bad decision making or investing time in mundane repetitive tasks.

Two lab studies leveraged experience sampling to identify the reasons for positive and negative emotions while programming [5], [6]. They found that developers get annoyed by low perceived productivity while they feel happy when *in flow*. Other reasons for negative emotions are related to cognitive difficulties, impossibility to fulfill information needs, and code not working [5], [6].

In the current paper, we ground our investigation of the triggers for developers' emotions on the findings of the aforementioned studies. We use their findings to compile an initial list of codes employed in the qualitative analysis of the developers' answers collected during our study. We extend the list and organize the results of our qualitative analysis in a taxonomy of emotion triggers (see Section 5).

Emotions and Productivity. Findings from recent studies appear to converge towards the claim that 'happy developers solve problems better' [19]. Graziotin and colleagues [4] found consequences of happiness and unhappiness experienced by software developers. Specifically, they report on the impacts of emotions and how they are beneficial or detrimental for the developers' well-being, the

software development process, and the quality of artifacts. Wrobel [20] surveyed 56 programmers and found that positive emotions increase developers' productivity while negative ones decrease it. Graziotin and colleagues conducted a series of studies on the correlation between developers' (un)happiness and their creativity, and analytical problem-solving skills of software developers [19]. In a controlled experiment, Graziotin et al. [8] empirically assess the correlation between developers' perceived productivity and self-reported valence, arousal, and dominance. They found a correlation between the productivity and both valence and dominance dimensions. These findings have been confirmed by Müller and Fritz [5] as well as Girardi et al. [6].

In this study, we reuse and adapt the experience-sampling approach from Müller and Fritz [5] and Girardi et al. [6] to perform an *in-vivo* study in five software development companies. By doing so, we aim at overcoming the limitations posed by the *in vitro* nature of lab studies.

2.3 Sensor-based Emotion Detection

The link between emotions and physiological feedback is widely investigated by affective computing research, which leverages a broad range of biometric measurements as predictors of emotions. *Electroencephalography* (EEG) records electrical activity of the brain through electrodes placed on the surface of the scalp. Variation in the EEG spectrum have been successfully used as a proxy for arousal or alertness [21] as well as emotional valence [22], [23] *Electrodermal activity* (EDA) is a measure of the electrical activity of the skin due to the variation in human body sweating. EDA varies consistently with intensity of emotions, with more evident changes for high arousal and emotional intesity [24]. Thus, EDA has been employed to detect excitement, stress, interest, attention, as well as anxiety and frustration [25] *Heart-related measurements* have been successfully leveraged for emotion detection [26]. They include heart rate (HR), which is number of contractions of the heart (beats) per minute as well as its variation in the time interval between two consecutive heartbeats, called heart rate variability (HRV). HR can be derived from the blood volume pulse (BVP) obtained by using of a photoplethysmography sensor. Bradley and Lang found that heart rate slows down when people feel negative emotions [24]. *Eye-related measurements* have been also used for emotion detection. For example, gaze duration is greater when people look at emotional pictures compared to neutral ones [27], while changes in the pupil dilatation indicate mental effort and affective responses [28]. *Electromyography* (EMG) captures the electrical activity in tissues, bones and skin due to the muscle contraction. Affective computing studies use facial muscle contraction, e.g., due to smiling and frowning, as an indicator of emotions [29].

Sensing developers' emotions. Vrzakova et al. [30] used eye gaze and EDA for classifying developers' valence and arousal during code review. They conducted an in-situ experiment with 37 software developers working on code reviews. The results show that the eye gaze is the most predictive measurement both for valence and arousal (accuracy=85.8% and 76.6%). However, considering the features of all signals in combination, including EDA, authors achieve even better results both for valence and arousal (accuracy=90.7% and 83.9%). Müller and Fritz trained a supervised emotion classifier able to distinguish between positive and negative valence with an accuracy of 71% [5]. They train a classifier using biometrics from 17 participants working on two programming tasks. The model achieves the best performance using a combination of EEG, EDA, HR, and eye-tracking metrics. Girardi et al. ran an empirical study aimed at identifying the minimal set of non-invasive biometric sensors for emotion recognition during programming tasks [6]. They trained two supervised classifiers for valence and arousal using as a gold standard the emotions self-reported by 23 participants during a Java programming task. They identified a minimum set of sensors—EDA, BVP, and HR measured using the Empatica E4 wristband—that can be used in an experimental protocol for detecting emotions during software development tasks. Specifically, using the wristband only they achieve an accuracy for valence ($.71$) and arousal ($.65$) comparable to the one obtained using the full sensors settings (i.e., wristband + EEG helmet). As such, in the present study, we use the Empatica wristband only for measuring both EDA and heart-related biometrics.

3 Study Design

3.1 Pilot study

The study design [31] was consolidated through a pilot study. Three professional developers working for a software development SME in Bari, Italy, were asked to wear the Empatica wristband for one week, during which they reported their emotions and activity using the pop-up application. After the pilot was concluded, we engaged with them in individual follow-up interviews, asking for their feedback about the study. The developers confirmed that reporting emotions once per hour through the pop-up was not annoying and that the Empatica wristband was comfortable to wear. In addition, they gave us two suggestions: i) adding "Just arrived" to the list of activities for filling the pop-up at the beginning of the day, when developers has not started any activity yet and ii) recommending to the future participants to insert the pop-up application in the startup folder of their own PC, in order to not forget to start it.

3.2 Companies and Participants

Five Dutch software development companies participated in the study: one startup (1 founder and 2 employees), two SMEs (between 20 and 200 employees), and two large companies (> 20.000 employees). The companies participating in our study develop software for a wide range of applications, including software for food-sorting technology, integrated circuits and semiconductor-based products, tools and data-driven products to support healthcare systems, manufacturing systems, and cyber-security applications, as well as a broad range of IT products also for domestic and personal use. All teams, whatever the company, follow an Agile software development method according to the principles behind the Agile Manifesto [32], which involves an iterative and incremental software process style to encourage fast reaction to changes and frequent communication. Among the agile principles there is the need for

running retrospective meetings, in which the team discusses problems, identify (un)effective solutions, and report individual feelings to improve self-organization. Retrospective meetings are conducted at the end of each iteration, which can be 2-4 week long. The term 'iteration' is also known as 'sprint', which has been popularized by Scrum[2], the most popular Agile method.

We recruited participants among professional software developers, on a voluntary basis. In order to recruit participants, the first author (hereinafter *experimenter*) organized dedicated meetings (one for each company) to explain the purpose of the study, the participants' role, the protocol of the experiment, and the possible risks and benefits in participating. Overall, 21 developers participated in the study (18 men, 3 women), with an average age of 33 years ($\pm$ 7.2, ranging from 23 to 50). Participants reported an average experience in software development of 8 years ($\pm$ 6.2, ranging from 1 to 25).

3.3 Instrumentation

Empatica E4. The Empatica E4 wristband[3] is equipped with an EDA sensor and a BVP sensor, where the latter is used to derive HR and HRV. Following the Empatica guidelines,[4] we excluded HRV because it is considered unreliable in presence of body movement as in our study. In fact, our participants wear the wristband for the entire day, including breaks, in which it is not uncommon for them, based on what they report, to take a walk. EDA and BVP are recorded with a sample frequency of 4Hz and 64Hz respectively.

Self-report of Emotions and Productivity. We use experience sampling [33] to collect developers' emotions and perceived productivity during working days. This choice is consistent with the methodology adopted by lab studies on emotions and productivity conducted by Müller and Fritz [5] and Girardi et al. [6]. We developed a pop-up[5] to self-report the valence, the arousal and the dominance scores using the Self-Assessment Manikin (SAM), in line with recommendation by Graziotin et al. [34]. SAM is an assessment technique for reliable self-report of emotions in terms of valence, arousal, and dominance. In their empirical study on measuring emotions, Bradley and Lang [35] demonstrated that the SAM approach is effective in measuring a person's affective reaction to a wide variety of stimuli in many contexts. SAM implements a non-verbal pictorial assessment technique consisting of 5 figures for each emotional dimension (see Figure 1). We use a 5-point pictorial scale for each emotional dimension, as this scale is easily understood and widely used in studies aiming at collecting affective states [3], [29].

In addition to the emotions, we ask participants to report the activity in which they were involved at the moment of the interruption and their perceived productivity. As for activity, we provide a drop-down list, based on previous work by Meyer et al. [7] including: *coding, bug fixing, testing, design, meeting, email, helping, networking, learning, administrative task, documentation, just arrived, other*. We also include

2. https://www.scrum.org/resources/scrum-guide
3. https://www.empatica.com/en-eu/research/e4/
4. https://support.empatica.com
5. https://github.com/collab-uniba/ExperienceSampling

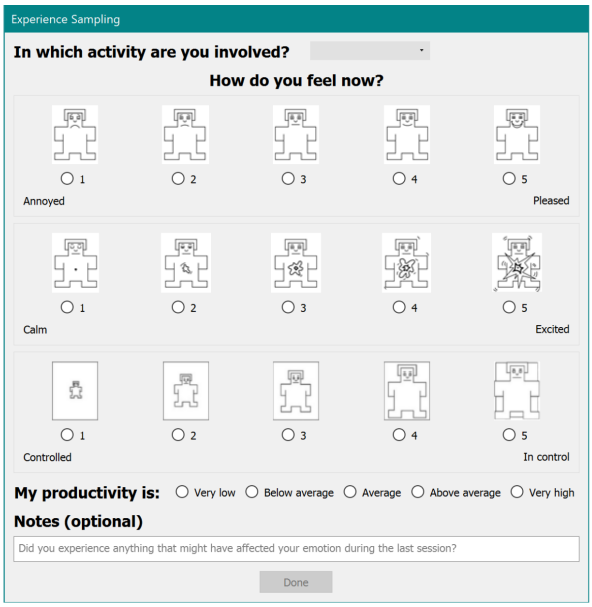

Fig. 1. The pop-up window to elicit perceived emotion and productivity

the 'just arrived' label that was added as a result of the pilot study (see Section 3.1), to be used at the beginning of the workday, upon arrival at the office. For perceived productivity, we use a 5-point Likert scale (from *Very low* to *Very high*), in line with previous studies [3], [6], [36]. Finally, we ask the participants to motivate the ratings provided, i.e. to explain the causes for the reported emotions.

