How Well Can a Wristband Provide Information About a Person's Emotional State?

Arianna Di Serio[1][0009-0006-9925-8598], Giulio Mori[1][0000-0001-9514-5066], and Fabio Paternò[1][0000-0001-8355-6909]

[1] CNR–ISTI, HIIS Laboratory, Pisa 56127, Italy
{arianna.diserio, giulio.mori, fabio.paterno}@isti.cnr.it

Abstract. The integration of technologies for collecting physiological data via wearable devices has introduced new ways to enhance User Experience (UX) evaluation also in terms of understanding users' emotions. By exploring emotional responses, UX professionals can create more adaptive, intuitive, satisfying, and emotionally aligned experiences that better meet users' needs. With this perspective, our study aimed to investigate the benefits and limitations of using a low-cost wristband to predict emotions. For this purpose, we developed an Android application for managing physiological data collected through an Empatica E4 wristband and employed machine-learning techniques for emotion classification of such data. We then conducted a user study with 30 participants who viewed seven videos stimulating specific emotions, and we compared the device and application results with the subjects' self-reported emotional states. The results provide useful information to better understand the possibilities and limitations when using a wearable device for emotional monitoring.

Keywords: Emotion Recognition, Wearable Devices, Usability Evaluation.

1 Introduction

Nowadays, the integration of technologies for collecting physiological data via wearable devices has introduced new ways to enhance User Experience (UX) evaluation, especially in terms of understanding users' emotions. Emotion recognition through physiological data can significantly improve the work of designers and developers with real-time insights into how users feel during interactions with a product or service. By understanding emotional responses - such as frustration, excitement, or engagement - UX professionals can create more intuitive, satisfying, and emotionally aligned experiences that better meet users' needs and adapt to them [12].

Emotion recognition is an ever-growing field of research that leverages various technologies and methods such as facial expression recognition through visual sensors, speech recognition through audio sensors, and radar technologies to analyse the chest micro-motion caused by breathing and heartbeats [4, 17, 18, 32]. Despite the advancements in these technologies, the field still grapples with defining and measuring emotions accurately [3, 9-11, 16, 19]. Indeed, studies often tend to infer emotions by observing the effects without clearly understanding the causes that stimulated them.

Emotions are complex because they depend on individual preferences, attitudes, moods, dispositions and interpersonal stances; “there is no single standard gold-method for their measurement” [27]. The issue of emotion detection is not a trivial task, and it becomes less trivial if this has to be done in real time. While emotion recognition in controlled environments has been extensively studied, translating these methods to real-world scenarios presents significant challenges, including the need for adaptable self-assessment methods and advanced machine learning techniques to address data imbalances and variability [25].

Bio signals emerge as an interesting field of research, as they present intrinsic and methodological advantages [24]. Firstly, they cannot be easily masked, and they are measurable with non-invasive sensors, making them suitable for multiple applications. Secondly, they provide more detailed and continuously updated insights into physiological changes compared to traditional emotion analysis methods, that rely solely on questionnaires, interviews, or expert opinions (which are more influenced by individual subjectivity and do not allow real-time measurement on their own). Unfortunately, there are still challenges in emotion recognition based on physiological signals [30]. Typically methods and tools in this area aim to monitor changes in a person’s internal physiological signals - Electroencephalogram (EEG), Electrocardiogram (ECG), Blood Volume Pulse (BVP), Electrodermal Activity (EDA), Infrared Thermography (IRT), etc. - or extrapolate changes from external aspects (such as facial expressions through software, self-evaluation methodologies, topographical self-report analysis [10], user’s utterance through words, stories, feelings, etc.). For instance, Speech Emotion Recognition (SER) analyses vocal characteristics to infer emotional states, though it can be affected by background noise and cultural differences. Facial Expression Recognition (FER) instead uses visual sensors and machine learning to interpret facial movements, but hidden expressions and variability in facial features are serious obstacles to emotion detection. Our experience with software able to detect facial expressions based on Ekman's six emotions classification (anger, disgust, fear, happiness, sadness, surprise) [8] emphasised several limitations: often, users had to exaggerate their facial expressions in order to enable emotion detection by these tools.

Wearable devices such as smartwatches, fitness trackers and EEG headsets can be used to collect continuous bio-signals data, also when users freely move about. They are capable of detecting, analysing and transmitting information related to, for example, bodily signals and environmental data and they can be utilised in various fields such as healthcare, security, entertainment and gaming. Their versatility, adaptability, and non-intrusiveness make them ideal for daily wear [20, 21], and suitable for mobile users. Thus, the integration of wearable devices for physiological data collection can change user experience evaluation by offering continuous, objective measurements. In dynamic and mobile environments, where emotions significantly shape satisfaction and engagement, wearables offer deeper, real-time insights into users' physical and emotional responses, surpassing traditional methods like surveys or interviews.

To provide a better understanding of the potential of wearable technologies, we focused on recognizing emotions from physiological signals such as skin electrodermal activity, blood volume pulse and body temperature. We have used the Empatica E4 bracelet [36], which allows developers to access sensor data for their own applications.

Specifically, we developed an Android application, capable of connecting with an Empatica E4 bracelet, measuring physiological signals and storing them in a database. Besides, using machine learning techniques, we implemented a classifier in Python labelling the gathered data through the Empatica E4 bracelet with some specific emotions (*amusement, stress, neutrality* and *meditation*). To achieve this, we used an online available training dataset (WESAD [28]), which collected physiological data using the Empatica E4 device.

We also developed a web platform to access the complete data of the measurement sessions and graphs monitoring trends of both physiological signals and the inferred emotions over time. Finally, we executed a user test to evaluate the effectiveness of the implemented system in recognizing the emotions perceived by the participants. We selected seven scenes drawn from a previous research [14] considered capable of evoking a specific emotion (amusement, anger, contentment, disgust, fear, neutrality, sadness and surprise). 30 users wore the bracelet (Empatica E4) while viewing the selected seven scenes. The users filled out a questionnaire, reporting their subjective emotional experiences for each scene, and their levels of pleasure and arousal.

In this way, we wanted to investigate the benefits and limitations of using the Empatica E4 wristband applied to the emotion recognition context, with the purpose of addressing the following research questions:

- RQ1: How useful is the data gathered through a wristband device for predicting emotions?
- RQ2: How well do the self-reported responses correlate with the predicted emotions based on physiological data and machine learning techniques?

