Multimedia Tools and Applications (2020) 79:35885–35907
https://doi.org/10.1007/s11042-020-09576-0

A user-centered approach for detecting emotions with low-cost sensors

Rita Francese[1] · Michele Risi[1] · Genoveffa Tortora[1]

Received: 10 December 2019 / Revised: 20 July 2020 / Accepted: 11 August 2020 /
Published online: 27 November 2020

Abstract

Detecting emotions is very useful in many fields, from health-care to human-computer interaction. In this paper, we propose an iterative user-centered methodology for supporting the development of an emotion detection system based on low-cost sensors. Artificial Intelligence techniques have been adopted for emotion classification. Different kind of Machine Learning classifiers have been experimentally trained on the users' biometrics data, such as hearth rate, movement and audio. The system has been developed in two iterations and, at the end of each of them, the performance of classifiers (MLP, CNN, LSTM, Bidirectional-LSTM and Decision Tree) has been compared. After the experiment, the SAM questionnaire is proposed to evaluate the user's affective state when using the system. In the first experiment we gathered data from 47 participants, in the second one an improved version of the system has been trained and validated by 107 people. The emotional analysis conducted at the end of each iteration suggests that reducing the device invasiveness may affect the user perceptions and also improve the classification performance.

Keywords Emotion detection · Affective analysis · Artificial intelligence · Low-cost sensors

1 Introduction

Recently, many research efforts have been devoted to the recognition of human emotions. This interest is mainly due to the fact that emotions impact on the uses' reactions and behaviours, and their understanding may be useful in various fields, such as human-computer interaction, software engineering [11], gaming, marketing and multimedia. In

✉ Rita Francese
francese@unisa.it

Michele Risi
mrisi@unisa.it

Genoveffa Tortora
tortora@unisa.it

[1] Department of Computer Science, University of Salerno, Fisciano, Italy

health-care it is particularly useful for specific types of patients, like autistic people or people which are not able to express their sentiments: to recognize their emotions may be useful for caregivers to prevent panic attacks or to better understand their needs.

Emotions are related to the mood of a person and last for few instants, so they are difficult to detect. Many emotion recognition approaches are based on the analysis of the expressions of the user's face, voice, and gestures, even if these user behaviors may be intentionally controlled or hide other emotions. In addition, the user expressions may also be affected by ethnicity and cultures. These problems may be overcome by adopting emotion recognition approaches based on physiological signals, which are not visible at the human eye and immediately reflect the emotional changes. These kinds of signals may be detected by sensors.

The use of low-cost sensors is a relevant additional challenge to let this technology be accessible to all [13]. In addition, it may be useful to evaluate if the device for acquiring the emotion is well accepted by the user and it does not influence his affective state. Indeed, an invasive device could affect the user's emotions. For this reason, emotional analysis may be adopted to collect the user perceptions on the use of the device.

This paper provides the following main contribution: a user-centered design approach to develop an emotion detection system by using: i) noninvasive, wearable, low-cost sensors, ii) artificial intelligence models for classifying emotions and iii) affective analysis trough the Self-Assessment Manikin (SAM) [4] questionnaire for assessing the impact of the system on the user affective state.

We describe in detail the process we follow for developing our emotion detection system. It has been performed in two successive iterations. At the end of each iteration, together with the performance evaluation and the comparison of among the various Machine Learning classification approaches, we conducted an emotional analysis, aiming at collecting the user's feelings on the detection device. The considered classification approaches include Multilayer Perceptron (MLP) neural network, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) neural network, Bidirectional Long Short-Term Memory (Bidirectional-LSTM) neural network, and Decision Tree.

The paper is structured as follows. Section 2 discusses related work concerning sensor-based emotion detection; Section 3 presents the proposed user-centered development process, based on the stimuli design cycle and the system design cycle; Section 4 discusses the threats to validity. Finally, Section 5 concludes the paper and proposes future work.

2 Background and related work

Let us examine how the relevant aspects concerning emotion recognition by using sensors have been treated in the literature, by considering the following key elements.

- ***Elicited emotion.*** Many emotion classification approaches have been proposed by psychologists. At the present, the most accredited are the following two:

    - Ekman et al. [10] consider the emotions as a discrete phenomenon and define a set of basic emotions (six): joy, disgust, anger, fear, surprise and sadness;
    - a (continuous) approach in which emotions are classified in terms of a continuous function on two or more dimensions. The two most accredited continuous models have been proposed by Plutchik [26] and by Russel [30]. In the Russel's Circumplex Model of Affect the emotional space is divided in four

quadrants by the valence and arousal axes. Considering this classification, emotions may be grouped in four classes, each one associated to one of the four quadrants.

- ***Stimuli & Datasets.*** The triggering mechanism for experimental investigations of emotion adopted to set up an emotion classification system may be of various nature:

  - *Pictures.* The International Affective Picture System (IAPS) is a collection of pictures of everyday life natural color photos depicting complex scenes. It inducts the following emotion types: gore, fear, disgust, excitement, erotica, and neutral [20]. These conditions cover the valence spectrum from negative (gore, fear, disgust) to positive (erotica, excitement). The use of this dataset is not public. It may be requested for research purposes.

    An open dataset is testImages_artphoto[1]. The pictures are taken from the Internet by searching the art sharing site and their emotion category was set by the person who uploaded the photo. According to [21], no clear difference in results exists in classification results for the IAPS and the artistic photo image sets.

    The Open Affective Standardized Image Set (OASIS) [19] was born to overcome the use difficulty of IAPS and is open. It consists of 900 color images depicting a broad spectrum of themes, including humans, animals, objects, and scenes, along with normative ratings on two affective dimensions: valence and arousal.
  - *Sound.* The International Affective Digitized Sound system (IADS) datasets is a collection of sounds of everyday life [33].
  - Video. The DEAP dataset contains videos classified in terms of valence and arousal values, on a scale from 1 to 9 [18].

    Another dataset created by using videos as stimuli is the AMIGOS dataset [8], which collects the mood, affect and personality of 40 participants through EEG, ECG, and GSR sensors,
  - *Games*. The emotions of players have been largely investigated in the literature [6] because they can be used to increase the involvement of a player by inducting specific emotional states, by varying the difficulty level or by proposing specific content. In games, emotions may be affected for two reasons: i) the difficulty increases when advancing to the next level; ii) the competence of the player increases with the experience.

- ***Sensors.*** Emotions affect the activity of the Autonomous Nervous System (ANS) which in turn influences various body parameters. Many physiological signals originated from different parts of the human body and are generally used for diagnostics purposes, but they may be also exploited to get information on the user's emotions [1]. These signals include blood volume pulse (BVP) and electrocardiogram (ECG) signals, which both have cardiovascular origin, while the electromyography sensor (EMG) measures the muscle electrical potential. Other signals are heart rate (HR) or heart rate variability (HRV), respiration sensor (RS) and electroencephalogram (EEG). Measures related to the skin are galvanic skin response (GSR), which measures the electrical activity of the skin when the body sweating varies, skin conductance (SC), skin humidity (SH)

[1] https://www.imageemotion.org

and temperature. It has been observed that GSR considerably varies when there are emotional changes with high arousal [5].
- ***Classification techniques.*** Generally, they are based on machine learning approaches, such as Support Vector Machines or Decision Tree. Also deep learning approaches are adopted, such as CNN and LSTM. Information Retrieval may be also used for correlating data [25].
- ***Metrics.*** How is the performance of the classifier measured? Accuracy is one of the most adopted metrics [9].
- ***Invasiveness.*** It concerns the parts of the body which are affected by the sensors, such as arms, head, belt and foot, and the movement impediment.
- ***Emotional assessment.*** The emotion detection device may influence the user's emotion. Thus, an emotional assessment may be useful to identify this kind of problem.
- ***Methodology.*** The description of the general process followed during the development of the emotion detection system.