3.4 Study set-up

The day before starting the experiment, the experimenter met the developers involved in the study. During this meeting, the experimenter demonstrated how to wear the wristband in order to get the proper acquisition of biometric signals. Then, she illustrated how to download and install both the pop-up application and the tool, called *E4 manager*, for transferring data from the wristband to the participants' computer. Next she explained how to use the SAM scales for self-reporting valence, arousal, and dominance. A printed image of the Circumplex Model of Affect was also provided to help participants in correctly rating their emotions. Then, the experimenter answered the participants' questions and then they signed the informed consent form. Finally, a private Dropbox folder was created for each participant, to allow them sharing data with the experimenter.

3.5 Experimental Protocol

The day after the set-up is completed, the experiment can start. For each participant, we observe and collect data for two or three weeks, based on the agile iteration length at the company, thus covering all key technical activities. Every day, the participants follow the steps reported in Figure 2.

Upon arrival at the office, participants wear the Empatica E4 and run the pop-up application. By default the pop-up appears on the participant's monitor once per hour. We define this interval inspired by the study design of Meyer et al. [37], who studied the developers' productivity

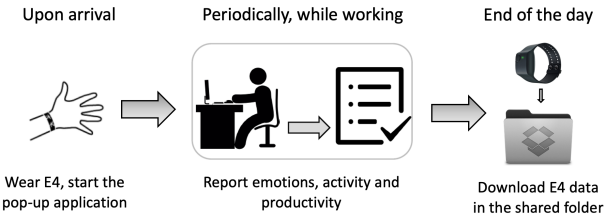

Fig. 2. The developer's working day during the study

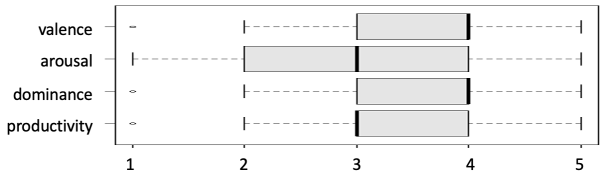

Fig. 3. Developers' valence, arousal, dominance, productivity at work.

using an analogous pop-up. Specifically, they report that 60 minutes was a good balance between the intrusiveness and the necessity of collecting as much data as possible, as also emerged during our pilot study. When the developers do not want to be interrupted, they can postpone answering the pop-up by specifying the delay in minutes. To reduce intrusiveness of the pop-up we follow the recommendations of Meyer et al. [37] and allow the participants to dismiss the pop-up for the entire day. Conversely, the participants can invoke the pop-up manually, when experiencing strong emotions that believe are important to be reported.

At the end of the day, the participants export the pop-up data as a .csv file, containing one row for every self-reported episode with the corresponding timestamp. Then, they download the data from the wristband using the E4 manager application. Finally, all data are added to the folder shared with the experimenter. At the end of the first week, the experimenter reviews the data obtained to check for consistency and completeness. Should additional information be required, she contacts the participants via e-mail. At the end of the study, the experimenter visits the company to collect the wristbands. Before providing the wristbands to the new participants, personal data are removed from the Empatica E4 devices. During the final meeting, the experimenter also provides a company-level overview of the emotions experienced by the participants including information about the main emotion triggers.

The study protocol has been approved by the ethical review board of Eindhoven University of Technology.[6] The main ethical concerns of the study were related to collection of personal data such as names and biometric measurements. To address privacy concerns, we enabled data sharing through Dropbox Business as it is compliant to the European General Data Protection Regulation policy. By doing so, we were able to solve privacy issues in terms of protection against unauthorised data access. This was also made explicit in the ethical review board application as well as in the consent form signed by the participants.

4 Self-reported Emotions and Correlation with Productivity

4.1 Dataset

Each participant reported emotions either for two or three weeks, depending on the duration of the agile iteration. Overall we have self-report data for 192 days out of 240 overall days of the study. The missing days are mainly due to participants not being at work, either for vacation, or health/personal issues (31 days). Among these, one of the participants was always off on Thursday (3 days overall over three weeks dedicated to the study). For three participants working at the same company we are missing the last day only (3 days overall), which might be due to internal reasons. Finally, we have 5 missing days for which the participants did not provide any explanation. Overall, we miss data for 42 days (17.5% of the total days of the study).

On average, participants filled the pop-up 5.4 times per day (sd = $\pm$ 1.24). Overall, we collected 1255 self-reports. For the purpose of this analysis we excluded the cases where the participant reported to be "just arrived", which identifies the first self-report made by the developers as soon as they arrive at the workplace. As such, information about self-perceived productivity is not available for these instances. Thus, we consider these instances as not relevant for the study of the correlation between self-reported emotions and productivity. After this step, 1098 self-reports remained, of which 404 collected during before 12PM (morning) and 694 after 12PM (afternoon).

4.2 Developers' Emotion at the Workplace (RQ1)

The boxplot in Figure 3 shows the SAM scores the developers reported during the study. We observe that the entire range of emotions is covered by the scores reported, in line with findings of the lab experiment by Girardi et al. [6]. For both valence and dominance the average self-report score is 4, indicating that developers mostly experience pleasant emotions and feel in-control of the situation. For arousal, developers report on average a neutral state (SAM score = 3) and the distribution is well balanced between low (SAM score = 1) and high arousal (SAM score = 5) arousal. For productivity, developers report the whole range of values from very low (score = 1) to very high (score = 5), with an overall tendency to report average (score = 3) to above average productivity (score = 4), as previously observed in the lab study by Müller and Fritz [5].

The observed range of valence scores is in line with findings by Graziotin et al. [18], reporting that a Github developers prevalently self-report positive emotions, thus depicting themselves as moderately happy population. Analogously, Müller and Fritz observed that the professional developers involved in their lab study reported the full range of valence values, with an average score corresponding to slightly positive valence [5]. In both studies, the participants hold an experience of about 7-8 years, which is comparable to the one of the developers involved in our field study. Conversely, Girardi et al. [6] mostly observed students reporting negative emotions and high arousal while coding, which can be explained by the students being less experienced and feeling less confident in solving the assigned programming task compared to professional software developers.

6. Approval number: 2019ECMCS02.

A possible interpretation of our results can be provided in the light of previous findings by Mäntylä et al. [16], presenting empirical evidence that novice developers are more inclined to negative valence and high arousal. Furthermore, experience has been reported as negatively correlated with effort—i.e., more experienced developers need less effort to complete a task [12], [38]. In line with this interpretation, the lower level of experience of the students in the previous lab study by Girardi et al. [6] is reflected in the fact that they mostly reported being stuck. Conversely, we observed a more balanced distribution, slightly skewed in favor of positive self-assessed productivity (see Fig. 3), in line with Müller and Fritz [5] reporting a more balanced distribution of progress with the majority of participants feeling in flow. As a further confirmation of this interpretation, we observe an average dominance score of 4 (see Fig. 3), indicating a good self-assessed confidence by the participants.

Takeaway message for RQ1 - Developers report a wide range of emotions at the workplace. We observe a prevalence of positive valence, neutral arousal, and high dominance, indicating they mostly experience pleasant emotions and feel in control of the situation.

4.3 Emotions and Productivity (RQ2)

We study the correlation between self-reported emotions and productivity by fitting a linear mixed model, which is robust in case of repeated measurements and longitudinal data [39]. To create the model, we used the `lme4` R package.[7] Consistently with the approach adopted in the former studies [5], [8], we consider productivity as the dependent variable and valence, arousal, dominance as fixed effects. Given our study design, we cannot exclude that the perceived productivity can be impacted by time, e.g. due to fatigue [17]. Therefore, time and its interaction with the emotional dimensions are also included in the model as fixed effects. Specifically, we model time as the part of the day (morning or afternoon) during which participants answered the pop-up. This choice is justified by findings of psychology research by Stone et al. [40], investigating diurnal rhythms of emotions during working days. They report a strong bimodal pattern for both positive and negative emotions, with differences in emotion peaks before and after lunchtime. In our study, we model as morning/afternoon the self reports made before/after 12PM, corresponding to lunch break in the Netherlands.

To account for individual differences in self-reporting emotions due to a personal perception of the SAM scale, we use Z-scores to standardize the raw scores, as already done in previous work [5], [6], [8]. Finally, to account for differences due to personal (e.g., personality) or environmental (e.g., company) factors, we also include participants and companies as random effects.