The paper is organised as follows. After the introduction, we discuss related work concerning emotion recognition that is relevant to our contribution. Next, we present an overview of the approach adopted and the implemented system. We then describe the user test, its results and the associated data analysis. Section 5 discusses what we learned and the limitations of the study. Lastly, we present some concluding remarks and indications for future work.

2 Related Work

Several studies have been conducted on emotion recognition using physiological signals: they offer the possibilities of different methodologies and approaches used in the field of affective computing to monitor and classify stress or other emotions through the analysis of these signals. In this section we discuss some contributions especially considering the same type of device that we have used in our work.

For instance, Di Lascio et al. [7] leveraged EDA sensors to monitor students' emotional engagement during lectures. By analysing physiological signals along-side self-reported data from 24 students across 41 lectures, they demonstrated that features representing general arousal, physiological synchrony with teachers, and momentary engagement could effectively discriminate between engaged and non-engaged students, achieving a recall of 81% with a Support Vector Machine classifier. Ketonen et al. [15]

examined the concurrent associations between self-reported emotions, such as excitement, calmness, anxiety and boredom, and students' biometric data such as heart rate (HR) and heart rate variability (HRV). This study involved 136 high school students who wore biometric sensors for three consecutive days, capturing data on their physical activity via the metabolic equivalent of task (MET) values, HR, and HRV, while also responding to experience sampling method (ESM) questionnaires. Their results showed that excitement was associated with higher HR and lower HRV, indicating increased sympathetic activation. Boredom, on the other hand, correlated with lower HR but showed no significant relationship with HRV. The study concluded that physiological markers, particularly HR, provide valuable insight into the emotional states experienced by students in naturalistic settings, while also emphasizing the importance of combining self-reported data with physiological data for more nuanced emotional assessments.

Going into more detail and focusing on studies that utilised the Empatica E4, a contribution from Wang et al. [31], leveraged the Empatica E4 wristband to distinguish different activities (e.g., sitting, walking, running) using the accelerometer sensor and extract fine-grained features characterising different emotions based on these activities using multi-mode sensory data.

In a study by Rahman et al. [23] only electrodermal activities (EDA) were analysed to recognize seven emotional categories while participants watched emotional videos. Twenty observers' EDA signals were processed to extract sixteen features per video (such as mean or variance). The analysis revealed that some emotions had significantly different electrodermal activity signals from one another, whereas others did not show such clear distinctions.

Engagement and emotion significantly impact a reader’s experience during reading tasks. A contribution [26] reported a study in which 18 university students read 14 documents and rated their engagement, valence, and arousal levels via questionnaire. Using a Tobii 4C eye-tracker and an Empatica E4 wristband, behavioural and physiological data were recorded. The research revealed some influences of engagement on the emotion felt, suggesting different affective responses to different texts.

Another pilot study by De Carolis et al. [6] explored the potential relationship between estimated stress levels based on physiological signals and the usability of an interactive system. Usability tests with 30 participants involved recording physiological signals with the Empatica E4 wristband. The Stress.io app analysed the data using a regression model trained on the AffectiveROAD dataset to predict stress levels. Participants evaluated two versions of a website interface: one poorly designed and one well-designed. During the task, participants' physiological data were collected (EDA, BVP and temperature), and afterwards, they completed the System Usability Scale (SUS) and Self-Assessment Manikin (SAM) questionnaires. The study found that greater variability in estimated stress levels inversely correlated with SUS scores, indicating lower perceived usability.

In summary, unlike other studies, we considered a broader set of physiological parameters, including EDA, BVP, and temperature, to capture a more comprehensive picture of emotional responses. Like some of these contributions, we also considered self-reported responses via questionnaires, such as the Self-Assessment Manikin. Rather than limiting our analysis to binary classifications like stress or non-stress, we explored a wider range of emotions (as described above in section 1).

3 Method

In our study, we adopted a comprehensive approach to address the research questions concerning the use of a wearable wristband device for emotion recognition. This approach combines conceptual models with tool implementations.

Conceptually, we drew on established multidimensional models of emotions, which provide a structured way to understand and categorise emotional states. Emotions can be distinguished based on factors such as [5]: a) pleasure (or valence), the pleasantness or unpleasantness of the emotion experienced; b) arousal (or activation), the intensity of the emotional state; c) dominance, the subject's sense of control over the emotion.

As a simplification, we considered Russell's Circumplex Model [33], which is one of the most influential frameworks for understanding emotional experiences. This model posits that emotions can be mapped onto a two-dimensional plane defined by the independent, bipolar dimensions of valence and arousal. Thus, each emotion can be interpreted as a linear combination of these two dimensions, or as varying degrees of both valence and arousal. For example, high arousal coupled with positive valence might correspond to excitement, while low arousal and negative valence could represent sadness [22]. We considered the dimensions of pleasure and arousal of this model in our user test, asking participants to express these concepts by describing their emotional experiences through a questionnaire.

On the tool side, we implemented a system based on different components (as illustrated in Figure 1), utilising the capabilities of the Empatica E4 device. This device includes: a photoplethysmography (PPG) sensor for measuring Blood Volume Pulse (BVP) and deriving heart rate variability at a sampling rate of 64Hz; a 3-axis accelerometer for detecting movement-based activity at 32Hz; an Electrodermal Activity (EDA) sensor to capture changes in skin's electrical properties at 4Hz; and an infrared thermopile sensor to monitor peripheral skin temperature at 4Hz. Our implementation comprises:

- an Android Studio app in Java, to connect with the Empatica E4 and store the physiological data in a Firestore database [37];
- a classifier in Python (a Multilayer Perceptron model), using the WESAD dataset [28] as the training dataset;
- a Flask server in Python to label the physiological signals of a measurement with the implemented classifier;
- a web platform to allow access to the complete data.