In the following, we concentrate our attention on research based on biometrics data, provided by one or more sensors. They are classified in Table 1, where for each work we listed the adopted emotional model, the collected emotions, the type of stimuli, the detected physical signals, the selected technology, Invasiveness, the adopted classification techniques, the type of experiment, Emotional Assessment(i.e., how the detected emotions are verified), the obtained results and the proposal of a development methodology. When more classifiers are used the one which reached best results is reported in bold.

In [17], few emotional states are detected (e.g., Sad, Dislike, Joy, Stress, Normal, No-Idea, Positive and Negative) by using a Decision Tree classifier. Four physiological sensors, BVP, EMG, GSR, and Skin Temperature are adopted. The input stimuli were 100 IAPS images, shown as a slide show (5 second for each image). Five different images from each group are shown and then the application asks participants about their current emotional state by using a Likert scale. The accuracy was 98%.

Gong et al. [14] propose to use fusion features from multiple physiological signals: ECG, EMG, respiratory changes (RSP) and skin conductivity (SC). By using the EEMD decomposition method for each signal they extract four kinds of features: time domain, time frequency, nonlinear and intrinsic mode function (IMFs) features. They also use Decision Tree classifier to choose which features largely contribute to the classification performance. These features are provided as input to another classifier for recognizing emotions.

Myroniv et al. [24] to collect user emotional data, connect heart rate, body temperature and galvanic skin response sensors to an Arduino Yun microcontroller. A sliding window-based segmentation method segments the data and candidate features are extracted from the segments. To classify the extracted features six classification algorithms have been trained: Random Tree, J48, Nave Bayes, SVM, KNN and MLP. A high recognition accuracy of 97.31% has been reached.

The Self-Assessment Manikin (SAM) questionnaire is adopted in [28] to assess the participants' emotions for training the emotion classifier. The selected emotions are: Displeasure, Neutral, Pleasure, Calm, Medium, and Excited measured in the Valence-Arousal model. The performance of three classification techniques is investigated (i.e., ANN, SVM, and Decision Tree). Best accuracy results have been reached by SVM. The sample was composed of 23 volunteers. The invasiveness is medium: the user has to wear belts and sensors are on an arm and on the foot.

Girardi et al. [13] adopt EEG, EMG, and GSR low-cost sensors to detect high vs. low emotional valence and arousal. They also evaluate what are the most relevant physiological

Table 1 Emotion detection based on sensors related work

Ref.	Model	Emotion	Stimuli	Sensors	Technology
Khan and Lawo [17]	–	Stress, joyhappy, sad, normalneutral, dislike, no-idea, positive and negative	IAPS images	EMG, BVP, GSR, ST	Raspberry Pi
Gong et al. [14]	Russel	Joy, anger, sadness, pleasure	Private	ECG. EMG, RSP, SC	–
Myroniv et al. [24]	Russel	Negative, Neutral, Positive	Geneva Affective Picture Database and IAPS	HR, BPM, GSR, T	Arduino Yun
Girardi et al. [13]	Russel	High vs. low valence and arousal	DEAP video	EEG, EMG, GSR	Neuroview acquisition and ConsensysPRO software
Chao et al. [7]	Russel	valence and arousal	predefined data (RECOLA), including sensor DATA	ECG and EDA	
Santamaria-Granados et al. [31]	Russel	High vs. low valence and arousal	AMIGOS	ECG and GSR	–
Rattanadoung et al. [28]	Russel	Fear, anger, sadness, joy, disgust, surprise, trust, anticipation	IAPS, IADS combination	HR, breathing pattern, temperature, skin humidity, SC	Arduino
Pollreisz and TaheriNejad [27]	–	Happiness, Sadness, Anger, Pain	one video	HR, SKT, and EDA	Empatica E4
Shu et al. [32]	–	neutral, happy and sad	CEVS	HR	Algoband F8
This paper	Russel	Sad, contentment, anger, fear	TestImages Artphoto Dataset	HR, audio, movement	Arduino Nano

Table 1 (continued)

Ref.	Invasiveness	Classification	Emotional Assessment	Results	Methodology
Chao et al. [17]	Arm and finger	Decision Tree	Emotion questionnaire	99.4% acc.	N
Gong et al. [14]	–	Decision Tree	Unspecified	88% acc.	N
Myroniv et al. [24]	Fingers	Random Tree, **J48**, Naive Bayes, SVM, KNN, MLP	–	J48 88.29% acc.	N
Girardi et al. [13]	Hand, harm, head	Naive Bayes, **SVM**, J48	The original DEAP scores for each video	0.585 F-measure with SVM and all the sensors	N
Chao et al. [7]	–	LSTM	original RECOLA scores	arousal = 0.718 (PCC) and valence = 0.627 (PCC)	N
Santamaria-Granados et al. [31]	–	**CNN**, Naive Bayes, AdaBoost, Random Forest, etc.	original AMIGOS scores for each video	81% acc. for arousal with ECG and & 75% for valence with GSR	N
Rattanadoung et al. [28]	Hand, foot, waist belt	ANN, **SVM**, Decision Tree	SAM59-63	acc: 59.54% valence 63% arousal	N
Pollreisz and TaheriNejad [27]	wrist	Decision Tree	SAM	65% success rate	N
Shu et al. [32]	wrist	KNN, RF, Decision Tree, GBDT, **AdaBoost**	–	96% acc.	N
This paper	Harm, hand	MLP, CNN, LSTM, Bidirectional-LSTM, **Decision Tree**	User feedback	91.47% acc.	User-centered design

measures. To this aim an empirical study on 19 participants have been conducted. Eight videos have been selected from the DEAP dataset, two for each quadrant of the Circumpflex model of affect. Performance has been measured in terms of Precision and Recall and three classification methods are compared while using the different sensors. Best results have been reached with SVM and all the three sensor data.

Chao et al. [7] adopted LSTM neural network for emotion recognition in terms of arousal and valence. The network is trained on existing dataset, RECOLA, containing emotional data of subjects who are recorded by audio, video, ECG and Electro-Dermal Activity (EDA). The dataset is annotated on terms of arousal and valence values. The approach is compared against a baseline in terms of Root Mean Square Error (RMSE) and Person Correlation Coefficient (PCC).

CNNs have been used for emotion detection in [31]. The Russel model has been selected for classifying emotions and the adopted dataset was AMIGOS, composed of 16 short and four long videos related to arousal and valence classification. The experiment involved 40 participants and compared the CNN results with respect to classical Machine Learning Algorithms, including Naive Bayes, AdaBoost, Random Forest. Best results were reached by CNN with both ECG and GRS signals. In particular, the former performs better for arousal detection, while the latter for valence.

Emotion detection based on Electroencephalography (EEG) signals have been proposed in [22] to select the optimal combination of these features for recognition. The sample was composed of 21 participants who were solicited by four types of emotional stimuli (happy, calm, sad, or scared). Balanced one-way ANOVA has been adopted to identify after calculating the Hjorth parameters for different frequency ranges. Features selected by this statistical method outperformed univariate and multivariate features. The optimal features were further processed for emotion classification using support vector machine, k-nearest neighbor, linear discriminant analysis, Naive Bayes, random forest and deep learning.

Smart Watches also have been adopted for emotion recognition. The approach proposed in [27] identifies the emotions with a success rate of 65% by using EDA, SKT and HR provided by the wearable device. The classification is done via a decision tree. Shu et al. [32] use heart rate for detecting three emotional states: neutral, happy and sad. Stimuli are taken from the China's Standard Emotional Video Stimuli Materials Library (CEVS). They evaluate five classifiers: k-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree, Gradient Boosting Decision Tree (GBDT), and Adaptive Boosting (AdaBoost). AdaBoost gets the higher accuracy, around 96%.