In Table 1.a, we report the parameter estimation for the mixed model and the percentage of deviance explained by each effect. We observe a statistically significant correlation with perceived productivity for valence, arousal, and dominance. The marginal R^2_m, that is the total variance explained by the model through the fixed effects, is 0.21, indicating that the changes in productivity are accounted by the differences between emotions. The conditional R^2_c, that is the proportion of total variance explained through both fixed and random effects, is 0.21. Thus, grouping the measurements by participants and by company does not contribute to the model explanation.

TABLE 1
Parameter estimation for the fixed effects on perceived productivity ($*$ indicates a statistical significance with $\alpha = 0.05$).

Fixed Effects	Estimate	p-value	Dev. explained
(a) Full model			
Valence (*)	0.30	0.00	12.15%
Arousal (*)	0.21	0.00	2.75%
Dominance (*)	0.19	0.00	3.91%
Time (*)	0.16	0.00	0.66%
Valence:Time (*)	-0.22	0.00	0.87%
Arousal:Time	-0.09	0.09	0.22%
Dominance:Time	0.07	0.24	0.10%
(b) Morning vs. Afternoon			
Morning (404 answers)			
Valence (*)	0.08	0.00	6.25%
Arousal (*)	0.11	0.02	1.30%
Dominance (*)	0.27	0.00	7.23%
Afternoon (694 answers)			
Valence (*)	0.30	0.00	15.57%
Arousal (*)	0.21	0.02	3.79%
Dominance (*)	0.19	0.00	2.52%

Specifically, valence shows the highest explanatory power with the 12.15% of deviance explained. Conversely, the effect of the arousal and the dominance appears negligible (respectively 2.75% and 3.91%), in line with results of previous lab studies [3], [5], [6]. Beyond confirming previous finding, we observe the impact of time, which is an additional finding. Indeed, the aforementioned lab studies were performed in a limited amount of time, ranging from 30 to 90 minutes of observation. Conversely, we could rely of data collected throughout the entire working day. In particular, we observe that time has a significant effect on the productivity in the interaction with valence.

To further investigate this aspect, we split the dataset in two subsets and repeat the analysis for morning vs. afternoon. The first subset includes 404 data points corresponding to answers provided during mornings, while the second one includes 694 answers collected in the afternoons. We report the results of the two separate models in Table 1.b. R^2_m is 0.21 and 0.15 for morning and afternoon, respectively. We observe that in the afternoon the correlation between emotions and productivity is stronger than during the morning. In fact, the coefficient estimate for valence is 0.30 in the afternoon, with a deviance explained = 15.57%, which is higher than what observed in the morning but also in the general model reported in Table 1. Conversely, dominance seems to have a stronger positive correlation with the perceived productivity in the morning (estimate = 0.27, deviance explained = 7.23%) than in the afternoon (estimate = 0.19, deviance explained = 2.52%).

While we could not provide any causal explanations—whether negative emotions reduce productivity or, conversely, lower productivity triggers negative emotion—this evidence suggests that fatigue might play a mediating role

7. https://cran.r-project.org/web/packages/lme4/index.html

in the relationship between negative emotions and perceived productivity, assuming that developers become more tired towards the end of the working day (i.e., in the afternoon). This suggestion is consistent with previous results reporting fatigue as a cause for negative emotions [17]. We can also interpret this results in light of previous findings by Sarkar and Parnin [41]: the authors found that fatigue harmed developers' productivity as well as the quality of their work, creativity, and motivation. Previous evidence in psychology [42] corroborates this interpretation, indicating that mental fatigue following cognitive tasks impairs emotion regulation. This would explain the stronger correlation between valence and productivity in the afternoon, suggesting that developers might be more successful in restoring their positive mood in the morning, when they are less tired.

Takeaway message for RQ2 - Valence is positively correlated with perceived productivity, with stronger correlation in the afternoon. Conversely, the correlation between dominance and productivity is stronger in the morning. This could be due to fatigue, which is known to impair emotion regulation.

5 Taxonomy of Emotion Triggers

5.1 Methodology

We manually analyzed the developers' answers to the open-ended question about the causes for the self-reported emotions. We performed qualitative data analysis in a semi-exploratory mode, by adopting the answer as annotation unit. Overall, we collected 350 answers describing the reasons for emotion scores. We decided to manually code all answers received, including those with neutral valence, to account for possible inconsistencies between the answers and the self-reported valence scores.

During the first coding iteration, two authors analysed 100 responses, randomly extracted from the full set of 350 answers. They used a closed coding approach [43], starting from a list of 48 codes derived from previous studies investigating the causes for positive and negative emotions of software developers [5], [6], [17], [18]. For example, Müller and Fritz identify *feeling-in flow* and *being stuck* as triggers for positive and negative emotions, respectively [5], while Ford and Parnin [17] found that *fatigue* is one of the triggers for frustration. Hence, we include *feeling-in flow*, *being stuck* and *fatigue* in the initial list, together with 45 others derived from the aforementioned studies. The two authors labeled each answer using codes from the initial list, thus creating a preliminary taxonomy. They could also assign to each answer more than one code, if needed. Furthermore, they could add new codes when they failed to find the most appropriate one in the list. Upon completion of the individual coding round, the entire team discussed the results to solve disagreements and reconcile the newly added codes. As result of this iteration, 11 new codes were added to the initial list, thus resulting in 59 codes overall.

We re-coded the initial 100 answers according to the final set of 59 codes and repeated the annotation on the remaining 250 answers. During this second round, a third author was involved in the coding such that each answer was coded by two people. We reached saturation after this second round, as only one code was added to denote answers describing emotion triggers related to the study itself (i.e., *Meta*, see Table 2). Once again, the entire dataset of 350 answers was re-coded to incorporate the new code.

After this coding round, we removed from the dataset the 68 answers for which a neutral emotion score was provided, in line with our goal of identifying a taxonomy of triggers for positive and negative developers' emotions at the workplace (RQ3). For the same reason, we filter out further answers. Specifically, we removed the 7 answers referring to *meta* topics, as they do not provide any useful information regarding the developers' activity in relation to their workday. Furthermore, we discarded 43 answers where the participants did not answer the question but rather provided a generic description of the activity performed rather than an explanation of the trigger for the self-reported emotion. Overall, this cleaning step resulted in a final dataset of 232 answers, of which 157 (68%) and 75 (32%) are associated with positive and negative emotions, respectively. 9 answers received two codes, thus resulting in 241 codes overall. We used this dataset to finalize the taxonomy. First, we included the code used for the 232 answers describing the triggers for either positive or negative emotions. Then, we grouped them to capture relationships and themes by applying axial coding [44]. This was done through two further iterations consisting in plenary meetings involving all the authors.

5.2 Results (RQ3)

The taxonomy of emotion triggers in shown in Table 2. Six themes emerge: *self* refers to the developers themselves, *developer-task relation* describes the link between the developer and the task, *artifacts and instrumentation* include properties of tools and source code as triggers for emotions, *social* refers to peers and collaborators, *work management* refers to issues with artifacts, design, and implementation of the task. *Non-work time* is also included to cluster individual and social breaks. Overall, we identified 18 triggers for positive emotions and 19—for negative emotions. In addition, we looked into data to verify what are the triggers reported for extremely negative (valence score = 1) and extremely positive (valence score = 5) emotions. In commenting Table 2, we report these observations as well, where relevant. Furthermore, we offer a comparison with previous work on causes for developers' emotions [5], [6], [17], [18] we build upon in defining our initial set of codes.

Self. The most frequent trigger for emotions refers to the *self* dimension (11 codes, 95 occurrences overall). The participants reported *feeling in-flow* as the main cause for positive emotions, that is a distraction-free state while writing code or performing coding-related activities (*"Organizing my workflow a little better than usual. Feel optimistic about getting a lot of work done on this feature today."*). Analogously, being able to conclude assignments (*task completed*) relates to a sense of accomplishment associated with positive emotional valence (*"Finished coding the today's task"*). Conversely, *being stuck* is a cause for negative emotions (*"Getting very annoyed by an annoying bug I can't solve."*), thus confirming

TABLE 2
Triggers for positive and negative emotions at workplace identified after coding. Code occurrences in the corpus are reported in parentheses.

Theme	High-level code	Codes denoting triggers for Positive emotions	Codes denoting triggers for Negative emotions
Self *11 codes (95)*	Productivity	Feeling in-flow (37) Start working (1) Task completed (11)	Being stuck (8)
	Perception of self	Feeling confident (4)	Feeling inadequate (3)
	Mental state		Fatigue (10)
	Personal issues	Personal positive facts (6)	Personal problems (11)
	Motivation	Boost of motivation (4)	Lack of motivation (1)
Developer-task relation *7 codes (28)*	Novelty of the task	Thrilled by new challenges (3)	Mundane or repetitive task (2)
	Code Comprehension	Understanding relevant code (2)	Problems in Mapping Behavior to Cause (11)
	Solution design	Clear solution design (7)	Complexity (2)
	Learning		Learning curve (1)
Artifacts and instrumentation *5 codes (33)*	Code quality	Working code with no errors (15)	Poor code (4)
	Tooling	Adequate tooling (2)	Poor tooling (8)
	Documentation		Unavailable or insufficient documentation (2)
Social *5 codes (26)*	Social feedback	Feeling appreciated (3)	
	Collaboration	Collaborative problem solving (11) Helping peers (8)	Peers not helping (2) Helping peers (2)
Work management *7 codes (26)*	Meetings	Constructive meeting (8)	Unconstructive meeting (3) Long meeting (1)
	Decision making		Bad decision making (3)
	Customers	Helping customers (1)	Problems with customers (4)
	Time management		Time pressure (6)
Non-work time *2 codes (34)*	Break	Individual break (25) Social break (9)	

the positive association between emotional valence and self-perceived productivity we report in Section 4 and also observed in previous lab studies by Müller and Fritz [5] and Girardi et al [6]. Being stuck in problem solving is reported as a major cause for negative emotions by Graziotin and colleagues based on a large-scale qualitative survey among developers [18].