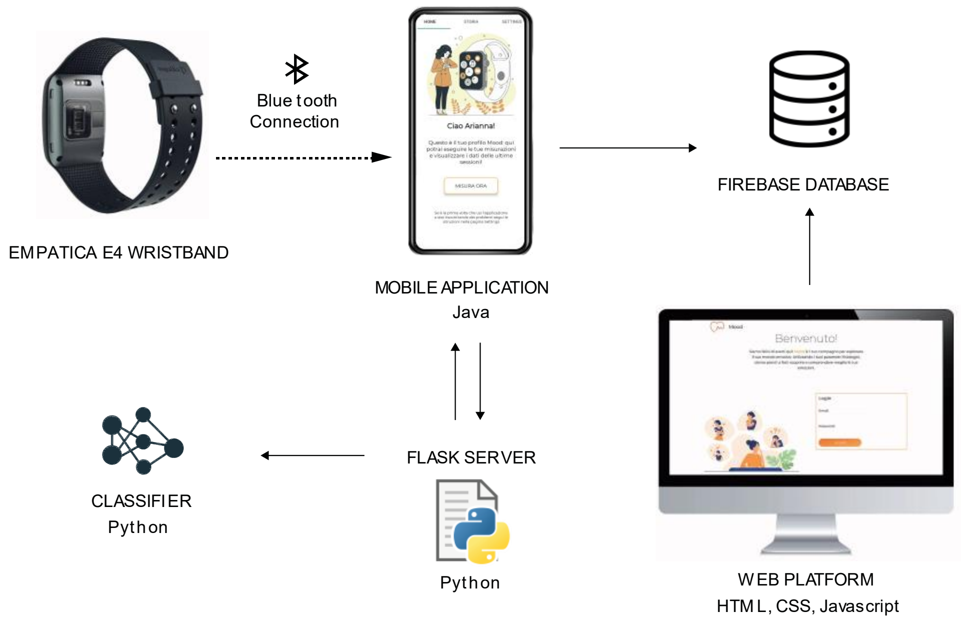

Fig. 1. The components of our system

3.1 The App, the Web Platform and the Database

The app was implemented in Android Studio, using the E4link framework (a sample project and a library provided by Empatica for Android). It allows users to register, log in and conduct a measurement session.

During a session, physiological signals are grouped by 10-second intervals and stored in a Firestore database, a service provided by Firebase. The choice of 10 seconds was based on established practices in research on sensors and the classification of emotional and cognitive states [13].

After a measurement session lasting at least 10 seconds, the data are sent to a Flask server to be analysed with the implemented classifier. The results are then displayed on the app, showing the percentage for each emotion class (*amusement, stress, neutrality* and *meditation*). For example, a session's output could be: *50% amusement, 25% neutrality, 20% meditation and 5% stress*.

To allow the full visualisation of all sessions, we also developed a web application. It shows: a table listing the user's measurements, with options to search by date or time range; detailed data for each measurement when clicked; line graphs showing trends for BVP, GSR, temperature, and acceleration during the selected session; a line graph displaying emotional state trends for the selected measurement; and two graphs illustrating emotion trends across all measurements (a bar graph of the frequency of each of the four emotions - amusement, stress, neutrality, and meditation - and a line graph of the predominant emotion for each day). Figure 2 shows some screen-dumps of the Android application.

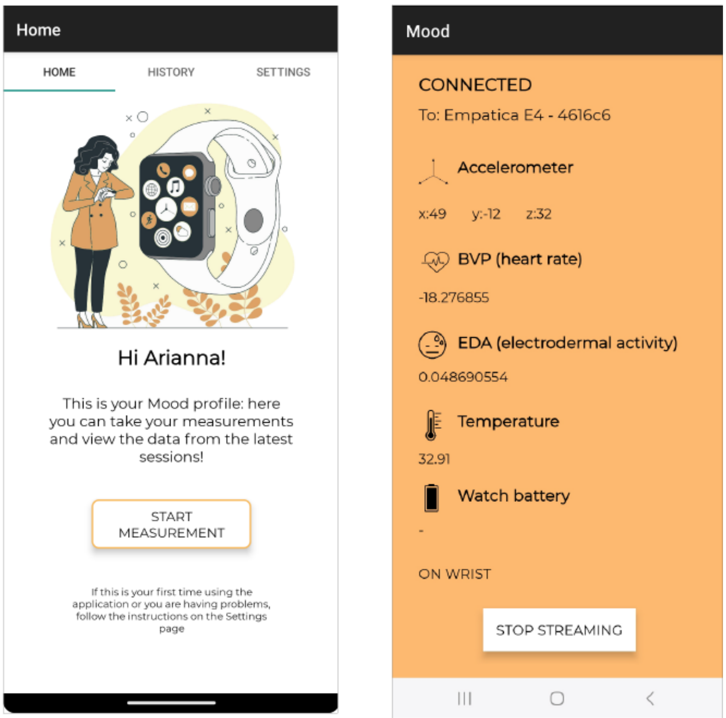

Fig. 2. The application (homepage on the left, measurement page on the right)

3.2 The Classifier and the Flask Server

To determine the emotion associated with physiological signals, we created a classifier with Python. As illustrated in Figure 3, we selected an existing dataset for training. After data preparation, we applied and compared the performance of multiple classification algorithms in correctly labelling the data. We selected the classifier based on performance metrics, trained the model, collected and classified user data, and then analysed the results obtained.

As highlighted by Ruta et al. [24], in the field of affective computing there is a scarcity of comprehensive datasets for the analysis of emotional states through the study of physiological signals. In our research we have reviewed several of them as potential training datasets for the classifier to be implemented: DEAP [35], DECAF [1], ASCERTAIN [34], LAUREATE [38] and WESAD [28].

DEAP focuses on a variety of physiological signals and frontal face videos to study users' arousal, valence, dominance and familiarity. DECAF interprets physiological responses to multimedia. ASCERTAIN uses a combination of physiological signals and facial activation units to explore correlations between personality traits and emotions. LAUREATE collects Empatica data, alongside self-reports and quizzes, to monitor psychophysiological states in students for educational purposes. WESAD captures multiple physiological signals, also using the Empatica device, to detect stress and emotions felt in human-computer interaction contexts.