In our paper, we propose an iterative user-centered development process for emotion detection based on low-cost sensors. Each step of the process has been detailed by considering a running example, where the training stimuli are images taken from the testImages_artphoto dataset, the data collected by the sensors HR and HRV, a gyroscope for detecting user movements, together with audio data, after pre-processing are provided as input to different classifiers: MLP, CNN, LSTM, Bidirectional-LSTM and Decision Tree. After the experiment, the SAM questionnaire is proposed for collecting the user emotions related to the use of our detection system, to verify if the adopted device may influence the experiment results. This is relevant because these emotions may interfere with the ones solicited by the stimuli provided during the experiments. The process is iterated until performance and affective states of the users are satisfying.

3 The proposed method

In this section, we describe the emotion detection system and the iterative user-centered process we adopted to develop it, depicted in the UML activity diagram with object-flow in Fig. 1. The process guides the implementation of an emotion detection system. It includes two research cycles: Stimuli Design cycle and System Design cycle.

A user feels emotions when exposed to external stimuli. One of the main challenges is to select the kind of stimuli on which the system has to be trained, including images, audios and videos. Generally, this material is extracted by public datasets. The stimuli design cycle involves expert users in focus group activity until the experts are satisfied with the proposed approach. In the system design cycle the hardware of the system is designed and implemented. Then it is used for collecting user emotions. In particular, the user reacts to the stimuli and his reaction is acquired by biological data sensors. The data are then pre-processed to be organized according to the requirements of the Emotion Classification activity. To this aim, many different classifiers may be adopted, as discussed in Section 2. There is the need of comparing the performance of many of them to select the one seems to be more suitable. New data are acquired until the accuracy of the emotion classification step is appropriate. Finally, the user's perceptions on the use of the system are collected and the process accuracy is determined. If the user's emotional analysis reveals problems or if the researchers are not satisfied with the system performance, the control-flow goes back to the System Hardware Design activity.

The process phases are detailed in the following of this section, where we explain the methodology we propose through a case study concerning the development of an emotion detection system based on low-cost sensors.

3.1 Emotion elicitation

Following the directions of [13], we decide to detect emotions with increasing arousal. We select the following set of emotions: *sad, contentment, anger* and *fear* which have increasing arousal and are used to classify images by the testImages_artphoto [21] dataset. We adopt a small set of emotions because the success rate in emotion detection decreases when increasing the number of emotions [2]. Details on the two design cycles are provided in the following.

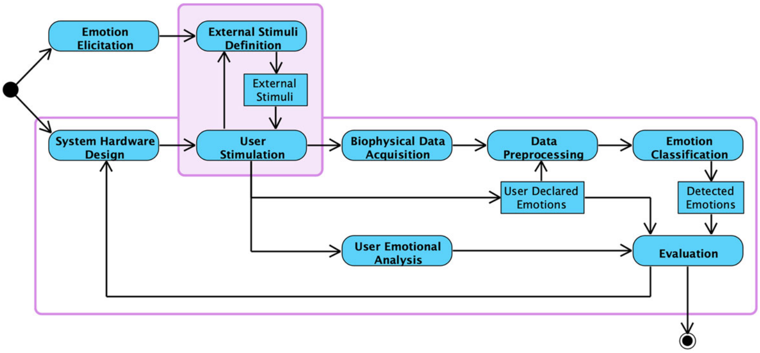


Fig. 1 The emotion detection design process

3.2 Stimuli solicitation design cycle

The focus of this cycle is on the creation of a highly usable interface producing the stimuli related to the emotions selected in the previous phase. We conducted two successive design phases, both in July 2019.

3.2.1 Stimuli solicitation design I

The goal of this first stimuli design session was to identify an appropriate set of stimuli for supporting the system building and to implement the subsystem devoted to the user stimulation. Among the images available in the testImages_artphoto [21] dataset, we selected a subset related to the emotions individuated in the previous phase. We performed a preliminary setting phase with six expert users (computer scientists) in a common room. We used four images for each emotion. They were proposed to the participants as a slide show, each one for ten seconds. Each participant used the right hand to press the buttons of the button pulse. She/he signaled her/his emotion by pressing the appropriate button (sad, contentment, anger and fear). We adopted buttons as input device instead of point-and-click with mouse or the keyboard by supposing that them would have been less distracting and with the aim of getting the expert opinion concerning their use.

The stimuli acquisition duration was 160 seconds, 10 seconds for each image. The final sequence has been ordered by increasing arousal. This choice is motivated by the fact that increasing arousal values are associated with higher arousal levels with respect to an individual baseline values [13, 35].

At the end of the session, we conducted a focus group, discussing the topics reported in Table 2.

3.2.2 Finding from stimuli solicitation design I

A user considered that the last images were observed with less interest. Other two agreed. The images were clear and well visible. The buttons were labeled in English and the label meaning was not clear. So, we translated them in Italian. Buttons were judged easy to press and not distracting, so we decided to adopt them as input device for the user interaction. The lab was an open space with many people. A user observed that a quieter environment may avoid distraction and highlighted the need of an ergonomic chair.

3.2.3 Stimuli solicitation design II

The same participants were enrolled in the the successive version of the stimulation subsystem at the mobile computing lab at the University of Salerno. The room was quiet with soft light. Participants sat on an ergonomic chair with armrests and headrest. The temperature was comfortable and no rumor or distraction occurred. Considering the findings of the

Table 2 Findings from the Stimuli design focus group

Category	Content
Picture Show	Clearness of the images, number of images
Input interface	Buttons, easiness of interaction
Environment	Influence of the environment

previous session, we reduced the image number to eight, two for each emotion. The eight selected images are shown in Fig. 2.

Figure 5 shows the setup of the experiment. The user watches the images which are displayed on the screen of a notebook in increasing arousal. She/he wears the acquisition sensor system on the left hand and presses the emotion buttons with the other hand. The stimuli acquisition duration was 80 seconds, 10 seconds for each image. All the participants agreed that this setting was appropriate.

3.3 System design cycle

The system design cycle aims at design and evaluate the emotion detection system on the base of the stimuli produced by the Stimuli Solicitation Design cycle. In this paper we describe the first two iterations we performed to improve the detection quality.

3.3.1 System design session I

The system architecture follows a client-server approach (see Fig. 3). It is composed of the Data Acquisition and the Data Management subsystems, the client and the server depicted in the left and in the right part of the figure, respectively, described in the following.

- *Data acquisition subsystem.* Our data acquisition hardware solution is based on Arduino Nano (see Fig. 4f), which is connected to various sensors for the acquisition of the biophysical parameters. To let users move their arm, components should not be wired. Thus, each Arduino through the communication component sends the data by a wi-fi ESP-8266 module, depicted in 4e, which provides the data to the server via HTTP. At first, we adopted a single Arduino Nano for all the sensors, but we noticed that each sensor collects and transmits the data at different times. So, we preferred to associate an Arduino Nano to each sensor.
- *Data management subsystem.* It has a repository-based architecture. The Data Collection components receives the data from the data acquisition system and store them in the database. The collected data are then pre-processed to get a form suitable for being analyzed by one of the Machine Learning Classifier (MLC). The report generator component extracts the data in Excel format.