In line with previous findings [18], we found that the developers' mental state and personal issues also impact their emotional state at the workplace. *Fatigue* is reported as a trigger for negativity (*"very tired from yesterday... double shift to finish stuffs"*), as well as personal issues (*"Not feeling particularly healthy in my mind, but it doesn't seem to get into the way of work right now."*). The developers' perception of self also triggers emotions. The solution to a programming task being perceived as easy or known (*feeling confident*) triggers positive emotions (*"I feel that I am close to solve the problem!! Feel excited!!"*). Conversely, *feeling inadequate* leads to negative emotions (*"Past and present failures on my mind. Feeling down."*). Perception of self as inadequate or under-qualified with respect to a given task was already reported as a cause for unhappiness by Graziotin et al. [18]. Similarly, Ford and Parnin mention the fear of failure as a trigger for frustration [17]. Fatigue and personal issues not related to work were also reported by these two studies among the causes for developers' negative feelings.

Developer-task relation. This dimension includes codes that link the developer with the task. Having difficulties in source code comprehension, i.e. wondering why the code does not produce the expected behavior, which part of code is causing a problem or not having a clear understanding of the code and its functioning (*Problems in mapping behavior to cause*) are reported as triggers for frustration and other negative emotions (*"Digging up old reproduction data for tickets that were implemented more than half a year ago. While circumstances for only being able to test this now were outside our control, this is still annoying."*). Conversely, being able to identify and *understand relevant code* is associated with positive emotions (*"I've found a starting point almost immediately. I have a good feeling that I'll make significant progress very soon."*), as well as having a *clear solution design*, i.e. a knowing what to do next and how to reach the solution (*"I am having a plan how I would like to implement the given task in the algorithm. Therefore I now what I need to do and in which order."*), which confirm evidence from previous lab studies [5], [6].

The codes included in this dimension have been broadly discussed by previous research [5], [6], [17], [18]. As for code comprehension, unexplained broken code and unexpected output are reported as causes for negative emotions by Graziotin et al. [18] and Girardi et al. respectively [6]. Analogously, Müller and Fritz observe a decrease of emotional valence when developers experience difficulties in understanding how parts of the code or API work. Conversely, a positive shift valence is observed when developers are able to localize relevant code [5], which we also confirm.

As for the novelty of the task, we observe that dealing with *mundane or repetitive tasks* is associated with negative emotions. This was already reported among the top ten causes for developers' unhappiness [18]. On the opposite side of the valence spectrum, we observe that dealing with new tasks can be a cause for positive emotions, with developers feeling *thrilled by new challenges*. This is in line with previous findings by Girardi et al. [6].

Finally, Ford and Parnin also report developers being frustrated when adjusting to a new project or environment, which relates to the *Learning curve* code included in our taxonomy [17].

Artifacts and instrumentation. Poor quality artifacts and inadequate instrumentation cause negative emotions (5 codes, 33 occurrences), with the most popular trigger being *poor tooling*, i.e., limited, inadequate, or buggy tools, programming languages, IDEs, or hardware: *"The computer*

froze during my work". Negative affects also occur when dealing with *poor code* that needs to be reused or changed (*"A bit frustrated because of the illogical current structure of the code"*). In particular, two out of four occurrences indicating poor code as a trigger for emotions is associated with strong negative emotions. Conversely, *working code with no errors* is associated with positive valence (*"It looks like my change is working!"*). In particular, in seven out of 15 occurrences reporting working code as a trigger for positive emotions, the self report is associated with strong positive valence.

These findings are in line with Graziotin et al. reporting bad code quality and coding practice among the most frequent causes for negative affect of developers [18]. Developers involved in their survey mostly complained about bad code written by colleagues and only a few times reported being annoyed by poor code written by themselves.

Developers also complain about *poor tooling* (*"The computer froze during my work"*). This confirms previous evidence that issues in the technical infrastructure trigger developers' unhappiness [18]. Finally, *unavailable or insufficient documentation* was reported by two developers as a cause for negative emotion, in line with previous work [6], [17].

Social. Collaborating with others to solve a development task either to receive (*collaborative problem solving*) or provide support (*helping peers*) is associated with positive emotions, in line with previous findings [6] (*"Reviewed and had constructive discussion over where to take a feature"*). In particular, four out of eight of occurrences reporting helping peers as trigger for positive emotions is associated with strong positive valence. Two participants reported being annoyed when *helping peers* (*"Helping an outsourced guy in his development, but he is just asking every little think to me. Really annoying"*) and by the fact that nobody is available for help or peers are described as incompetent (*peers not helping*, i.e. *"some walking around to find the correct people was involved"*). This concurs with under-performing colleagues being among the top ten triggers for developers' unhappiness [18] and incompetent peers—among the triggers for developers' frustration [17].

Work management. Effective work management appears to positively impact developers' feelings, and vice versa. Developers report being happy when they invest their time in productive activities, as in *constructive meetings* (*"Review went well, though ran into one more problem."*). Conversely, *long* or *unconstructive meetings* are perceived as a waste of time and trigger negative emotions (*"Just had a terrible meeting!"*). Analogously, *time pressure* due to interruptions, approaching deadlines, or limited amount of time for completing tasks, causes sadness and stress (*"A bit sad for not having enough time to finish my work!"*). In particular, for three out of six occurrences reporting time pressure, the self report is associated with strong negative emotions (as in *"very tired from yesterday... double shift to finish stuffs"*). Negative emotions are also triggered by a manager' or colleague's (*bad decision making*), either because they are uninformed decisions or because task complexity was underestimated (*"Some decisions in management are being challenged, leading to arguments."*). Both, time pressure and bad decision making were previously reported as responsible for developers negative feelings [17], [18].

Non-work time is related to *individual* or *social* breaks during the working hours, e.g., for lunch or regaining focus and taking a rest. More than 40% of reportings associate break with an extremely positive emotion.

> *Takeaway message for RQ3* - Positive emotions are mostly triggered by the developers' perception of being productive, either because they feel-in flow or completed their tasks. Other causes are working code with no errors, successful collaborative problem solving, and constructive meetings. Negative emotions are mostly triggered by code comprehension issues, poor tooling, and fatigue. Personal issues not related to work are also a major cause, as well as developers' being stuck and dealing with poor tooling. Both social and individual breaks emerge as useful for restoring a positive mood.

6 Sensor-based Emotion Recognition

6.1 Our Vision

Supporting emotion awareness in software development could benefit both developer teams and individual developers. At the team level, companies could implement strategies to support emotion awareness, by self-reporting emotions during meetings. Indeed, emotional self-awareness is an antecedent of team effectiveness, as suggested by research in psychology [45]. In a case study conducted by Andriyani et al. [46], developers openly discussed their feelings during Agile retrospective meetings. Using biometrics, developers' emotions could be shared anonymously, also in an aggregate fashion, enabling the managers to measure the mood of the project and allowing developers to gain awareness of their colleagues' emotions, while avoiding the need for self-disclosure through discussion, to preserve privacy.

At the individual level, awareness of emotions could positively impact a developer's progress in their tasks [47]. Based on biometrics, developers can receive suggestions on how to regain focus and restore positive moods when negative emotional episodes or prolonged stress is observed using biometrics. In this perspective, detection of negative emotions while coding can be used as a prompt for recommender systems suggesting breaks to prevent bug introduction, the need for code reviews or pair programming, or links to Stack Overflow or curated documentation. We also envision the possibility to enhance the developers' self-emotional awareness by enabling the analysis of the log of their own emotions as detected by the biometric sensors, e.g. at the end of the day or at the end of the week. Such a scenario grounds in psychological research using biofeedback to raise emotion awareness and improve emotion regulation [48], which is the result of a continuous adaptive process [49], [50].

6.2 Dataset

Our goal is to investigate to what extent we can predict the emotions of software developers at workplace using lightweight non-invasive biometric sensors (RQ4). Since the

results of our regression analysis show a strong positive correlation between valence and perceived productivity of professional software developers at the workplace (see RQ2 in Section 4), we focus on the recognition of emotional valence using biometrics. We observe that developers mostly report positive valence, neutral arousal, and high dominance scores, i.e., a positive emotional state appears to be the usual condition of professional developers at the workplace. Thus, early detection of non-positive emotional episodes might enable just-in-time corrective actions in order to restore positive affect and productivity. As such, we design our machine learning study around the task of distinguishing between positive vs. non-positive emotions.

