Virtual Reality (2024) 28:107
https://doi.org/10.1007/s10055-024-00989-y

S.I. : NEW TRENDS ON IMMERSIVE HEALTHCARE

Multimodal emotion classification using machine learning in immersive and non-immersive virtual reality

Rodrigo Lima[1,2,3] · Alice Chirico[4] · Rui Varandas[6,7] · Hugo Gamboa[6,7] · Andrea Gaggioli[4,5] · Sergi Bermúdez i Badia[1,2,3]

Received: 30 June 2022 / Accepted: 20 March 2024 / Published online: 6 May 2024

Abstract

Affective computing has been widely used to detect and recognize emotional states. The main goal of this study was to detect emotional states using machine learning algorithms automatically. The experimental procedure involved eliciting emotional states using film clips in an immersive and non-immersive virtual reality setup. The participants' physiological signals were recorded and analyzed to train machine learning models to recognize users' emotional states. Furthermore, two subjective ratings emotional scales were provided to rate each emotional film clip. Results showed no significant differences between presenting the stimuli in the two degrees of immersion. Regarding emotion classification, it emerged that for both physiological signals and subjective ratings, user-dependent models have a better performance when compared to user-independent models. We obtained an average accuracy of 69.29 ± 11.41% and 71.00 ± 7.95% for the subjective ratings and physiological signals, respectively. On the other hand, using user-independent models, the accuracy we obtained was 54.0 ± 17.2% and 24.9 ± 4.0%, respectively. We interpreted these data as the result of high inter-subject variability among participants, suggesting the need for user-dependent classification models. In future works, we intend to develop new classification algorithms and transfer them to real-time implementation. This will make it possible to adapt to a virtual reality environment in real-time, according to the user's emotional state.

Keywords Affective computing · Emotions · Wearables · Physiological signals · Machine learning · Virtual reality

Alice Chirico and Rui Varandas have contributed equally to this work.

✉ Rodrigo Lima
rodrigo.lima@arditi.pt

Alice Chirico
alice.chirico@unicatt.it

Rui Varandas
r.varandas@campus.fct.unl.pt

Hugo Gamboa
hgamboa@fct.unl.pt

Andrea Gaggioli
andrea.gaggioli@unicatt.it

Sergi Bermúdez i Badia
sergi.bermudez@uma.pt

[1] Faculdade de Ciências Exatas e Engenharia, Universidade da Madeira, Campus Universitário da Penteada, 9020-105 Funchal, Portugal

[2] NOVA Laboratory for Computer Science and Informatics, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Setúbal, Portugal

[3] ARDITI - Agência Regional para o Desenvolvimento da Investigação, Tecnologia e Inovação, Caminho da Penteada, 9020-105 Funchal, Portugal

[4] Dipartimento di Psicologia, Università Cattolica del Sacro Cuore, Largo Agostino Gemelli 1, 20123 Milan, Italy

[5] Applied Technology for Neuro-Psychology Lab, I.R.C.C.S, Istituto Auxologico Italiano, Milan 20149, Italy

[6] LIBPhys (Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Setúbal, Portugal

[7] PLUX Wireless Biosignals S.A., Avenida 5 de Outubro 70, 1050-059 Lisboa, Portugal

1 Introduction

Affective computing (AC) is the computing that relates or influences emotions and focuses on understanding the psychophysiological mechanisms underlying the way humans recognize emotions (Bota et al. 2019). Traditional methodologies to recognize emotions in AC usually rely on facial expressions and speech. However, these methods lack feasibility during in-field experiments (Chanel et al. 2007).

Physiological signals are an alternative method to assess emotions, using a wide variety of wearable sensors in an unobtrusive approach (Bota et al. 2019).

AC has been used to recognize emotional states through physiological signals, and it has been applied to patients with cognitive disorders and depression to improve their mental state.

Research in AC has been mostly performed by presenting emotional stimuli in 2D environments. However, virtual reality (VR) is becoming more popular in the field of emotion recognition (Marín-Morales et al. 2020). VR is used to study human emotions in controlled conditions that simulate real-life situations and environments, providing high levels of presence, which relates to the feeling of being in the virtual environment (Rose et al. 2018). Previous studies have demonstrated the effectiveness of VR in eliciting specific emotional states in controlled settings by comparing the emotional response between virtual and real environments while assessing the physiological response through physiological signals (Marín-Morales et al. 2020).

Therefore, the goal of this study was to automatically detect and recognize emotional states using machine learning and VR. To achieve our goal, the following objectives were proposed: (1) to evaluate the effect of the degree of immersion in the presentation of the emotional stimuli; (2) to classify emotional stimuli from self-reported ratings and physiological signals using user-dependent and user-independent classification models. Thus, we formulated the following research questions: (1) How does the degree of immersion of the emotional stimuli affect the emotional response?; (2) How accurate is it possible to classify the emotional stimuli using machine learning from physiological signals and self-reported ratings?