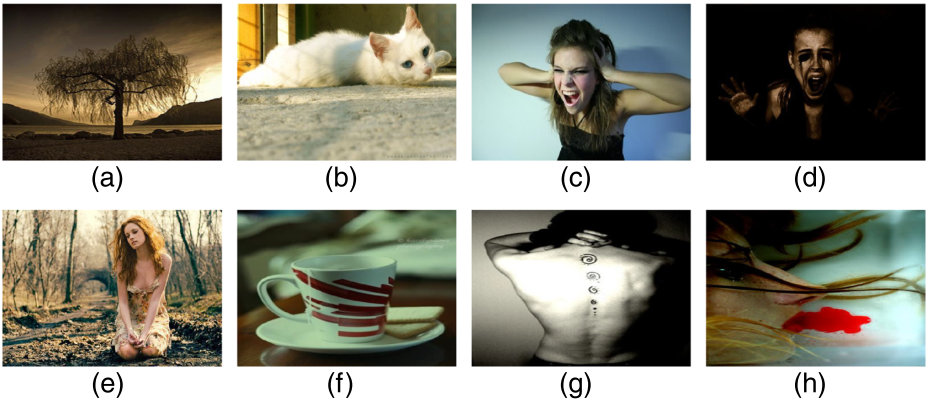


Fig. 2 Images soliciting (**a**)–(**e**) sadness, (**b**)–(**f**) contentment, (**c**)–(**g**) anger and (**d**)–(**h**) fear [21]

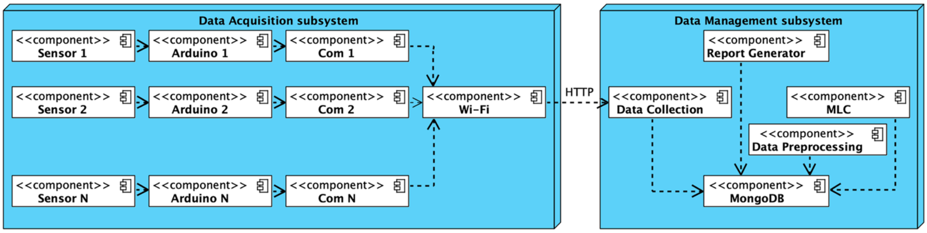


Fig. 3 The system architecture

System hardware design Low-cost and low invasiveness were the design objectives we aim at pursuing. All sensors except the heart rate which is set on the user index are integrated in a single device. Figure 5a shows the first version of the system prototype. The physiological measures we selected for our emotion detection system are the following.

- Heart rate. We tested three different hearth-rate sensors: MAX 30102, KY-039, and HGJVBFGH1 pulse-sensor heart rate. These sensors have been tested taking into account the precision and the accuracy of detected values, and their wearability. MAX-30102 and KY-039 provided good results in terms of accuracy and precision, whilst HGJVBFGH1 had a high response time and low precision. We decided to use the KY-039 for our experiments because it was more stable when mounted on a finger like a ring (high wearability) during hand/arm movements. It is composed of a lighting led and by a photodiode (see Fig. 4a);
- movement, collected through a gyroscope on 3-axis and an accelerometer on 3-axis are on the same chip, MPU-6050 (see Fig. 4b);
- audio, acquired by the sensor KY-037. It may be used to detect a sound over a given threshold (see Fig. 4c).

Biophysical data acquisition During the training and evaluation of the system for each user we collected 20 samples for each emotion, i.e., 80 samples for each user. Each sample was

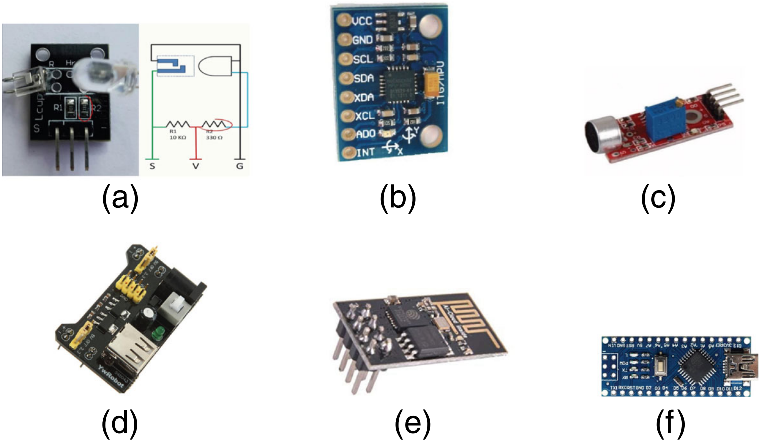


Fig. 4 The selected sensors

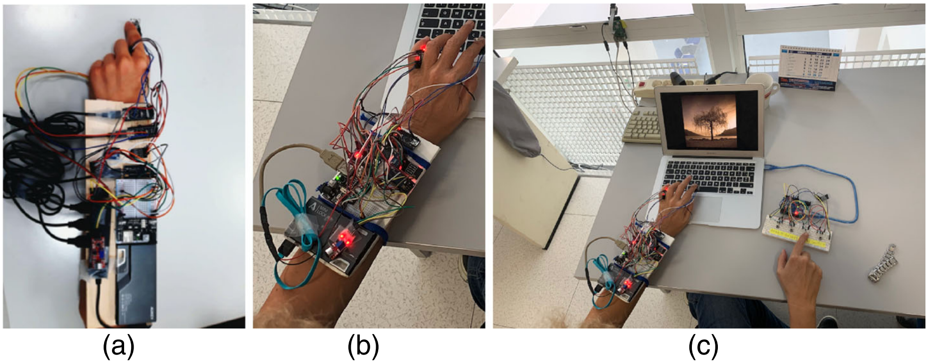


Fig. 5 The (**a**) first and the (**b**) second release of the glove prototype, and (**c**) a user during the data collection

associated to a row of the input data table, where all the data collected by the same sensor are on the same column. In particular, a sample was composed of the following information:

We consider each image as shown separately. Our objective is to collect the user perception on the single image.

Data preprocessing Before providing the input data to the classifier they have to be first synchronized and then normalized.

- *Synchronization.* Each sensor takes a different time for the data computation and sends them at different instants and quantities. To overcome this problem we segmented the sensor data in temporal windows of t seconds each and assign a collection to each sensor. To be able to classify the data, each row has to contain the data of all the sensors detected at the same time, as shown in Fig. 6. To solve this problem the sensor data have to be synchronized as follows. *i)* together with each sensor data also the time in which the datum has been detected is stored; *ii)* define a temporal range and add to the same dataset row (associated to a sensor) all the data in that range. The second step creates new problems: for each sensor it is possible that in the given range no value or more than a value occurs. In the first case a default value has been assigned to indicate

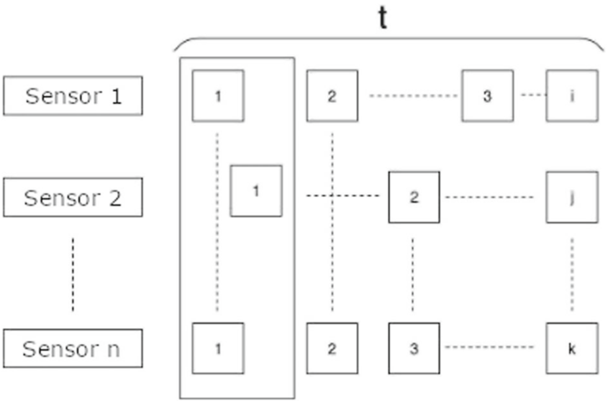


Fig. 6 Data collection in a temporal window of t seconds

that the value has not been detected, in the second one the average value of the sensor data in the range is considered. If at a given time no data is detected it is very likely that the previous data is still valid, because the biometric data does not vary very fast. Thus, we provide a prediction of the missing entries by repeating a previously detected value when in a successive time no value is detected. Whereas in the case in the first instant no value is detected, the first valid value is assigned to the previous ones.

- *Data normalization.* To avoid that input data be on different scales there is the need of normalizing them. The normalization process aims at letting all the values be in the range $[-1, 1]$, while preserving their meaning. The normalization z of a given input value x belonging to the vector $\mathbf{x}$ of input features collected by a given sensor is computed as follows.

$$z = \frac{x\text{-}average(\boldsymbol{x})}{max(\boldsymbol{x})\text{-}min(\boldsymbol{x})} \tag{1}$$

The *minmax normalization* does not handle anomalous values well. To overcome this limitation, data are normalized in blocks of 80 rows, corresponding to the data collected by a single user.

Emotion classification To classify the emotions, we decided to adopt and compare different classifiers widely used in the literature. In particular, we considered the following classifiers:

- Multilayer Perceptron (MLP);
- Convolutional Neural Network (CNN);
- Long Short-Term Memory (LSTM);
- Bidirectional Long Short-Term Memory (Bidirectional-LSTM);
- Decision Tree.