We build our dataset using the self-reported scores for valence as our ground truth. Specifically, we map the valence scores provided to the SAM questionnaires during the study to a binary rating of either *positive* (score > 3) or *non-positive* (score ≤ 3). [8] Among 21 study participants, 19 have shared with us the biometric data. Two participants did not wear the Empatica wristband during the experiment because the software for downloading the data was not supported on their operating system. From the self-report data points provided by the 19 participants, we excluded data points for which the biometrics were missing. This happened either because the participants forgot to turn on the device or because the wristband lost the signal due to the lack of contact between the sensors and the skin. Unfortunately, the device is not designed to send an alert to the user when the signal is missing. As a result, our dataset is composed of 759 self-reports with biometrics, of which 58% are labeled as positive and 42% as non-positive. To balance our dataset, we apply SMOTE [51] using the *SmoteClassif* function of the R UBL package.

6.3 Preprocessing and Features extraction

The biometric signals are recorded during the entire experimental session for all the participants. However, we only consider the signals recorded in proximity of the stimulus of interest—i.e., the signals collected in the 10 seconds before the participants provide the self-report about emotional valence using the pop-up. The choice of the interval is inspired by previous work on sensor-based classification of emotions [5], [6], [52]. In addition to considering a 10-second interval, we investigate a larger time frame because Züger et al. [53] found that a 3-minute interval might be optimal for extracting heart-related features. For the purpose of the machine learning study, we create two different versions of the datasets by considering features extracted in the two different time windows (i.e., 10″ or 3′). In 10 out of 759 responses, the data were not available for the three minutes before the interruption (for example, when the participant switched-on the wristband only one or two minutes before the self-report). The final distribution of the labels for valence is reported in Table 3.

To synchronize the measurement of the biometric signals with the self-reported emotions, we: *(i)* save the timestamp in which the participant fills the pop-up `(t_self-report)`, *(ii)* calculate the timestamp for relevant time interval for each interruption—i.e., 10 seconds (or 3 minutes, depending on the setting) before the self-report `(t_start)`, *(iii)* select each signal sample recorded between `t_start` and `t_self-report`.

8. We also experimented with the normalized scores, consistently with the approach adopted for the correlation analysis reported in Section 4, obtaining a performance comparable to the one reported in Table 5.

TABLE 3
Gold Standard in the two time windows for feature extraction.

10″	Positive	Non-positive	3′	Positive	Non-positive
	442 (58%)	317 (42%)		435 (58%)	314 (42%)

To account for differences in the biometrics between individuals, we normalize the raw signals following the approach previously used in [54], [55], which accounts for baseline fluctuations between days: $S_i^{\text{norm}} = (S_i - \mu)/(\max - \min)$. Then, we perform signal-specific preprocessing by following consolidated approaches. Specifically, we extract the *tonic* and *phasic* EDA components using the cvxEDA algorithm [56]. As for heart-related metrics, we filter the BVP signal using a band-pass filter, following the approach used by Canento et al. [26].

After signal pre-processing, we extract the features presented in Table 4, which we use to train our classifiers. We select features based on previous studies using the same signals [5], [6], [52], [57]. Some of these features are based on differences between the signals collected during the experiment and the signals collected while participants watched a neutral video (*baseline*). Since in our setting it was not possible to show participants videos, we consider as baseline the signals collected over the entire experiment, as done by Jacques et al [55].

TABLE 4
Machine learning features grouped by physiological signal.

Signal	Features
EDA	- tonic: mean, phasic: AUC, min, max, mean, sum peaks amplitudes
BVP	- min, max, sum peaks amplitudes - mean peak amplitude (diff. between baseline and task)
HR	- mean, sd. deviation (diff. between baseline and task)

6.4 Machine learning

We experimented with four machine learning classifiers, i.e., Support Vector Machine (SVM), k-nearest neighbor (knn), Decision Trees (J48) and Random Forest (rf), since they resulted the best performing algorithms in previous studies using biometrics for emotion detection in software development [5], [6], [30], [55]. In line with the previous lab study by Girardi et al. [6], we evaluate the classifiers in the *hold-out* and *leave-one-subject-out* (LOSO) settings.

In the *Hold-out* setting, we split the gold standard into train (90%) and test (10%) sets, in line with consolidated practice in affective computing [55] and using the stratified sampling strategy implemented in the *R* `caret` package [58]. We perform hyper-parameters optimization [59] using leave-one-out cross validation, as recommended for small training sets [60], such as ours. We evaluate the best model resulting from hyper-parameter optimization on the hold-out test set, to assess its performance on unseen data. We repeat this entire process 10 times to further increase the validity of the results. We evaluate the overall performance of the classifier by computing the mean of precision,

recall, F-measure, and accuracy over the different runs. This setting is directly comparable to the one implemented by Müller and Fritz [5], which includes data from the same subject in both training and test sets. Furthermore, it is directly comparable with the hold-out setting implemented by Girardi et al. [6], with the only notable difference that in both previous studies the task was to classify negative vs. positive emotions.

In the LOSO setting, we assess the classifiers performance on data obtained from unseen developers. We repeat the evaluation on a test set 19 times, i.e., the number of subjects in our dataset. At each iteration, we train the model on all the observations from the 18 participants, and we test the performance on the remaining one.

6.5 Classification Performance (RQ4)

In Table 5, we report the classifier with the best F1-measure, together with its precision, recall, and accuracy, for the two time windows considered. Furthermore, we provide the average over the performance of the ten runs for the best setting, together with standard deviation. We compare the classifier performance with the baseline classifier always predicting the majority class (in our case the *positive* class). In Table 6 we report the performance by class.

In the *hold-out setting*, the valence classifier outperforms the performance of the *baseline*. The choice of the time window considered for the feature extraction (10 seconds vs. 3 minutes) has a negligible effect on the average performance. As for the best performance, the classifier is substantially more precise than the baseline (+.38 and +31 for the 10-second and 3-minute settings, respectively). The improvement is smaller for Recall (+.17 and +10) with values ranging from .60 (3′ setting) to .67 (10″ setting). Overall, there is an improvement in F1 of +.30 and +23 for the 10-second and the 3-minute settings, respectively. As for the average performance, we observe lower precision, recall, and F1. The *valence* classifier is substantially more precise than the baseline (+.27 and +25 for the 10-second and 3-minute settings, respectively). The improvement is smaller for Recall (+.06 and +04). Nevertheless, there is an improvement in F1 of +.18 and +17 for the 10-second and the 3-minute settings, respectively. This is comparable to the average performance observed in our lab study using Empatica E4 only [6], where we observed substantially higher precision (.70) but comparable recall (.59), and F1 (.59).

Looking at Table 6 we observe that the major cause of error is due to misclassification of non-positive cases. While *positive* cases are recognized with good precision (.71) and recall = (.75), the *non-positive* cases tend to be misclassified as positive, as demonstrated by the lower recall (.58). Such difference in performance between the two classes might be due to a bias towards the majority class (i.e., *positive*), in spite of the use of SMOTE to balance the training set. Another possible explanation is that clustering neutral and negative valence in the *non-positive* class might introduce noise in the training. A better performance could be achieved by removing under-represented or irrelevant polarity classes, as done in the field study by Jaques et al. [55]. Specifically, they focus on classifying happy vs. unhappy days, after removing the 40% of instances of the dataset for which the participants reported average (i.e., neutral) emotional scores and report an overall accuracy of 64% using Empatica only. While beneficial in terms of noise reduction for the machine learning, such filtering would not be feasible in a natural setting, as the just-in-time emotion detection at the workplace scenario that we envision as our long-term goal.

The best *LOSO* setting results are comparable to the ones reported in the *hold-out* settings. However, we confirm the drop in performance, compared to the hold-out setting, already observed in the lab study by Girardi et al. [6], with the average of the LOSO setting. This is due to the variability for the individual performance on each participant test set, suggested by the higher standard deviation compared to the hold-out setting. Indeed, the accuracy among the 19 subjects varies from 0.29 (worst-performing model) to 0.79 (best-performing model). Differently from the hold-out setting, in the LOSO condition we observe better performance when extracting features in the 3-minute window before the self report, with peaks of precision, recall, and F1 up to .75, .86 and .75, respectively, for the best-performing model. Consistently with what observed for the hold-out setting, we report a better performance for the positive class (see Table 6). Again, we report comparable performance with respect to our lab study (Prec = .45, Rec = .61, F1 = .50) [6].

Overall, our results confirm that non-invasive sensors can be used for valence classification, as already observed [5], [6]. Specifically, we use the minimum set of sensors—GSR, BVP, and HR measured using the Empatica E4 wristband—that can be used in an experimental protocol for detecting emotions during daily activities of software developers at the workplace. Using machine learning, we are able to distinguish between positive and non-positive valence. However, differences of physiology can significantly impact the performance, thus confirming the need for individual, dedicated training of emotion classification models. Of course, better performance could also be achieved by relying on other high-definition sensors. It is the case, for example, of Nogueira et al. [61] achieving up to .91 of accuracy for valence using facial EMG electrodes. Analogously, Vrzakova et al. [30] recently reported achieveing F1 = .79 for valence recognition during code review using the Shimmer GSR+[9] mounting GSR sensors on finger strap-rings. However, these studies rely on either invasive sensors, as in the case of facial electrodes used by Nogueira and colleagues, or sensors that might be perceived as less comfortable to wear, as the Shimmer GSR ring sensors that could impair typing and other developers' movements.