The developed experimental procedure consists of presenting emotional film clips (using the EMDB database (Carvalho et al. 2012)) in two degrees of immersion and monitoring users' physiological signals. The participants' emotional states were monitored by measuring the following signals: electrocardiogram (ECG), electrodermal activity (EDA), electroencephalogram (EEG), skin temperature (SKT), respiration (RESP), and functional near-infrared Spectroscopy (fNIRS). Then, we applied specific signal processing techniques to extract a combination of features from these signals. Additionally, the participants reported subjective ratings for each presented film clip using the self-assessment manikin (SAM) (Bradley and Lang 1994) and an ad-hoc self-reported appraisal instrument developed for this study (Scherer et al. 2001; Smith 1989; Reisenzein et al. 2019; Aue et al. 2007; Denson et al. 2009).

Machine learning algorithms and models were applied to the features extracted from the physiological signals to classify the emotional category of the film clip accurately. The classification models were applied separately for the subjective ratings and the physiological signals, using user-dependent and user-independent models to detect the emotional category of the corresponding film clip.

This paper is divided into the following sections: Sect. 2 presents the state of the art; Sect. 3 describes the software, hardware, experimental procedure, emotional database, physiological signals, and machine learning algorithms used in the study; Sect. 4 shows the results for the two degrees of immersion and the classification results for the subjective ratings and physiological signals; Sects. 5 and 7 discuss the results obtained and present conclusion, as well as future work.

2 State of the art

2.1 Virtual reality

Virtual reality (VR) has been used extensively in several rehabilitation areas, from motor to cognitive rehabilitation. It is an alternative to conventional rehabilitation therapies, with potential limitations, such as the transfer to everyday activities. Thus, VR can be a means to deliver personalized therapy, as it can induce specific emotions, leading to a better prognosis (Stasieńko and Sarzyńska-Długosz 2016).

In this work, we used head-mount displays as the presentation modality, where users wear special glasses or helmets on the head (head-mounted displays (HMD)) (Trojan et al. 2014). Juvrud et al. (2018) developed the Virtual Lab, a large-scale in-home testing that combines physiological measures and VR. Meehan et al. (2002) studied the influence of the level of presence on the physiological response to a stressful environment.

Another important concept of VR is the degree of immersion. Immersion is the extent to which the VR system delivers sensations from the real world to a virtual world (Rose et al. 2018). Previous studies (Borrego et al. 2016; Goncalves et al. 2021) compared the sense of immersion between an HMD and 3D space projection of 2D information, the CAVE system. They concluded that participants reported a higher level of immersion in the HMD system compared to the CAVE.

Thus, previous studies support the use of VR to study the effect of immersion in physiological responses since higher levels of immersion reported using the HMD have been shown to elicit higher physiological responses (Meehan et al. 2002). Moreover, it would be interesting to evaluate how the different degrees of immersion in the presentation mode of the emotional stimuli, i.e., the screen size and angle, would affect the emotional response of the participants.

2.2 Emotion regulation

Emotion regulation can be defined as "the process of modulating one or more aspects of an emotional experience or response" (Rey et al. 2014). Individuals differ in how they use emotion regulation strategies, as deficits in this regulation affect mental health (Inwood and Ferrari 2018). Therefore, it is an important factor to maintain social relationships and adapt to new activities of daily living (Rey et al. 2014).

It is a major research topic in several mental health disorders and psychotherapy research, such as borderline personality disorder (BPD), depression, anxiety disorders, eating disorders, and somatoform disorders (Berking and Wupperman 2012). People with BPD have less emotional awareness and clarity, as well as the tendency to use self-injury as an emotion regulation strategy. Individuals with depression reported difficulties in identifying and accepting negative emotions, and people with anxiety disorders have a higher reaction to fearful situations (Berking and Wupperman 2012).

Studies usually have different approaches in terms of conceptualization and assessment of emotion regulation, thus creating limitations to comparing the results across studies (Berking and Wupperman 2012). Primary to emotion regulation, it is necessary to detect emotions properly. Emotion detection is often performed by detecting facial expressions, speech, gestures, and physiological states (Garcia-Garcia et al. 2017). Facial expressions reflect the emotions one person is feeling, by detecting a face, identifying points in the face that reveal emotions, and processing those points to decide which emotion is being detected (Garcia-Garcia et al. 2017). Speech is another method to detect emotions by gathering emotional information from the user's voice, through the information generated to the link associated with paralinguistic information (Garcia-Garcia et al. 2017). Body gestures also reveal what a person is feeling, although it is still not clear how to create systems to detect body gestures to recognize emotions. Finally, emotions can be detected by monitoring physiological states and creating a classification system to predict the response of our body associated with emotions (Garcia-Garcia et al. 2017).