Table 3 Classifiers models implemented with Keras and Scikit-learn libraries

MLP	**CNN**
Dense(9, activation=’tanh’, input_dim=9) Dense(500, activation=’relu’) Dense(300, activation=’relu)) Dense(200, activation=’relu)) Dense(100, activation=’relu)) Dense(50, activation=’relu’) Dense(20, activation=’relu’) Dense(4, activation=’softmax’)	Convolution1D(input_shape=(9, 1)) Activation(’relu’) Flatten() Dropout(0.4) Dense(400, activation=’relu’) Dense(300, activation=’relu’) Dense(200, activation=’relu’) Dense(100, activation=’relu’) Dense(50, activation=’relu’) Dense(20, activation=’relu’) Dense(4) Activation(’softmax’)
LSTM	**Bidirectional-LSTM**
LSTM(256, input_shape=(input_reshaped[1], input_reshaped[2])) Dense(128, activation=’relu’) Dense(64, activation=’relu’) Dense(32, activation=’relu’) Dense(16, activation=’relu’) Dense(4, activation=’softmax’)	Bidirectional(LSTM(256, input_shape=(input_reshaped[1], input_reshaped[2]))) Dense(128, activation=’relu’) Dense(64, activation=’relu’) Dense(32, activation=’relu’) Dense(16, activation=’relu’) Dense(4, activation=’softmax’)
Decision Tree	
DecisionTreeClassifier(2 ** 13)	

The adopted classifier architectures are summarized in Table 3, where we show the main layers with considered parameters developed with the Keras[2] and Scikit-learn[3] libraries. All the neural networks have been trained for 500 epochs, with a batch size equal to 80. Moreover, we used the Adam optimizer and categorical cross-entropy as loss function, in order to update the weights on hidden layers to reduce error. For the Decision Tree we set the maximum depth to 2^{13}.

To assess the predictions of the models, we performed a *k-fold* cross validation ($k = 5$) to split samples into training and testing sets. This kind of validation is widely used to assess how the results of a statistical analysis can be generalized to an independent dataset [12]. Moreover, to provide an unbiased evaluation of a model by tuning the model hyperparameters, the training set has been spitted by considering the configuration 80-20 to produce a validation set.

3.3.2 Experimental evaluation

Experimental unit The prototype shown in Fig 5a has been evaluated on a sample of 47 participants, 28 male and 19 female. Average age 25.44, standard deviation 9.99. At first, participants filled-in a demographic questionnaire. They were all informed on the experiment aim and the experimental procedure. They were volunteers.

Experimental setup and material The experiment was conducted in the Mobile Computing Laboratory of the University of Salerno in September 2019. The room was quiet and the lights were comfortable. We asked participants to wash and accurately dry their hands, because the heart rate sensor put on their index could have been imprecise if the finger was sweaty. As experimental objects, we proposed the eight images we selected, described in Section 3.2.3.

The user feedback on each image is provided by pressing the button corresponding to the mood he felt. This feedback (the last one is considered if multiple feedback are provided) is assigned to all the samples collected for that user on the depicted image.

At the end of the experiment participants had to fill in the SAM questionnaire [4], to collect affective reactions on the whole experiment.

SAM considers the following dimensions: valence (or pleasure), arousal, and dominance, scored by a nine-point rating scale. The SAM questionnaire is depicted in Fig. 7. In particular, pleasure ranges from unhappiness/sadness to happiness/joyfulness (as shown in the first row). Arousal varies from stimulated/excited to calm/bored (it is reversed as shown in the second row). Finally, dominance scale varies from submissive to dominant, i.e., from “without control” to “with control” (see the third row). The three SAM variables may drive the development process: valence should be around the medium score (5), arousal and dominance values should be high. These values assure that the user is in a neutral emotional state, is calm and has the control of the situation.

We also added the liking dimension: a nine-point rating scale ranging from from “dislike” to “like” (see the fourth row in Fig. 7).

Picture was displayed for ten seconds. During this time, the user has to select the button corresponding to the mood he felled.

[2] https://keras.io/api

[3] https://scikit-learn.org

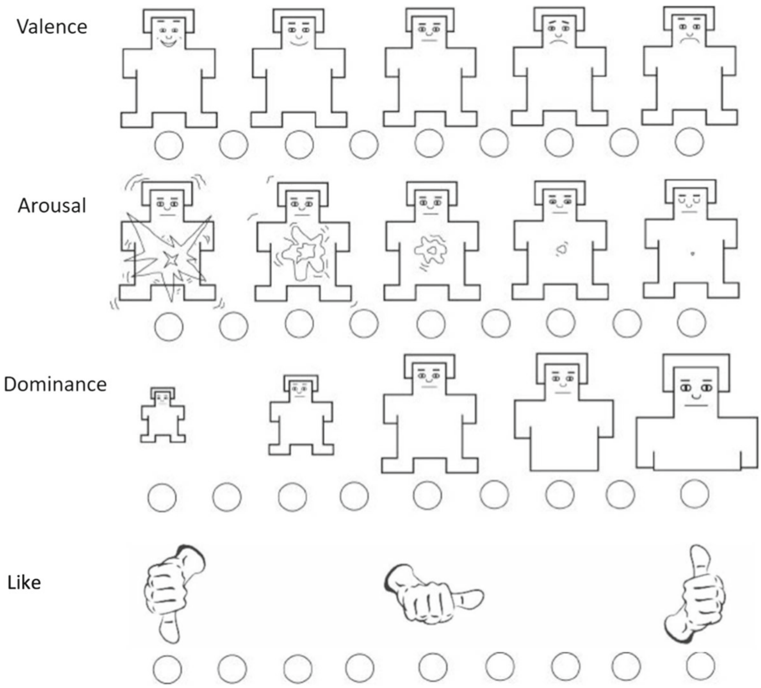


Fig. 7 The SAM [4] and Like scales

Variables and data analysis In this study we used Accuracy as a performance measure of the classification system, defined as follows.

Let T_p and T_n the number of true positives and true negatives, respectively, and N be the total number of examples.

$$Accuracy = \frac{T_p + T_n}{N} \tag{2}$$

The Receiver Operating Characteristic (ROC) curve [3] is often used to measure the classification performance. In particular, we performed the ROC curves analysis on the Accuracy measure. The area under the ROC curve (AUC) provides a metric of the utility of the instrument, with AUC values of 1.0 indicating perfect discrimination. AUC values greater than .80 are indicative of good discrimination, whereas those less than .70 do not provide sufficient discrimination [29].

To measure affective reactions, we used four variables (one for each dimension of SAM): valence (VAL), arousal (ARS), dominance (DOM), and kike (LIK). Median values and frequencies have been adopted as descriptive statistics of these variables [34].

Results The dataset is composed of 3,760 records. A detail of the performance of the classifiers in the first experiment is reported in Table 4 where in bold is highlighted the best results reached between the first and the second experiment, for each classifier. As it is possible to see, best results in the testing set are reached by the Decision Tree classifier with accuracy 84.41%. It is worth noting that we considered the accuracy obtained for the training set to compare that result with the one obtained by the training set when new data is considered.

Left side of Fig. 8 shows the accuracy results reached by the different classifiers while the number of epochs (number of times the model is exposed to the training set) increases,

Table 4 Accuracy improvement

Classifier	Variable	First experiment	Second experiment
MLP	Accuracy training set	88.56%	**90.06%**
	Accuracy testing set	80.98%	**87.68%**
CNN	Accuracy training set	83.38%	**87.05%**
	Accuracy testing set	80.45%	**85.05%**
LSTM	Accuracy training set	88.00%	**88.14%**
	Accuracy testing set	80.32%	**86.21%**
Bidirectional-LSTM	Accuracy training set	87.20%	**88.35%**
	Accuracy testing set	81.12%	**86.97%**
Decision Tree	Accuracy training set	**97.69%**	95.77%
	Accuracy testing set	84.41%	**91.47%**

for the training and the testing sets (blue and red lines, respectively). The ROC curve are shown in the left side of Fig. 9.