> *Takeaway message for RQ4* - Biometrics can be used as predictor for emotions at workplace. The observed variability of performance between participants suggests that emotion recognition might be enhanced by training emotion classification models on an individual basis.

9. www.shimmersensing.com

TABLE 5
Best valence classifiers performance. Improvement over the baseline reported in parenthesis. For comparison, we also report the classifier performance for the positive vs. negative valence classification as observed in our previous lab study [6]

	Hold-out setting Train: 90% + 10-fold cross-validation Test: 10% (10 times)						Leave-one-subject out setting Train: all-1 subject + LOO cross validation Test: 1 held-out subject (19 subjects)					
Time Window	**Alg.**	**Prec**	**Rec**	**F1**	**Accuracy**	**stdev**	**Alg.**	**Prec**	**Rec**	**F1**	**Accuracy**	**stdev**
Best run												
10 seconds	rf	.67 (+.38)	.67 (+.17)	.67 (+.30)	.68 (+.18)	–	knn	.64 (+.35)	.64 (+.14)	.59 (+.22)	.59 (+.09)	–
3 minutes	rf	.60 (+.31)	.60 (+.10)	.60 (+.23)	.61 (+.11)	–	rf	.75 (+.46)	.86 (+.36)	.75 (+.38)	.79 (+.29)	–
Average over the runs of the best setting												
10 seconds	rf	.56 (+.27)	.56 (+.06)	.55 (+.18)	.58 (+.08)	.05	knn	.48 (+.19)	.48 (-.02)	.43 (+.06)	.46 (-.04)	.08
3 minutes	rf	.54 (+.25)	.54 (+04)	.54 (+.17)	.56 (+.06)	.05	rf	.51 (+.22)	.53 (+.03)	.46 (+.09)	.50 (–)	.14
Baseline		*.29*	*.50*	*.37*	*.50*			*.29*	*.50*	*.37*	*.50*	
Classifier performance in the lab setting [6]												
10 seconds (average)	knn	.70	.59	.59	.71	.07	rf	.45	.61	.50	.68	.27

TABLE 6
Performance by class for the best and average of train-test rounds in the Hold-out and LOSO setting, respectively.

Hold-out setting					*LOSO setting*				
Class	**Prec**	**Rec**	**F1**	**Acc**	**Class**	**Prec**	**Rec**	**F1**	**Acc**
Best run									
Positive	.71	.75	.73	.68	Positive	1.00	.72	.84	.79
Non-positive	.62	.58	.60	.68	Non-positive	.50	1.00	.67	.79
Average of runs for best setting									
Positive	63	67	65	58	Positive	60	65	57	50
Non-positive	48	45	46	58	Non-positive	45	40	37	50

7 Discussion

7.1 Implications

Emotion as a proxy for productivity and job satisfaction. Our correlation study provides evidence that a relationship exists between emotional valence and self-perceived productivity (see Section 4), thus confirming previous findings in literature [3], [5], [6]. The results of our correlation analysis are corroborated by the findings of our coding study, leading to the definition of a taxonomy of emotion triggers at the workplace (see Section 5). In fact, most of the emotion triggers are associated to productivity, with the feeling of being *in-flow* and the ability to complete the daily tasks among the top causes for positive emotions. Conversely, being stuck is associated to negative feelings. Furthermore, we found that the perception of effective use of time (e.g., constructive meetings) and the ability to complete their tasks are among the most frequently reported triggers for (un)happiness. Overall, our results suggest that emotions might act as a proxy for productivity, towards bridging automated measures and self-report for productivity assessment [62]. For example, positive emotions can indicate that a developer is *in flow* and should not be disturbed or that he/she was able to successfully accomplish the daily tasks. Similarly, the identification of negative emotions can indicate a developer requiring support because is stuck or has problems in comprehending code. In this view, our taxonomy of emotion triggers might guide and inspire to the definition of *ad hoc* interventions to enhance their productivity and thus supporting their well-being.

By specifically focusing on triggers for developers' emotions at work, we are able to complement previous results of two large-scale survey studies at Microsoft investigating the developers' satisfaction and well-being. In the first study, Storey et al. [63] develop a theory revolving around the bidirectional relationship between job satisfaction and the perceived productivity. They identify the social and technical factors, challenges, and contextual aspects, all playing a role in this relationship. In the second study, Meyer et al. [7] report that the perceiving work as important and valuable is a key to developers' satisfaction. The authors propose a conceptual framework for good working days based on three main factors including value creation, efficient use of time and considerations of affective states. The results of our study fits in the frame of such previous findings and complement them by providing specific insights on the role played by emotions in the developers' well-being and on how emotions correlate with perceived productivity.

Collaboration and organization of work. In our taxonomy, we observe how work management and social factors play a role in triggering emotions at work. Developers mostly reported being happy when engaging in fruitful interactions with peers, e.g., during collaborative problem solving, while helping peers, or when participating in constructive meetings. This is in line with recent findings [63] reporting collaborative team culture among the top factors for job satisfaction. Surprisingly, we also observed a minority of developers being annoyed by helping peers, which can be a consequence of introverted personalities who find difficult to help others. In this view, personality assessment could be used as a tool informing effective team composition, as envisioned by previous research on personality in software development teams [64].

The role of fatigue. The results of our time-wise correlation analysis (see Section 4) suggest that fatigue may be a mediating factor in the relationship between emotions and self-assessed productivity. In addition, developers reported fatigue among the most frequent causes for negative emotions. Conversely, breaks are used to restore positive mood (see Section 5). Fatigue has been already reported as a cause for negative emotions by software developers [17] and should be taken into due consideration as it might impair cognitive abilities and performance [41], thus lowering code quality. We advocate in favor of follow-up research, also leveraging biometrics, towards early-detection of symptoms of fatigue and stress, in order to enable just-in-time implementation of strategies for restoring positive mood and regain focus during the workday.

Sensor-based emotion recognition at workplace. This study represents the first attempt to assess the performance of

a supervised machine learning classifier for developers' emotion at work, by leveraging developers' biofeedback collected during the entire working day and across the different activities, beyond programming. Previous studies either focused on specific tasks, such as code review [30], or attempt to classify emotions during development tasks in a lab setting [5], [6]. While promising, our classifier performance is still not robust enough for practical usage. Further data collection is required to ensure the reliability of our approach, also leveraging individual training of emotion models for each participant. Yet, we believe our findings pave the way to follow-up studies to empower emotion awareness in software development by using sensor technology. We believe the machine learning protocol we designed and used for collection of biofeedback and self-reported emotions and productivity can stimulate future *in-vivo* research, towards reaching the maturity required for the deployment and adoption of sensor-based emotion classifiers. In our vision, sensor-based emotion recognition could be integrated in context-aware approaches that leverages multiple sources of information to prompt just-in-time suggestions for developers. For example, Kaur et al. [65] propose an approach for modeling opportune moments for transitions and breaks based on affect- and task-related data. They build models to predict whether a worker should continue their task, move to a new task, or take a break. Züger et al. [53] also report computer interaction data are effective in predicting interruptibility while at work. Beyond emotion recognition, observing developers at the workplace also opens opportunities to build classifiers for identification of bad days (i.e., when mostly negative emotions are identified) or negative working conditions of developers (i.e., when negative affect is observed over a long period of time).

The importance of emotion awareness at team level. At team level, companies are recently implementing strategies to support emotion awareness [14], [66]. For example, during agile retrospective meetings, developers could self-report their emotions on a whiteboard and leverage them as starting point for the discussion. By doing so, the team can better identify what are the activities and events relating to positive and negative emotions. Recently, Andriyani et al. [46] conducted a case study by interviewing software practitioners from different agile teams about retrospective meetings. They found that beyond discussing problems and strategies adopted to address them, developers also discuss their positive and negative feelings about events and difficulties encountered. Along the same line, we believe our taxonomy could guide self-report towards the inclusion of emotional feedback in agile meetings. In the long run, we envision the adoption of biometrics-based emotion recognition to support and enhance retrospective meetings by including emotional information collected day-to-day.

Instrumentation. Among the most frequent cause of negative emotion, our participants reported being annoyed by poor tooling, such as non-working or not adequate hardware or software infrastructure, or buggy source code they have to modify. At the time of writing, we conjecture this problem might be further amplified by the working condition of many software developers that are forced to work from home due to the Covid-19 pandemic [67], [68].

7.2 Threats to validity

Threats to *external validity* relate to the generalizability of the results, which recently emerged as an open challenge of software engineering research [69]. We managed to involve a fairly diverse pool of five companies ranging from startup to large international companies. However, we are aware that we need to be cautious in claiming generalizability of our findings as our sample might not be representative of the software industry as a whole. Furthermore, the pool of participants is imbalanced with respect to gender distribution. Further replications should involve more women and non-binary participants to account for potential differences in the emotional reaction.

As for the biometrics study, the *validity of our conclusions* ground on the robustness of the generalized linear model and machine learning models. We mitigated such threat by running and comparing several algorithms, applying hyper-parameters tuning, and evaluating the approach in two different settings —i.e., Hold-out and LOSO. Nevertheless, the results we report here are limited by the sample size and the imbalance of data used for training.