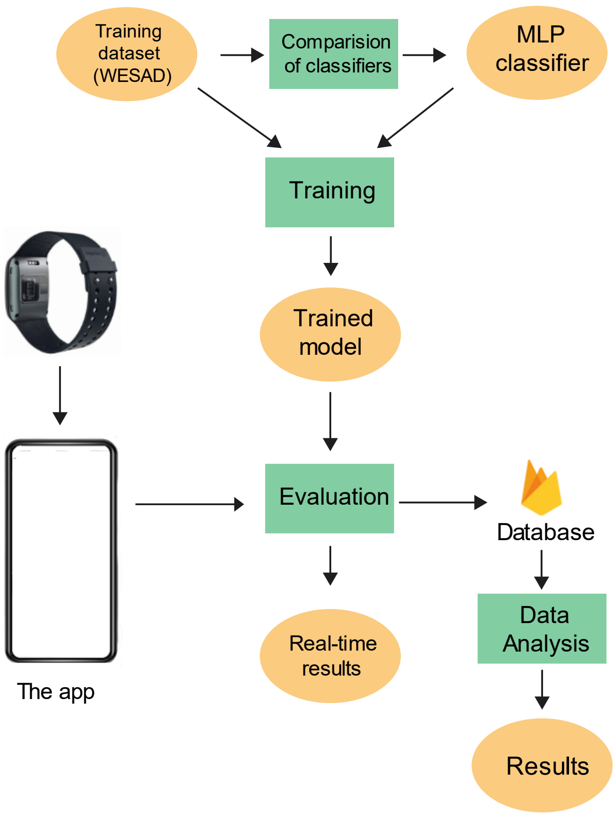

Fig. 3. The classification and evaluation scheme

This one was the dataset chosen for our classifier because it was freely available online, relatively recent (2018), and created using all the sensors of the Empatica E4 (unlike the DEAP, DECAF and ASCERTAIN datasets, which did not use the E4).

WESAD collected data from 17 test participants using Empatica E4 and RespiBAN, a chest-worn device. However, the data collected with RespiBAN were not considered in our analysis. Unfortunately, the WESAD supported only amusement, stress, neutrality and meditation emotions.

To create the WESAD dataset, participants were subjected to various activities designed to induce stress and emotional changes while simultaneously recording biometric signals and annotations of emotional states. Subsequently, the data were preprocessed and made available as an annotated dataset for the scientific community [28].

Once we chose the dataset, the second step involved preparing the data for model training. The WESAD dataset consists of separate folders for each user, each containing CSV files with physiological data. First, we combined the data from all users into an initial dataset. We then normalised the data to the lowest sampling frequency (4Hz), corresponding to the EDA and TEMP data, due to their initially different sampling rates – as described in Section 3. Next, the data were grouped into 10-second intervals (since the data were sampled at 4Hz and we wanted a 10-second window, each window contains 40 rows). To determine the label to assign to the grouped data, the most frequent label within the window was calculated by computing the statistical mode.

We then split the data into a training set and a test set, with 20% of the dataset allocated to the test set. As the last step, we tested multiple classifiers - Multi-layer Perceptron (MLP) [41], Random Forest [43], Support Vector Machines (SVM) [44], and multinomial logistic regression [40] - to compare their effectiveness in solving the specific problem at hand. We chose them as they represent some of the most widely used methods in the field. Our evaluation criteria included metrics such as the confusion matrix, accuracy and F1-score. Based on the results (see Table 1), we selected one of the best-performing models, the MLP model (also for the practical advantage of using a framework with which we were already familiar).

Table 1. Classifiers evaluation results

CLASSIFIER	ACCURACY	F1-SCORE
Multilayer perceptron (MLP)	0.983	0.983
Multinomial logistic regression	0.40	0.225
Random Forest	0.9967	0.9967
Support Vector Machine (SVM)	0.356	0.3088

To label the physiological data of measurements and return the predicted emotion percentages from our classifier, we used a Flask server, a Python-based micro web framework. In the Android application, the connection to Flask is made using the Volley library.

The Flask server receives the physiological data of a measurement in JSON format from the Android application via an HTTP POST request. Then it extracts and normalises the data: since the classifier was trained using data grouped into 10-second intervals at a 4Hz sampling rate (resulting in 40 rows per window), the EDA and Temp data, sampled at 4Hz, are concatenated. For all signals, a normalization function ensures that each row contains exactly 40 values.

The pre-trained MLP model is then loaded to make predictions, returning probabilities for each output class (the four labels *neutrality, stress, meditation, amusement*). The results are organised into a final dataframe, where each row contains the four emotions with their respective percentages and the corresponding timestamp. The dataframe is converted to JSON format and is returned as an HTTP response.

The results are displayed in a text element of the user interface and are also saved in the Firestore Database.

4 User Test

To evaluate the effectiveness of a wrist-worn wearable device (Empatica E4) in emotion recognition, with the support of an associated classifier, we conducted a targeted study within a controlled laboratory setting. We selected seven scenes drawn from previous research [14] considered capable of evoking one specific emotion (amusement, anger, contentment, disgust, fear, neutrality, sadness and surprise). The previous research [14] involved five years of work including an evaluation of over 250 films shown to 494 subjects to identify the relevant videos. The selected clips for our experiment and their corresponding emotions are as follows:

- Amusement: When Harry Met Sally, scene “Discussion of orgasm in café”
- Anger: Cry Freedom, scene “Police abuse protesters”
- Satisfaction: scene of waves on the beach
- Neutrality: scene of abstract shapes
- Fear: The Shining, scene “Boy playing in hallway”
- Sadness: Bambi, scene “Mother deer dies”
- Surprise: Capricorn One, scene “Agents burst through door”

We conducted a user test with 30 subjects (16 men and 14 women). The average age was 35 (range 21-73, SD 13.77). Of these participants, 80% (24 individuals) had a bachelor's degree, 10% (3 individuals) held a doctoral degree, and 10% (3 individuals) had completed high school. Participants were recruited by asking for volunteers among those present in our laboratories, without receiving any form of compensation. Before participation, all subjects signed a consent form allowing us to collect and process their data. This form provided an explanation of the experiment, detailing that participants’ physiological data would be collected using the wristband while they watched movie scenes, and that they would be asked to fill out a questionnaire. However - to avoid any influences - the users were not informed beforehand that the test was about emotion recognition, so that their responses were genuine and spontaneous.

The test required about 20 minutes and was conducted as follows:

- An account was created in the application for each user.
- Users were asked to sign the consent form and equipped with the E4 wristband.
- Each participant watched the seven video scenes, one at a time, while the wristband-generated data were recorded, analysed by the classifier and saved on the app. To avoid the influence of a predetermined order of the video scenes, the videos were shown in a different cyclical order for each user.
- After watching each scene, the users were asked to complete the questionnaire.