Table 5 shows the median, the minimum and the maximum values for VAL, ARS, DOM and LIK. The frequencies of the participant answers are shown in Fig. 11a. They are mainly characterized by low valence values, which denote users a little stressed. Low arousal values denote a state of excitement, agitation or anger. Also, dominance is low for many participants. This means that they perceived as not having the complete control of the situation. Nevertheless, most participants like the idea of the experience, as shown by the Like dimension results ($median = 6$ and $min = 4$). First three emotional states may interfere with the emotion detection system. So there is the need to improve the system for making people feel at ease.

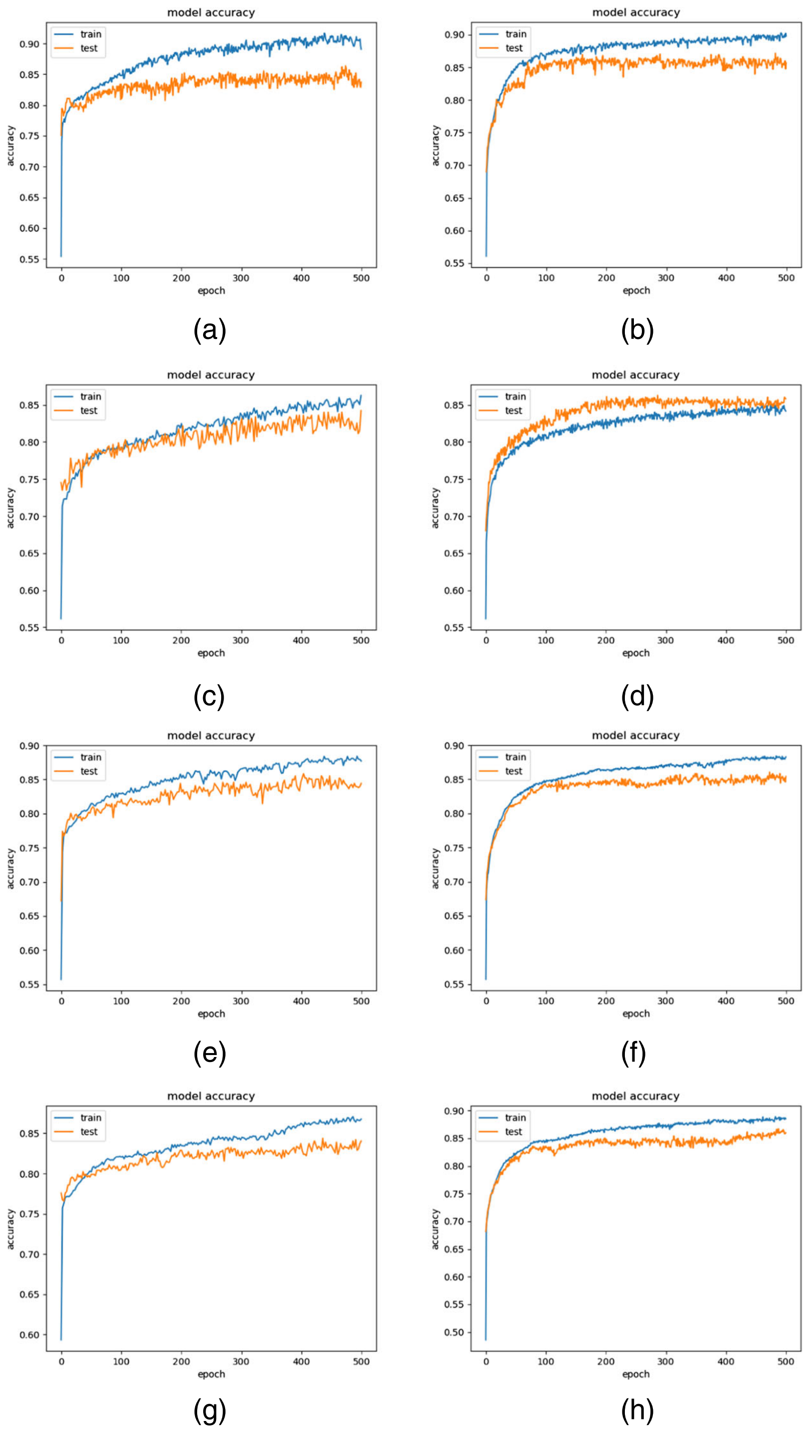
3.3.3 System design second iteration

On the base of the emotional analysis and performance reached in the testing set of the first experiment (i.e., under the 85%), we decided to perform a second iteration to intervene on the following aspects of the system hardware design:

- overweight. The system was too heavy;
- dimension. The device needed a long arm to be easily used. In case of a short arm, the device was annoying on the elbow or on the wrist section.

We investigated the main reasons for the system being overweight: the sensors laid on a one centimeter thick wooden tablet and a hub was adopted for supplying power to the three Arduinos. Thus, we eliminated the hub and the power supply cables by transferring SVs from the battery to the breadboard and by powering the Arduino through GND and Vin pins. The hub elimination allowed us to recover much space and to reduce the prototype dimension by 50% and the weight by 30%.

Second experiment. The prototype shown in Fig. 5b has been evaluated on a sample of 107 participants, (61 male, 46 female, average age 24.75, standard deviation 11.87), by following the same procedure of the first iteration. The final dataset is composed of 8,560 records.



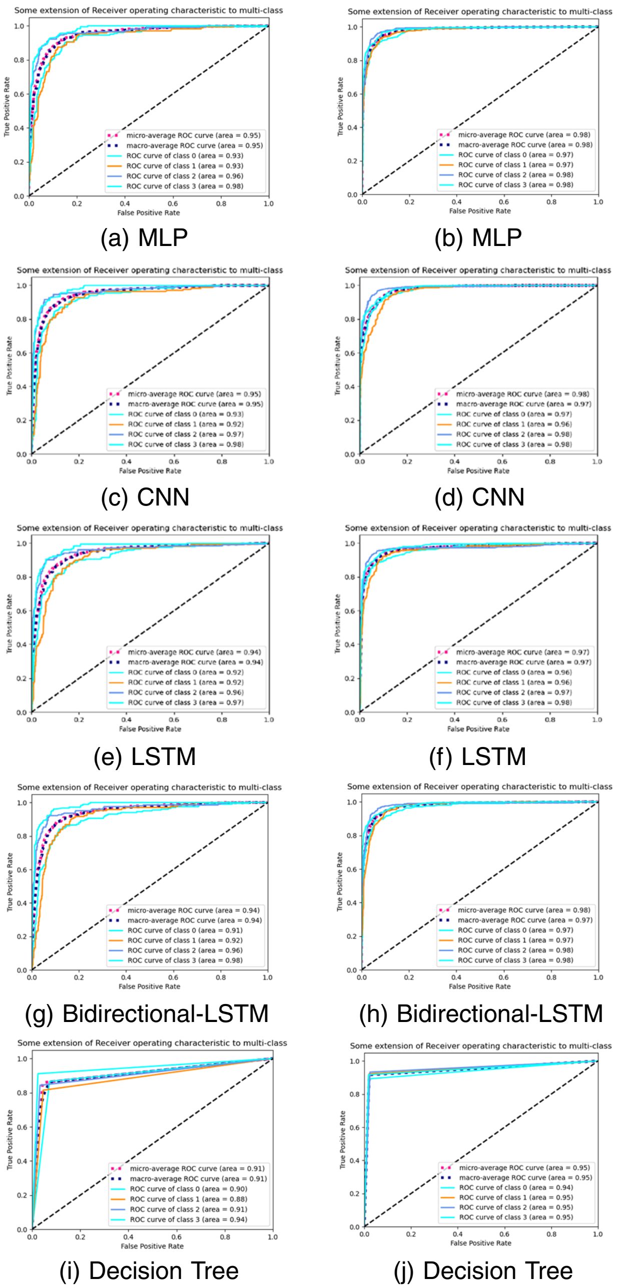
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig. 8 Accuracy of the classifiers in the first (**a**) = mlp_1,-(**c**) = cnn_{1bis},-(**e**)LSTM$_1$,-(**g**) = $blstm_1$ and second experiment (**b**) = mlp_2,-(**d**) = cnn_{2bis},-(**f**) lstm$_2$,-(**h**) blstm$_2$ when the number of epochs increases