Our study suffers from threats to *construct validity*—i.e., the reliability of our measures in capturing emotions and progress, mainly due to the *in situ* nature of the data collection protocol. In this study, we employed low-cost, lightweight sensors that are comfortable to wear at work. This might have lowered the quality of data collected by the sensors with respect to those collected in a controlled setting as in previous lab studies. To mitigate this threat, we performed a careful quality assessment of the collected data to compensate the impossibility to supervise the experiment in person at the company. Specifically, one of the authors performed a daily check of the correctness of the data shared by participants, by synchronizing their reports via Dropbox. Furthermore, we performed data quality assurance and did not consider participants who misinterpreted the concept of valence, arousal, and dominance—e.g., who reported always the same score also during the experiment.

Threats to *internal validity* concern confounding factors that can influence the results. Using the self-report pop-up involved interrupting developers during the task, which can have interfered with their work, thus eliciting negative emotions. We mitigate this threat by interrupting the developers every hour, in line with the suggestion received by the developers participating in our pilot study. Furthermore, the developers could always skip the self-report if they did not want their activity to be interrupted. Finally, as regards the impact of time as it emerged from the analysis of correlation between self-reported emotions and productivity, we are aware that individual differences in participants' circadian rhythms could have played a role. As such, we cannot exclude that we would have observed differences between people due to their circadian rhythms and different levels of alertness in the morning vs. afternoon. Unfortunately we could not control for this factor, which we believe would be worth investigating in follow-up studies.

8 Conclusion

Emotions are known to play an important role in problem solving as well as to influence job performance. In this paper

we report the findings of a longitudinal study of emotions experienced by software developers at their workplace. Twenty-one software developers from five companies have been observed during two or three weeks, depending on the duration of the agile iteration. Emotion data have been collected by means of self-reporting and biometric sensing. Participants reported their emotions in terms of valence, arousal, and dominance, as well as their perceived productivity during the workday. Developers mostly reported positive valence, neutral arousal, and high dominance, indicating they mainly experience pleasant emotions and feel in control of the situation while at work. The analysis of the correlation between emotions and productivity demonstrates a significant association between positive valence and self-assessed productivity, which becomes stronger in the afternoon, probably due to the effect of fatigue.

Other than assessing their emotions, developers were required to explain the causes for the emotional scores they provided. We coded these open answers and obtained a taxonomy of the emotion triggers at the workplace. The results of the coding study demonstrate how positive emotions are mostly caused by developers' feeling productive or being able to successfully collaborate with peers. Problems in code comprehension, poor tooling, fatigue, as well as personal issues not related to work, are reported as triggers for negative feelings.

Results of both our qualitative and quantitative analysis suggest that emotions might act as a proxy for productivity, in line with findings from previous studies on developers' emotions. We believe taxonomy of emotion triggers can drive enhancement interventions of developers' productivity by guiding and informing consideration of affect and its causes into daily practice, e.g. through integration of emotional feedback in retrospective meetings or in planning activities to improve organization of work.

In the long run, we envision the use of biometrics for emotion recognition to support and enhance emotion awareness both at an individual and the team level. Towards this long-term goal, we experimented with a minimum set of non-invasive biometric sensors can be used as predictor for emotions. Results are encouraging, yet not applicable in everyday practice. Further studies are required to collect additional data and improve classification performance, also leveraging training on an individual basis. Both the approaches adopted in the present study, i.e. the experience sampling and biometrics, can be used to achieve a shared goal, that is to support developers' emotional awareness, thus enhancing their well-being and productivity. In the next future, we envision studies based on biometrics that can lead to a refined version of the taxonomy of emotion triggers we present in this paper, e.g. by asking people to self-report triggers for positive and negative emotions according to biometrics. Conversely, the taxonomy can be used in combination with biometrics, by leveraging triggers as one of the predictors in a classifier.

REFERENCES

[1] T. M. Amabile, S. G. Barsade, J. S. Mueller, and B. M. Staw, "Affect and creativity at work," *Administrative Science Quarterly*, vol. 50, no. 3, pp. 367–403, 2005.

[2] A. Murgia, P. Tourani, B. Adams, and M. Ortu, "Do developers feel emotions? an exploratory analysis of emotions in software artifacts," in *MSR*, 2014, pp. 262–271.

[3] D. Graziotin, X. Wang, and P. Abrahamsson, "Are happy developers more productive? - the correlation of affective states of software developers and their self-assessed productivity," in *PROFES*, 2013, pp. 50–64.

[4] D. Graziotin, F. Fagerholm, X. Wang, and P. Abrahamsson, "What happens when software developers are (un)happy," *Journal of Systems and Software*, vol. 140, pp. 32–47, 2018.

[5] S. C. Müller and T. Fritz, "Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress," in *ICSE*, 2015, pp. 688–699.

[6] D. Girardi, N. Novielli, D. Fucci, and F. Lanubile, "Recognizing developers' emotions while programming," in *ICSE*, 2020, p. 666–677.

[7] A. Meyer, E. T. Barr, C. Bird, and T. Zimmermann, "Today was a good day: The daily life of software developers," *IEEE Transactions on Software Engineering*, pp. 1–1, 2019.

[8] D. Graziotin, X. Wang, and P. Abrahamsson, "Do feelings matter? on the correlation of affects and the self-assessed productivity in software engineering," *J. of Software: Evol. and Proc.*, vol. 27, no. 7, pp. 467–487, 2015.

[9] S. C. Müller and T. Fritz, "Using (bio)metrics to predict code quality online," in *Proc.of the 38th Int'l Conf. on Software Engineering, ICSE 2016, Austin, TX, USA, May 14-22, 2016*, 2016, pp. 452–463.

[10] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[11] C. Foster and J. Sayers, "Exploring physiotherapists emotion work in private practice," *New Zealand Journal of Physiotherapy*, vol. 40, pp. 17–23, 01 2012.

[12] M. Mäntylä, K. Petersen, T. O. A. Lehtinen, and C. Lassenius, "Time pressure: A controlled experiment of test case development and requirements review," in *ICSE*, 2014, pp. 83–94.

[13] D. Graziotin, X. Wang, and P. Abrahamsson, "Software developers, moods, emotions, and performance," *IEEE Software*, vol. 31, no. 4, pp. 24–27, 2014.

[14] E. Marcos, R. Hens, T. Puebla, and J. M. Vara, "Applying emotional team coaching to software development," *IEEE Software*, pp. 1–8, 2020.

[15] J. Russell, "Culture and the categorization of emotions," *Psychological Bulletin*, vol. 110 (3), pp. 426–450, 1991.

[16] M. Mäntylä, B. Adams, G. Destefanis, D. Graziotin, and M. Ortu, "Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity?" in *MSR*, 2016, pp. 247–258.

[17] D. Ford and C. Parnin, "Exploring causes of frustration for software developers," in *CHASE*, 2015, pp. 115–116.

[18] D. Graziotin, F. Fagerholm, X. Wang, and P. Abrahamsson, "On the unhappiness of software developers," in *EASE '17*. New York, NY, USA: ACM, 2017, p. 324–333.

[19] D. Graziotin, X. Wang, and P. Abrahamsson, "Happy software developers solve problems better: psychological measurements in empirical software engineering," *PeerJ*, 2014.

[20] M. R. Wrobel, "Emotions in the software development process," in *HSI*, 2013, pp. 518–523.

[21] A. E. Kramer, *Physiological Metrics of Mental Workload: A Review of Recent Progress*, D. T. I. Center, Ed., 06 1990.

[22] B. Reuderink, C. Mühl, and M. Poel, "Valence, arousal and dominance in the eeg during game play," *Int. J. of Autonomous and Adaptive Communic. Syst.*, vol. 6, no. 1, pp. 45–62, 2013.

[23] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Aff. Comp.*, vol. 7, no. 1, pp. 17–28, 2016.

[24] M. M. Bradley and P. J. Lang, "Measuring emotion: Behavior, feeling, and physiology," in *Cognitive Neuroscience of Emotion*, ser. Series in Affective Science, R. D. Lane and L. Nadel, Eds. Oxford University Press, 2000, ch. 11, pp. 242–276.

[25] W. Burleson and R. W. Picard, "Affective agents: Sustaining motivation to learn through failure and state of "stuck"," in *Social and Emotional Intelligence in Learning Environments Workshop*, 8 2004.

[26] F. Canento, A. Fred, H. Silva, H. Gamboa, and A. Lourenço, "Multimodal biosignal sensor data handling for emotion recognition," in *SENSORS*. IEEE, 2011, pp. 647–650.

[27] E. Carniglia, M. Caputi, V. Manfredi, D. Zambarbieri, and E. Pessa, "The influence of emotional picture thematic content on exploratory eye movements," *J. Eye Mov. Res.*, vol. 4, pp. 1–9, 2012.

[28] K. Muldner, R. Christopherson, R. Atkinson, and W. Burleson, "Investigating the utility of eye-tracking information on affect and reasoning for user modeling," in *UMAP 2009*, 2009, pp. 138–149.

[29] S. Koelstra, C. Mühl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[30] H. Vrzakova, A. Begel, L. Mehtätalo, and R. Bednarik, "Affect recognition in code review: An in-situ biometric study of reviewer's affect," *J. Syst. Softw.*, vol. 159, 2020.

[31] D. Girardi, F. Lanubile, N. Novielli, L. Quaranta, and A. Serebrenik, "Towards recognizing the emotions of developers using biometrics: The design of a field study," in *SEmotion Workshop*, 2019, pp. 13–16.