In the questionnaire, participants first provided demographic information; then for each clip viewed they had to indicate:

1. Which of the seven emotions from the research [14] the video evoked;
2. Which of the four emotional states identified by WESAD (*amusement, stress, neutrality* and *meditation*), they experienced;

3. Their levels of pleasure and arousal using the Self-Assessment Manikin (SAM) scale, ranging from 1 to 9 (with 1 indicating the minimum level and 9 the maximum). This followed the Russell's Circumplex Model and its representations of emotions in a two-dimensional space defined by the dimensions of valence and arousal [33].

The user test was designed to better understand potential correlations between the emotion perceived watching the seven scenes and the physiological signals collected by the Empatica E4 wristband; and to validate the classifier by comparing its classifications with users' responses in the questionnaire.

4.1 Test Results and Data Analysis

After the user test, we analysed the collected data (the users' physiological data, their questionnaire responses, and the classifier results). This section provides an overview of the analysis in relation to the user test goals.

Correlation between Physiological Data and Reported Emotions

To understand whether there was a correlation between the collected physiological signals and the emotion the users reported in the questionnaire (among the seven associated with the videos), we applied a multiple regression analysis [39].

This analysis predicts the dependent outcome variable (in this case reported emotions) on the basis of several independent predictor variables (physiological data). Thus, for each physiological signal, all values received within each 10-second interval were aggregated to compute a mean value. We used the one-hot encoding method [42] to transform categorical variables (the perceived emotions) into numerical format, with the aim of converting categorical variables into a series of binary variables but ensuring that the categorical nature of the data was preserved.

Next, we calculated the coefficient of determination R^2, the R^2 adjusted and the p-value. The coefficients from the regression indicate how much each physiological signal influences the reported emotions. Signals with the highest coefficients represent those that most contribute to the correlation, considering only the values that are statistically significant. The results obtained are illustrated in table 2.

Looking at R^2 and R^2 adjusted, we can see a moderate correlation between the physiological signals and the perceived emotions (for twenty-five users $R^2 > 0.50$; for thirteen users $R^2 > 0.60$; for twenty-one users R^2 adjusted > 0.50; for twelve users R^2 adjusted > 0.60).

Temperature (*temp*) and galvanic skin response (*GSR*) had the greatest influence on supporting the correlation with perceived emotions. This confirms that sweating and temperature changes are direct physiological responses to emotion-al stimuli (the primary controller of sweating is the integration between internal and skin temperatures [29]).

Table 2. Results of the multiple regression analyses

User	R^2	R^2 Adjusted	Signals with max significant coefficient
1	0.462	0.412	gsr
2	0.582	0.543	gsr
3	0.505	0.459	gsr
4	0.672	0.641	temp, Z
5	0.652	0.6196	gsr
6	0.189	0.114	temp
7	0.663	0.632	gsr
8	0.678	0.648	gsr
9	0.589	0.5498	gsr
10	0.661	0.6295	temp
11	0.513	0.467	Y, Z, temp
12	0.363	0.303	temp, gsr
13	0.545	0.503	gsr
14	0.585	0.546	temp
15	0.599	0.561	temp
16	0.567	0.525	temp, gsr
17	0.502	0.456	gsr, temp
18	0.422	0.366	gsr
19	0.644	0.6098	temp
20	0.480	0.431	gsr
21	0.635	0.601	temp
22	0.608	0.571	gsr
23	0.516	0.470	temp
24	0.581	0.541	temp
25	0.690	0.661	temp
26	0.671	0.639	Z, temp
27	0.556	0.519	Z
28	0.719	0.692	X
29	0.794	0.774	gsr
30	0.659	0.627	bvp

We also explored the sensitivity of physiological signals, such as electrodermal activity (EDA) to the stimuli provided by the selected videos. As indicated by Babaei et al. [2] EDA is extensively used in psychophysiology to assess arousal, attention and stress levels. This physiological response is particularly valuable in emotion research, as it provides insights into the intensity of emotional arousal, even though it cannot differentiate between positive and negative emotions. This makes EDA an essential tool for understanding the physiological underpinnings of emotional experiences, as demonstrated by by the variations in EDA levels observed in our study.

The bar chart in Figure 4 provides a visual comparison of the mean EDA levels for the seven scenes, each designed to evoke a specific emotion. This visualisation helps to identify patterns in the physiological response associated with each emotional stimulus. By analysing the bar-chart, it is evident that scenes with strong emotional stimuli, such as "Cry Freedom" (anger) and "The Shining" (fear), have significantly higher average EDA values, reflecting heightened physiological arousal reported by users. Conversely, scenes with milder emotional content, such as "Waves" (contentment) and "Abstract Shapes" (neutrality), show lower average EDA values, consistent with less intense emotional engagement.

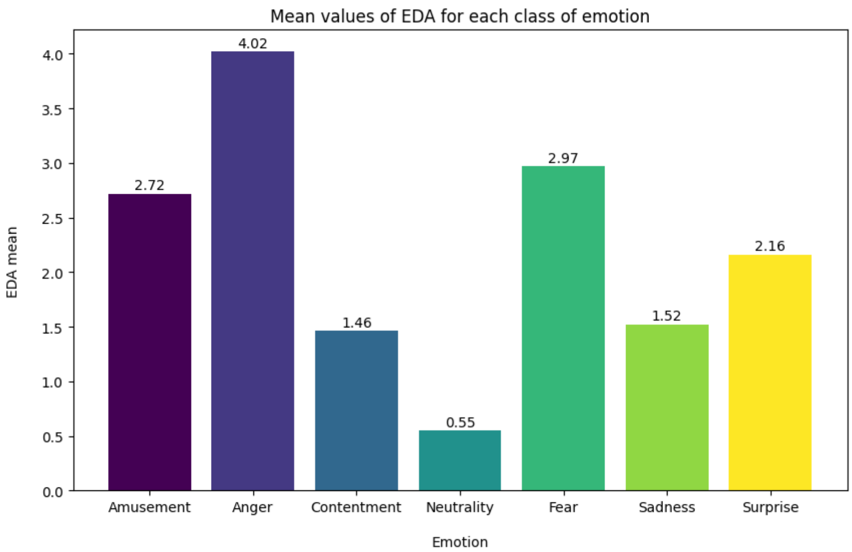

Fig. 4. Mean values of EDA for each video

Performance of the Automatic Classifier

In order to assess the results of the classifier, users were explicitly asked to indicate which of the four emotions detected by the WESAD dataset they would associate with each video.