(a) MLP (b) MLP

(c) CNN (d) CNN

(e) LSTM (f) LSTM

(g) Bidirectional-LSTM (h) Bidirectional-LSTM

(i) Decision Tree (j) Decision Tree

Fig. 9 ROC curves in the first (**a**) mlp_roc_1,-(**c**) cnn_roc_{1bis},-(**e**) $lstm_roc_1$,-(**g**) $blstm_roc_1$,-(**i**) dt_roc_1, and second (**b**) mlp_roc_2,-(**d**) cnn_roc_{2bis},-(**f**) $lstm_roc_2$,-(**h**) $blstm_roc_2$,-(**j**) $dt_roc_{2(a)}$ experiment

Second experiment results. Again, Decision Tree reached the best accuracy on the testing set (i.e., 91.47%), as reported in Table 4. Right side of Fig. 8 shows the obtained accuracy results while the number of epochs increases. The AUC of Decision Tree reached a value greater than 0.91 (see Fig. 9).

Table 5 Summary of the results for affective reaction

Experiment	Variable	Median	Min	Max
First experiment	VAL	3	1	8
	ARS	4	2	8
	DOM	4	2	8
	LIK	6	4	9
Second experiment	VAL	5	1	9
	ARS	4	2	8
	DOM	5	3	9
	LIK	7	4	9

Table 5 shows the descriptive statistics including median, minimum and maximum values for VAL, ARS, DOM and LIK of the second experiment. The frequencies of participant answers are shown in Fig. 11b. The valence values are split into two halves in the value range (47 of them were negatively activated, 49 are positively activated and 11 neutral). Arousal values denote still an excitement state. Dominance median is 5, several participants perceived to have control of the situation (61, e.g., 57%). Most participants liked the experience ($min = 4$, $median = 7$ for the LIK dimension in Table 5).

3.3.4 Comparison

The classification accuracy reached best results with the Decision Tree in both the experiments. For the testing set improves from 84.41% in the first experiment to 91.47% in the second one (see Table 4). Also the other classifiers improved their accuracy in the second experiment. This is also shown by the ROC curve shown in Fig. 9, where there is in the second experiment an improvement of the curve areas related to the classification with respect to the first experiment ones.

Concerning the affective analysis, the boxplot in Fig. 10 compares the distributions of VAL, ARS, DOM and LIK in both the experiments. As shown in Fig. 11a, Valence is concentrate among the intermediate values, the best value is reached for valence =5, which is the median value. Dominance increase, this means that the person controls better the situation. Higher arousal values also are positive: the individual is less agitated, calmer. The users liked to use the system in both the experiments ($median = 6$ in the first experiment and 7 in the second one).

To verify if there is no statistically significant difference between the participants' perceptions concerning VAL, ARS, DOM and LIK in both the experiments we used the Mann-Whitney U test [23], which does not require the assumption of normal distributions.

The statistical analysis presented in Table 6 reports the p-values of the statistical test performed (in bold the significant p-value results) and the effect size, when applicable. In particular, the Mann-Whitney U test highlighted that there exists a significant difference ($p - value < 0.05$) for VAL, ARS and DOM in the two experiments, while there is no significant difference for LIK. We also evaluated the magnitude of performance difference between the two experiments. To this aim, we used the Cliff's Delta effect size (or d) [15]. The effect size is small for $d < 0.33$, medium for $0.33 \leq d \leq 0.474$ and large for $d > 0.474$. The magnitude of the performance difference is medium for VAL and small for ARS and DOM. So we can say that the new version of the system does not cause neither very

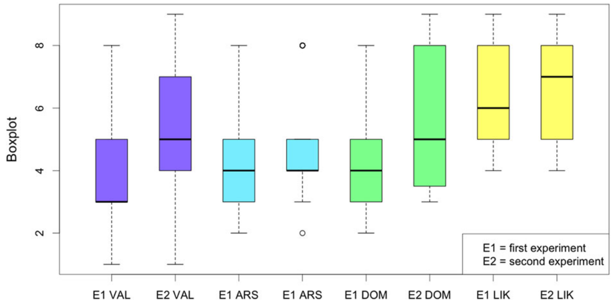


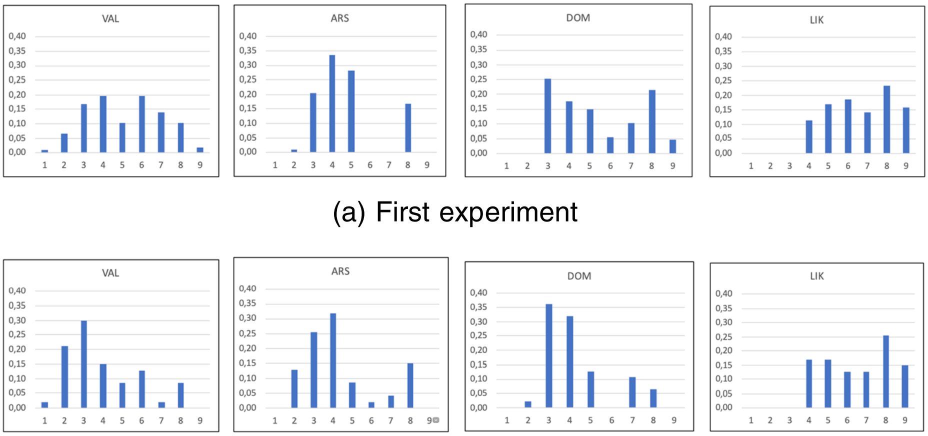
Fig. 10 Boxplots of the emotional variables

unpleasant and nor very pleasant sensations (corresponding to lower and higher valence values, respectively), but maintains the user in an intermediate state. It also improves the arousal and dominance variables. This is relevant because arousal mainly may interfere on the anxiety level of the person related to the system perception and a greater dominance provides the user with the awareness of mastering the situation.

4 Threats to validity

In this section, threats that could affect results and their generalization are presented and discussed to better comprehend strengths and limitations of our experiment. Despite our effort to mitigate as many threats as possible, some threats are unavoidable. We discuss the threats to validity following the guidelines proposed by Wohlin et al. [36].

Conclusion validity. We do not violate the assumptions of statistical tests because we chose tests which do not require normally distributed data (e.g., Mann-Whitney U test



(a) First experiment

(b) Second experiment

Fig. 11 Frequencies for the affective reactions related to the first (Fig. 11.a) and second (Fig. 11.b) experiment

Table 6 Affective reaction analysis

Variable	p-value	Effect size
VAL	**0.001**	−0.346 (medium)
ARS	**0.042**	−0.2 (small)
DOM	**0.002**	−0.308 (small)
LIK	0.514	NA

and Cliff's Delta effect size). We maintain the implementation of the treatment similar for the different participants, with the same light and chair, without external noise.

Internal validity. We reduced the maturation threats by proposing an experiment with a short duration. Participants were volunteers. They may be more motivated than traditional users (selection threats). This threats does not affect the results of the system when used in a real context because it is adopted for recognizing emotions through sensors, without the user intervention.

Construct validity. This threat concerns the design of the experiment. A threat may be to have reduced the number of emotions to only four, but on the other side, a higher number of emotions could have reduced the classification accuracy. Another threat may be to consider as user's emotion the input she/he provides by pressing the button. An alternative could have been to use the SAM questionnaire for evaluating also the user's emotion after seeing each image, but this does not produce specific emotions to provide as input to the classifier. Another threat may be due to the user involvement for pressing the button that could produce anxiety. The user expressed her perception by pressing the button during the stimuli acquisition, whose duration was 10 seconds for each image. So there is no hurry in performing this activity.