[32] J. Highsmith and M. Fowler, "The agile manifesto," *Software Development Magazine*, vol. 9, no. 8, pp. 29–30, 2001.

[33] R. Larson and M. Csikszentmihalyi, *The Experience Sampling*. Springer Netherlands, 2014, pp. 21–34.

[34] D. Graziotin, X. Wang, and P. Abrahamsson, "Understanding the affect of developers: Theoretical background and guidelines for psychoempirical software engineering," in *SSE Workshop*. New York, NY, USA: ACM, 2015, p. 25–32.

[35] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. of Behav. Therapy & Experim. Psych.*, vol. 25, no. 1, pp. 49–59, 1994.

[36] S. C. Müller and T. Fritz, "Using (bio)metrics to predict code quality online," in *ICSE 2016*. ACM, 2016, pp. 452–463.

[37] A. N. Meyer, G. C. Murphy, T. Zimmermann, and T. Fritz, "Design recommendations for self-monitoring in the workplace: Studies in software development," *Proc. ACM HCI*, vol. 1, Dec. 2017.

[38] M. Kuutila, M. Mäntylä, U. Farooq, and M. Claes, "Time pressure in software engineering: A systematic review," *Inf. Softw. Tech.*, vol. 121, 2020.

[39] R. Gueorguieva and J. H. Krystal, "Move over anova: progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry." *Archives of general psychiatry*, vol. 61, no. 3, pp. 310–317, 2004.

[40] A. Stone, J. Schwartz, D. Schkade, N. Schwarz, A. Krueger, and D. Kahneman, "A population approach to the study of emotion: diurnal rhythms of a working day examined with the day reconstruction method." *Emotion*, vol. 6 1, pp. 139–49, 2006.

[41] S. Sarkar and C. Parnin, "Characterizing and predicting mental fatigue during programming tasks," in *SEmotion*, 2017, pp. 32–37.

[42] C. Grillon, D. Quispe-Escudero, A. Mathur, and M. Ernst, "Mental fatigue impairs emotion regulation," *Emotion*, vol. 15, no. 3, pp. 383–389, 2015.

[43] T. Zimmermann, "Card-sorting: From text to themes," in *Perspectives on Data Science for Software Engineering*, T. Menzies, L. Williams, and T. Zimmermann, Eds. Morgan Kaufmann, 2016, pp. 137–141.

[44] P. Martin and B. Turner, "Grounded theory and organizational research," *J. Appl. Behav. Sci.*, vol. 22, no. 2, pp. 141–157, 1986.

[45] P. Jordan and N. Ashkanasy, *Emotional Intelligence, Emotional Self-Awareness, and Team Effectiveness*. Lawrence Erlbaum Associates Publishers, 2006, pp. 145–163.

[46] Y. Andriyani, R. Hoda, and R. Amor, "Reflection in agile retrospectives," in *Agile Processes in Software Engineering and Extreme Programming*, H. Baumeister, H. Lichter, and M. Riebisch, Eds. Cham: Springer International Publishing, 2017, pp. 3–19.

[47] A. Fountaine and B. Sharif, "Emotional awareness in software development: Theory and measurement," in *2017 IEEE/ACM 2nd Int'l Workshop on Emotion Awareness in Software Engineering (SEmotion)*, 2017, pp. 28–31.

[48] C. Repetto, A. Gaggioli, F. Pallavicini, P. Cipresso, S. Raspelli, and G. Riva, "Virtual reality and mobile phones in the treatment of generalized anxiety disorders: A phase-2 clinical trial," *Personal and Ubiquitous Computing*, vol. 17, pp. 253–260, 02 2013.

[49] H. Stegge and M. Terwogt, *Awareness and Regulation of Emotion in Typical and Atypical Development*. The Guilford Press, 2006, pp. 269–286.

[50] J. Lambie and A. Marcel, "Consciousness and the varieties of emotion experience: A theoretical framework," *Psychological Review*, vol. 109, no. 2, pp. 219–259, 2002.

[51] N. V. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Intell. Res. (JAIR)*, vol. 16, pp. 321–357, 06 2002.

[52] D. Girardi, F. Lanubile, and N. Novielli, "Emotion detection using noninvasive low cost sensors," in *ACII 2017*, 2017, pp. 125–130.

[53] M. Züger, S. C. Müller, A. N. Meyer, and T. Fritz, "Sensing interruptibility in the office: A field study on the use of biometric and computer interaction sensors," in *Proc.of the 2018 CHI Conf. on Human Factors in Computing Systems, (CHI 2018)*, 2018, p. 591.

[54] J. Healey and R. Picard, "Digital processing of affective signals," in *ICASSP '98*, vol. 6, 1998, pp. 3749–3752 vol.6.

[55] N. Jaques, S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano, and R. Picard, "Predicting students' happiness from physiology, phone, mobility, and behavioral data," in *ACII 2015*, vol. 2015, 09 2015, pp. 222–228.

[56] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing," *IEEE Trans. on Biom. Eng.*, vol. 63, no. 4, pp. 797–804, 2016.

[57] D. Fucci, D. Girardi, N. Novielli, L. Quaranta, and F. Lanubile, "A replication study on code comprehension and expertise using lightweight biometric sensors," in *ICPC 2019*, 2019, pp. 311–322.

[58] M. Kuhn, "The caret package," http://topepo.github.io/caret/index.html, 2009.

[59] C. Tantithamthavorn, S. McIntosh, A. E. Hassan, and K. Matsumoto, "The impact of automated parameter optimization on defect prediction models," *IEEE Trans. on Softw. Eng.*, vol. 45, no. 7, pp. 683–711, 2019.

[60] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *CoRR*, vol. abs/1811.12808, 2018.

[61] P. A. Nogueira, R. A. Rodrigues, E. C. Oliveira, and L. E. Nacke, "A hybrid approach at emotional state detection: Merging theoretical models of emotion with data-driven statistical classifiers," in *IAT 2013*. IEEE, 2013, pp. 253–260.

[62] M. Beller, V. Orgovan, S. Buja, and T. Zimmermann, "Mind the gap: On the relationship between automatically measured and self-reported productivity," 2020.

[63] M. Storey, T. Zimmermann, C. Bird, J. Czerwonka, B. Murphy, and E. Kalliamvakou, "Towards a theory of software developer job satisfaction and perceived productivity," *IEEE Trans. Softw. Eng.*, pp. 1–1, 2019.

[64] F. Calefato, G. Iaffaldano, F. Lanubile, and B. Vasilescu, "On developers' personality in large-scale distributed projects: The case of the apache ecosystem," in *ICGSE '18*. ACM, 2018, p. 92–101.

[65] H. Kaur, A. C. Williams, D. McDuff, M. Czerwinski, J. Teevan, and S. T. Iqbal, "Optimizing for happiness and productivity: Modeling opportune moments for transitions and breaks at work," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–15.

[66] N. Novielli and A. Serebrenik, "Sentiment and emotion in software engineering," *IEEE Software*, vol. 36, no. 5, pp. 6–23, Sep. 2019.

[67] P. Ralph, S. Baltes, G. Adisaputri, R. Torkar, V. Kovalenko, M. Kalinowski, N. Novielli, S. Yoo, X. Devroey, X. Tan, and et al., "Pandemic programming," *Empirical Software Engineering*, vol. 25, no. 6, 2020.

[68] C. Miller, P. Rodeghero, M.-A. Storey, D. Ford Robinson, and T. Zimmermann, "'how was your weekend?' software development teams working from home during covid-19," in *ICSE 2021*, 2021. [Online]. Available: https://arxiv.org/abs/2101.05877

[69] S. Baltes and P. Ralph, "Sampling in software engineering research: A critical review and guidelines," 2021. [Online]. Available: https://arxiv.org/abs/2002.07764

ACKNOWLEDGMENTS

This work was partially supported by the Italian Ministry of University and Research under grant PRIN 2017 "EMPATHY: EMpowering People in deAling with internet of THings ecosYstems" (project H94I19000280001). We would like to thank Apuliasoft for participating in the pilot study, and the companies and the developers who participated in the field study. We thank Giuseppe Antonio Nanna and Arcangelo Saracino for their support in developing the pop-up application.

Daniela Girardi received a PhD in Computer Science in 2021 from the University of Bari. Italy. During her PhD, she focused on investigating the use of biometrics for automatic recognition of developers' emotions.

Filippo Lanubile is a Full Professor of computer science at the University of Bari, Italy, where he leads the Collaborative Development Research Group. He is also the CEO and co-founder of the academic spin-off company PeoplewareAI. His research interests include: human factors in software engineering, collaborative software development, and software engineering for AI/ML systems. He has won two awards from IBM and one from Microsoft Research. He is the Chair of the IEEE Software Advisory Board.

Nicole Novielli is an Assistant Professor at the University of Bari, Italy. Her research interests lie at the intersection of software engineering and affective computing with a specific focus on mining emotions and opinions from developers communication traces and sensor-based recognition of developers cognitive and affective states.

Alexander Serebrenik is a Full Professor of Social Software Engineering at Eindhoven University of Technology, The Netherlands. His research goal is to facilitate evolution of software by taking into account social aspects of software development. He has co-authored a book Evolving Software Systems (Springer Verlag, 2014) and circa 200 scientific papers and articles. He has won several distinguished paper and distinguished review awards.