By comparing the users’ responses to the questionnaire with the emotions detected by the classifier, we categorised results as follows: total correspondence, partial correspondence and no correspondence. The classifier detected automatically the percentages of each of the four emotions (*amusement*, *stress*, *meditation* and *neutrality*). Total correspondence means that the emotion identified by the classifier with the highest value matched the emotion indicated by the user in the questionnaire; partial correspondence means that the emotion selected by the user in the questionnaire was between the emotions detected by the classifier, but not with the highest percentage; no correspondence means that the emotion indicated by the user was not detected at all by the classifier.

Globally, across the 7 tested scenes, the percentage of total and partial matches was over 50% for 5 scenes (‘When Harry met Sally’, ‘Waves’, ‘Abstract Shapes’, ‘Shining’, ‘Bambi’), while ‘Capricorn One’ had about 47% of matches and ‘Cry Freedom', the most problematic, had about a 20% match rate. The lower match rates for these two scenes could be attributed to factors such as the challenging nature of their emotional content (surprise and anger), which may not align well with the classifier’s detection capabilities.

Effectiveness of the seven scenes in eliciting specific emotional responses

We have also analysed the actual effectiveness of each of the seven scenes in stimulating the pre-established emotion. For this purpose, we compared the emotional state indicated by the users in the questionnaire (choosing between amusement, anger, sadness, neutrality, surprise, satisfaction and fear) and the emotion associated with the video scene, according to the research of Gross and Levenson [14]. The results showed a strong correspondence in these cases: the match was 80% for the scene designed to evoke amusement, 87% for neutrality, 87% for sadness, and 67% for surprise. Instead, a lower correspondence was found for: anger (the match was 57%), contentment (the match was 47%) and fear (the match was 42%).

We also aimed to explore how the affective pleasure and arousal perceived by participants for each of the seven scenes relate to Russell's model of emotions. For each of them, the users provided a value between 1 and 9 for the perceived pleasure and arousal (where 1 indicated the lowest value and 9 the highest). Figure 5 shows scatterplots (b, c, d) with the distributions of pleasure and arousal values (the ovals have the mean values as centre and their sizes correspond to the standard deviation). To simplify the presentation and focus on the most representative results, we chose to display only the cases of amusement, contentment, and sadness. Comparing these with the positioning of emotions in Russell's model (a), it is possible to see that there is relatively good correspondence, indicating that the videos were effective in evoking the intended emotions in these cases.

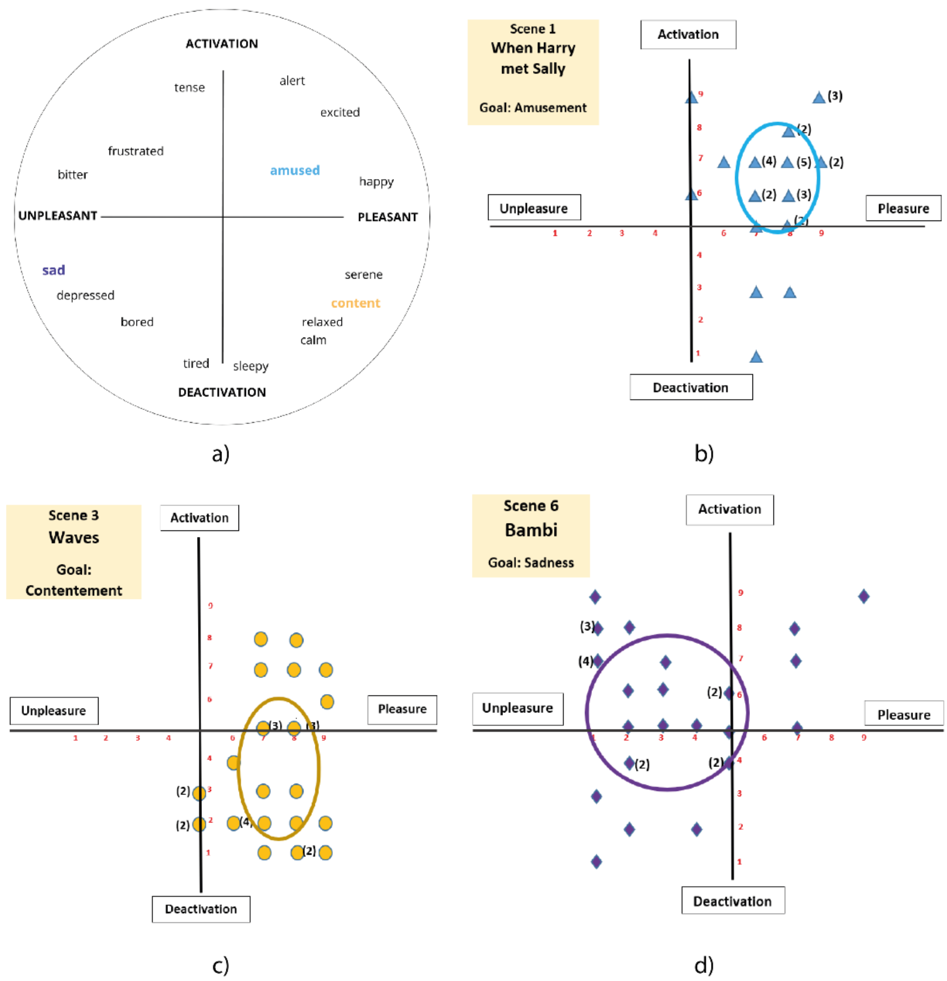

Fig. 5. Russell's circumplex model (a) compared with scatter plots of the distributions of pleasure and arousal for the emotion videos of amusement (b), contentment (c), and sadness (d).

5 Discussion and Limitations

Our study aims to investigate the actual possibilities of wrist-worn wearable technology by collecting real-time physiological data, while participants engage in a natural viewing experience. This approach allows for continuous monitoring and provides dynamic insights into how physiological signals correspond to emotional experiences over time.

The analysis of the test results suggests several considerations. First, the discrepancy between users' responses and the expected emotions in some cases (such as fear and contentment) could be attributed to various factors, including subjectivity in interpreting emotions and the impact of familiarity with the scenes (a few participants had previously seen several of the scenes).