External validity. Concerning the selected people, they do not belong to a specific target of population. 30% were students, the others of various ages and employment. This should reduce the *threats of selection and treatment*. The used experimental objects are selected by a public dataset and this should avoid threats related to *interaction of setting and treatment*.

5 Conclusion

In this paper we presented a user-centered development process of an emotion detection system based on low-cost biometric sensors and Artificial Intelligence, which analyzes the emotional perception of the user on the detection system. Emotional perceptions may be useful to evaluate the interference of the system with the user affective state. We presented a case study in which two iterations of the process have been performed by using different Machine Learning classifiers. The second iteration reduced the weight and the dimension of the system hardware.

Second experiment results revealed that the second system release improves both the accuracy and the affective reactions of the users with respect to the heavier and more cumbersome previous one. This preliminary evaluation revealed that the user affective reactions are still influenced by the system, i.e., the user still perceived anxiety, even if in a more reduced form. We can try to further miniaturize the device. Another factor that can affect the user perceptions is the interaction modality. We chose external buttons, and when asked the opinion of the experts in the external stimuli definition phase, they considered the buttons appropriate. Nevertheless, this aspect needs further investigation: better results may be obtained by taking into account affordances [16] to evaluate and improve the design of the interaction modality in the external stimuli definition phase. In particular, different alternatives may be considered, such as different physical buttons, with different aesthetics, dimensions and labels, or other interaction modalities, such as point-and-click, touch-based, gesture-based and vocal interfaces.

In the future, we plan to try to improve the methodology by working in single emotion and modifying the invasiveness of the setting and to train the system with different kinds of inputs, such as videos or particularly realistic stimuli and on a wider sample of users.

Funding Open access funding provided by Universit degli Studi di Salerno within the CRUI-CARE Agreement.

References

1. Agrafioti F, Hatzinakos D, Anderson AK (2011) ECG Pattern analysis for emotion detection. IEEE Trans Affect Comput 3(1):102–115
2. Alonso-Martin F, Malfaz M, Sequeira J, Gorostiza J, Salichs M (2013) A multimodal emotion detection system during human–robot interaction. Sensors 13(11):15549–15581
3. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 30(7):1145–1159
4. Bradley MM, Lang PJ (1994) Measuring emotion: the self–assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry 25(1):49–59
5. Bradley MM, Lang PJ (2007) Emotion and motivation. Handbook of Psychophysiology
6. Chanel G, Rebetez C, Bétrancourt M., Pun T (2011) Emotion assessment from physiological signals for adaptation of game difficulty. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 41:1052–1063. 12
7. Chao L, Tao J, Yang M, Li Y, Wen Z (2015) Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In: Proceedings of the 5th international workshop on audio/visual emotion challenge, pp 65–72
8. Correa JAM, Abadi MK, Sebe N, Patras I (2018) AMIGOS: A Dataset for affect, personality and mood research on individuals and groups IEEE Transactions on Affective Computing
9. Di Bitonto P, Laterza M, Roselli T, Rossano V (2010) An evaluation method for multi-agent systems. In: KES International symposium on agent and multi-agent systems: Technologies and applications, Springer, pp 32–41
10. Ekman P, Friesen WV, Ellsworth P (2013) *Emotion in the human face: Guidelines for research and an integration of findings*, vol 11, Elsevier
11. Fritz T, Begel A, Müller SC, Yigit-Elliott S, Züger M (2014) Using psycho–physiological measures to assess task difficulty in software development. In: Proceedings of the 36th international conference on software engineering (ICSE), ACM, pp 402–413
12. Geisser S (1993) *Predictive* inference: An introduction. Chapman and hall/CRC Monographs on Statistics and Applied Probability Series Chapman and Hall
13. Girardi D, Lanubile F, Novielli N (2017) Emotion detection using noninvasive low cost sensors. In: Proceedings of the 7th international conference on affective computing and intelligent interaction (ACII), pp 125–130

14. Gong P, Ma HT, Wang Y (2016) Emotion recognition based on the multiple physiological signals. In: Proceedings of the IEEE international conference on real–time computing and robotics (RCAR), IEEE, pp 140–143
15. Grissom RJ, Kim JJ (2005) *Effect sizes for research: A broad practical approach* Taylor & Francis Group
16. Hartson R (2003) Cognitive, physical, sensory, and functional affordances in interaction design. Behaviour & Information Technology 22(5):315–338
17. Khan AM, Lawo M (2016) Wearable recognition system for emotional states using physiological devices. eTELEMED 2016:131–137
18. Koelstra S, Muhl C, Soleymani M, Lee J-S, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I (2011) DEAP: A database for emotion analysis using physiological signals. IEEE Trans Affect Comput 3(1):18–31
19. Kurdi B, Lozano S, Banaji MR (2017) Introducing the open affective standardized image set (OASIS). Behav Res Methods 49(2):457–470
20. Lang PJ, Bradley MM, Cuthbert BN et al (1999) International affective picture system (IAPS): Instruction manual and affective ratings. The center for research in psychophysiology University of Florida
21. Machajdik J, Hanbury A (2010) Affective image classification using features inspired by psychology and art theory. In: Proceedings of the 18th ACM International Conference on Multimedia, ACM, pp 83–92
22. Majid Mehmood R, Du R, Lee HJ (2017) Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors. IEEE Access 5:14797–14806
23. Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, pp 50–60
24. Myroniv B, Wu C-W, Ren Y, Christian A, Bajo E, Tseng Y-C (2017) Analyzing user emotions via physiology signals. Data Science and Pattern Recognition 1(2):11–25
25. Pellecchia MT, Frasca M, Auriemma Citarella A, Risi M, Francese R, Tortora G, De Marco F (2019) Identifying correlations among biomedical data through information retrieval techniques. In: 2019 23Rd international conference information visualisation (IV), pp 269–274
26. Plutchik R (1984) Emotions: a general psychoevolutionary theory. Approaches to Emotion 1984:197–219
27. Pollreisz D, TaheriNejad N (2017) A simple algorithm for emotion recognition, using physiological signals of a smart watch. In: Proceedings of the 39th international conference of the ieee engineering in medicine and biology society (EMBC), IEEE, pp 2353–2356
28. Rattanadoung K, Champrasert P, Aramkul S (2018) The emotional state classification using physiological signal interpretation framework. In: Proceedings of the international conference on signals and systems (ICSigSys), IEEE, pp 79–85
29. Read JP, Haas AL, Radomski S, Wickham RE, Borish SE (2016) Identification of hazardous drinking with the young adult alcohol consequences questionnaire: Relative operating characteristics as a function of gender. Psychol Assess 28(10):1276
30. Russell JA (2003) Core affect and the psychological construction of emotion. Psychol Rev 110(1):145
31. Santamaria-Granados L, Munoz-Organero M, Ramirez-Gonzalez G, Abdulhay E, Arunkumar N (2018) Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS). IEEE Access 7:57–67
32. Shu L, Yu Y, Chen W, Hua H, Li Q, Jin J, Xu X (2020) Wearable emotion recognition using heart rate data from a smart bracelet. Sensors 20(3):718
33. Stevenson RA, James TW (2008) Affective auditory stimuli: Characterization of the international affective digitized sounds (IADS) by discrete emotional categories. Behavior Research Methods 40(1):315–321
34. Sullivan GM, Artino AR (2013) Analyzing and interpreting data from likert–type scales. Journal of Graduate Medical Education 5(4):541–2
35. Valenza G, Scilingo EP (2013) *Autonomic* nervous system dynamics for mood and emotional-state recognition: Significant advances in data acquisition, signal processing and classification. Springer Science & Business Media
36. Wohlin C, Runeson P, Hst M, Ohlsson MC, Regnell B, Wessln A (2012) Experimentation in software engineering. Springer Publishing Company Incorporated

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.