Eliciting an intense emotion in just a few minutes of video is also challenging. In addition, to fully appreciate some scenes and immerse oneself completely in the situation, it might be necessary to know the broader context of the film or have seen the entire movie. In addition, a laboratory environment with the presence of some observers (the authors of the paper) could have influenced the participants' feelings.

The performance of the emotion classifier, while promising, also indicates areas for improvement. The discrepancy between the seven emotional states associated with the scenes [14] and the four labels of the WESAD dataset has introduced some ambiguity in the interpretation of emotional labels. For example, if a scene was intended to evoke sadness (which is not one of the four WESAD categories), users had to select another emotion that might not correspond well, potentially leading to a misalignment between their perceived emotion and the available options. However, the classifier achieved over 50% correspondence for five of the seven tested scenes, demonstrating a baseline level of effectiveness but highlighting the need for further refinement.

Despite these challenges, the analysis of average Electrodermal Activity (EDA) values across different scenes revealed distinct patterns of physiological arousal corresponding to various emotional stimuli. This approach aligns with the study by Rahman et al. [23] which only focused on EDA to recognize emotional states. While both studies analysed fear, anger, sadness, and neutrality, our study also included amusement, surprise and contentment. Conversely, the study by Rahman et al. [23] examined surprise, happiness and disgust, which were not included in our study. Our findings indicate that anger elicited the highest EDA mean value, followed by fear and amusement. Rahman et al. [23] also identified anger and fear among the higher EDA responses. Both studies observed that emotions such as neutrality and sadness are associated with lower EDA values. The findings confirm the utility of EDA as a reliable indicator of physiological responses to emotional stimuli, highlighting its effectiveness in capturing variations in emotional intensity.

Interestingly, the study by Wang et al. [31] employed an adaptive emotion recognition approach that integrated context aware analysis by adjusting the emotion classifier according to the user's activity scene (e.g. walking, running, sitting). The study's scene-specific classifiers achieved an average accuracy of 74.3%, with the highest accuracy in sitting scenes and the lowest in running scenes, underscoring the impact of activity context on data quality. This comparison suggests that while our approach captures emotional responses in a controlled setting, incorporating context-aware elements like those used in the adaptive approach (the user's activity) [31], could further refine the accuracy and reliability of emotion detection in real-time scenarios.

In light of our findings regarding the limitations of using only a wearable device for emotion recognition, it becomes evident that a single-source approach may not capture the full spectrum of emotional responses. Indeed, an important area of development concerns the integration of multimodal data. When combined with multiple physiological signals, additional data from environmental sensors, audio or video recordings and behavioural observations can provide a more comprehensive understanding of users' emotional dynamics. Additional data streams such as blood pressure, respiratory rate, EEG waves for brain activity, and eye tracking metrics (e.g., pupil dilation) can offer deeper insights of internal changes in relation to emotional responses.

Wearable emotion recognition technology, as explored in this study, could find applications in fields such as mental health or education. In the domain of UX evaluation, this technology can help how designers and developers assess emotional responses during user interactions. By providing real-time physiological data, wearable devices enable the detection of users' emotional responses, such as frustration during a confusing user interface interaction or satisfaction during an intuitive and easily understandable one. Such insights allow UX professionals to refine products and services to create more intuitive, emotionally engaging, and user-centred experiences, improving overall satisfaction.

As we mentioned, there are still challenges in accurately distinguishing emotions and accounting for the contextual dependency of physiological responses. Therefore, it may be necessary to complement wearable data with observational methods, self-reports, and contextual analysis to develop a more comprehensive framework for UX evaluation.

6 Conclusion and Future Work

In conclusion, our study underscores the evolving landscape of emotion recognition technology, particularly through wearable devices such as the Empatica E4 bracelet. By addressing the two main research questions, we have identified both the potential and limitations of current wristband-based technologies in capturing nuanced emotional states. Our findings indicate that while wearable devices offer valuable real-time data on physiological responses and benefit from their non-invasiveness, they are limited in their ability to fully capture the complexity of human emotions, an area that necessitates a more comprehensive approach which integrates multiple data sources.

Thus, this work offers practical recommendations for leveraging such technologies. These include incorporating multimodal data, refining emotion classification techniques, and addressing contextual dependencies. Addressing these gaps could lead to tools capable of not only predicting emotions with greater precision but also responding adaptively to users' nuanced emotional states.

The data generated by the users during the test (their physiological data detected by the Empatica E4 device and their self-reported emotions) have been used to create a new publicly available dataset, which can be exploited for further studies.

Moving forward, the integration of multimodal data promises to enhance the accuracy and contextuality of emotion recognition systems. Future work will consider several directions. One will aim to identify the most relevant combinations of physiological data with additional data streams, such as environmental sensors, audio or video recordings and behavioural observations, to provide a more comprehensive understanding of users' emotional dynamics.

The emotional labels collected during testing will be used to further refine our classification algorithm, allowing for emotion classification with a broader set of categories. Subsequent testing of the refined algorithm will enable the system to more accurately and personally learn the relationships between physiological signals and perceived emotions. Additionally, this system can be exploited in adaptive interactions, for example with social robots or Web applications, enhancing their ability to respond to human emotions dynamically and contextually.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abadi, M.K., Subramanian, R., Kia, S.M., Avesani, P., Patras, I., Sebe, N.: DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. IEEE Trans. Affective Comput. 6, 209–222 (2015). https://doi.org/10.1109/TAFFC.2015.2392932
2. Babaei, E., Tag, B., Dingler, T., Velloso, E.: A Critique of Electrodermal Activity Practices at CHI. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–14. Association for Computing Machinery, New York, NY, USA (2021)
3. Cabanac, M.: What is emotion? Behavioural Processes. 60, 69–83 (2002). https://doi.org/10.1016/S0376-6357(02)00078-5
4. Cai, Y., Li, X., Li, J.: Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. Sensors. 23, 2455 (2023). https://doi.org/10.3390/s23052455
5. Cosoli, G., Poli, A., Scalise, L., Spinsante, S.: Measurement of multimodal physiological signals for stimulation detection by wearable devices. Measurement. 184, 109966 (2021). https://doi.org/10.1016/j.measurement.2021.109966
6. De Carolis, B., Loglisci, C., Giuseppe, M., Trufanova, K.: Analyzing Stress Responses Related to Usability of User Interfaces. In: Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter. pp. 1–9. Association for Computing Machinery, New York, NY, USA (2023)
7. Di Lascio, E., Gashi, S., Santini, S.: Unobtrusive Assessment of Students' Emotional Engagement during Lectures Using Electrodermal Activity Sensors. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2, 103:1-103:21 (2018). https://doi.org/10.1145/3264913
8. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology. 17, 124–129 (1971). https://doi.org/10.1037/h0030377
9. England, R.: The Cognitive/Noncognitive Debate in Emotion Theory: A Corrective From Spinoza. Emotion Review. 11, 102–112 (2019). https://doi.org/10.1177/1754073918798038
10. Epstein, S.: Controversial issues in emotion theory. Review of Personality & Social Psychology. 5, 64–88 (1984)
11. Gendron, M.: Defining Emotion: A Brief History. Emotion Review. 2, 371–372 (2010). https://doi.org/10.1177/1754073910374669
12. Ghiani, G., Manca, M., Paternò, F.: Dynamic user interface adaptation driven by physiological parameters to support learning. In: Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems. pp. 158–163. Association for Computing Machinery, New York, NY, USA (2015)
13. Girardi, D., Novielli, N., Fucci, D., Lanubile, F.: Recognizing developers' emotions while programming. In: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. pp. 666–677. Association for Computing Machinery, New York, NY, USA (2020)

14. Gross, J., Levenson, R.W.: Emotion Elicitation Using Films. Cognition and emotion. 9, 87–108 (1995)
15. Ketonen, E.E., Salonen, V., Lonka, K., Salmela-Aro, K.: Can you feel the excitement? Physiological correlates of students' self-reported emotions. British Journal of Educational Psychology. 93, 113–129 (2023). https://doi.org/10.1111/bjep.12534
16. Lench, H.C., Reed, N.T.: Comment: Can We Model What an Emotion Is? Comment on Suri & Gross (2022) - Heather C. Lench, Noah T. Reed, 2022, https://journals.sagepub.com/doi/10.1177/17540739221089692
17. Li, J., Haeb-Umbach, R., Gong, Y.: Robust Automatic Speech Recognition [Book], https://www.oreilly.com/library/view/robust-automatic-speech/9780128026168/
18. Li, S., Deng, W.: Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing. 13, 1195–1215 (2022). https://doi.org/10.1109/TAFFC.2020.2981446
19. Mohammadi, G., Van De Ville, D., Vuilleumier, P.: Brain networks subserving functional core processes of emotions identified with componential modeling | Cerebral Cortex | Oxford Academic, https://academic.oup.com/cercor/article/33/12/7993/7081419
20. Motti, V.G., Caine, K.: Human Factors Considerations in the Design of Wearable Devices. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 58, 1820–1824 (2014). https://doi.org/10.1177/1541931214581381
21. Motti, V.G., Caine, K.: An overview of wearable applications for healthcare: requirements and challenges. In: Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers. pp. 635–641. Association for Computing Machinery, New York, NY, USA (2015)
22. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and psychopathology. 17, 715 (2005). https://doi.org/10.1017/S0954579405050340
23. Rahman, J.S., Hossain, Md.Z., Gedeon, T.: Measuring Observers' EDA Responses to Emotional Videos. In: Proceedings of the 31st Australian Conference on Human-Computer-Interaction. pp. 457–461. Association for Computing Machinery, New York, NY, USA (2020)
24. Ruta, M., Scioscia, F., Cinquepalmi, A., Cipriani, S., Di, E.: Dai biosegnali agli stati emotivi: un approccio semantico.
25. Saganowski, S., Perz, B., Polak, A.G., Kazienko, P.: Emotion Recognition for Everyday Life Using Physiological Signals From Wearables: A Systematic Literature Review. IEEE Transactions on Affective Computing. 14, 1876–1897 (2023). https://doi.org/10.1109/TAFFC.2022.3176135
26. Santhosh, J., Pai, A.P., Ishimaru, S.: Toward an Interactive Reading Experience: Deep Learning Insights and Visual Narratives of Engagement and Emotion. IEEE Access. 12, 6001–6016 (2024). https://doi.org/10.1109/ACCESS.2024.3350745
27. Scherer, K.R.: What are emotions? And how can they be measured? Social Science Information. 44, 695–729 (2005). https://doi.org/10.1177/0539018405058216
28. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 400–408. Association for Computing Machinery, New York, NY, USA (2018)
29. Shibasaki, M., Crandall, C.G.: Mechanisms and controllers of eccrine sweating in humans. Front Biosci (Schol Ed). 2, 685–696 (2010)

30. Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., Yang, X.: A Review of Emotion Recognition Using Physiological Signals. Sensors. 18, 2074 (2018). https://doi.org/10.3390/s18072074
31. Wang, Z., Yu, Z., Zhao, B., Guo, B., Chen, C., Yu, Z.: EmotionSense: An Adaptive Emotion Recognition System Based on Wearable Smart Devices. ACM Trans. Comput. Healthcare. 1, 20:1-20:17 (2020). https://doi.org/10.1145/3384394
32. Yang, B., Wu, J., Hattori, G.: Facial Expression Recognition with the advent of face masks. In: Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia. pp. 335–337. Association for Computing Machinery, New York, NY, USA (2020)
33. Yarwood, M.: Russell's (1980) Circumplex Models. Psychology of Human Emotion: An Open Access Textbook.
34. ASCERTAIN dataset, https://ascertain-dataset.github.io/
35. DEAP: A Dataset for Emotion Analysis using Physiological and Audiovisual Signals, https://www.eecs.qmul.ac.uk/mmv/datasets/deap/
36. E4 wristband | Real-time physiological signals | Wearable PPG, EDA, Temperature, Motion sensors, https://www.empatica.com/en-eu/research/e4/
37. Firebase | Google's Mobile and Web App Development Platform, https://firebase.google.com/
38. LAUREATE: A Dataset for Supporting Research in Affective Computing and Human Memory Augmentation: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies: Vol 7, No 3, https://dl.acm.org/doi/10.1145/3610892
39. LinearRegression, https://scikit-learn/stable/modules/generated/sklearn.linear_model.LinearRegression.html
40. LogisticRegression, https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
41. Neural network models (supervised), https://scikit-learn/stable/modules/neural_networks_supervised.html
42. OneHotEncoder, https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.OneHotEncoder.html
43. RandomForestClassifier, https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
44. Support Vector Machines, https://scikit-learn/stable/modules/svm.html