Most of these approaches have the limitation that facial expressions, speech, and gestures can be controlled to hide a specific emotional state (Chen et al. 2020). To overcome this problem, it is relevant to classify and distinguish between different emotion dimensions, using standardized self-report scales and physiological signals to identify different emotional states.

2.3 Emotional models

AC is a growing field due to its importance in real-world applications by recognizing and interpreting emotions (Sharma and Mathew 2020). Throughout the years, many emotional models appeared to categorize emotional states. These can be defined according to two dimensions: a discrete dimension and a continuous dimension (Scherer et al. 2001). In the discrete emotional dimension, emotions are categorized into specific states, such as fear and anger (Cacioppo and Tassinary 1990). Ekman (1992) stated that emotions are universal and shared between different cultures. Thus, people can easily recognize expressions in people from different cultures. From these facial expressions, Ekman enumerated the following basic emotions: happiness, sadness, anger, disgust, fear, and surprise. Plutchik (1982) created the wheel of emotions to classify eight basic emotions: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust. This wheel is composed of low-intensity emotions on the outer edges of the wheel. Then as you move towards the center, the primary emotions appear until the center of the wheel is reached, holding the most intense emotions one can experience. In this model, every emotion has its own opposite emotion (e.g., sadness vs joy, anger vs fear, etc.). On the other hand, regarding the continuous emotional dimension, Russell (1980) developed the circumplex model of emotion (Fig. 1), a two-dimensional circular space where the vertical axis represents arousal, and the horizontal axis represents valence. In this model, an emotional state can be represented by any level of valence and arousal. Finally, Mehrabian and Russell (1974) added a third dimension—dominance. The Dominance scale was added to characterize control and dominance over one's emotions.

2.4 Emotional induction datasets

Inducing emotional states in experimental settings has been widely studied in terms of psychological research, using methods like hypnosis, mental imagery, music and facial expressions. These methods are usually based on deception and cannot be standardized (Uhrig et al. 2016).

An alternative method to induce specific emotional states that can be easily standardized and do not involve using deception is using pictures and films to generate these emotional states.

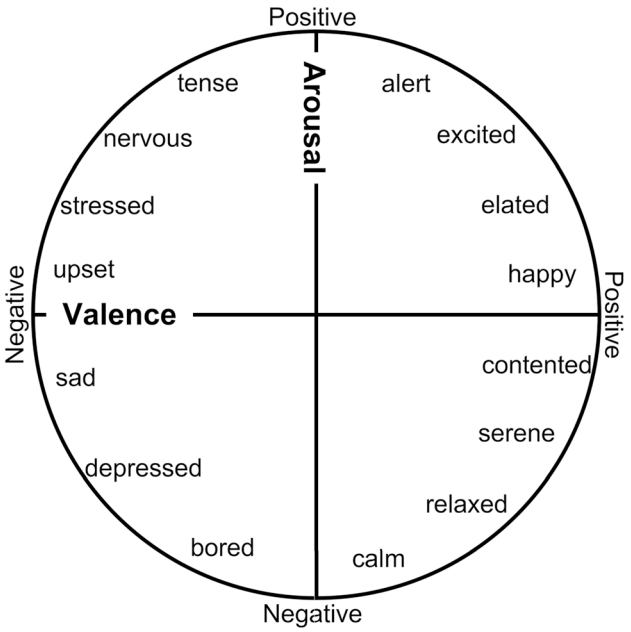

Fig. 1 Circumplex model of affect. Based on:Liapis and Xenos 2013

The international affective picture system *(IAPS)* (Lang et al. 2005) is a standardized dataset of emotional pictures divided into several emotional categories, shown for 6 s. It was developed to assess emotional states using the dimensional approach of emotions to classify them into valence, arousal, and dominance, using the Self-Assessment Manikin (SAM) scale (Bradley and Lang 1994).

The Geneva affective picture database *(GAPED)* (Dan-Glauser and Scherer 2011) is a new database composed of 730 pictures, created to increase the number of available emotional stimuli presented by the IAPS. Its content is divided into four specific negative contents (spiders, snakes, and scenes that induce emotions related to the violation of moral and legal norms), positive (humans and animal babies) and neutral pictures (inanimate objects). The rating of the images was performed according to the arousal, valence, and congruence with internal and external norms.

The *ArtEmis* dataset (Achlioptas et al. 2021) focuses on the affective experience triggered by visual artworks. It contains thousands of human emotional attributions and explanations regarding paintings from WikiArt. The emotional classification is subjective and personal, accompanied by a caption for each picture. Although the subjectivity of the responses, namely the captions, this dataset uses the following eight categorical emotional states: anger, fear, disgust, sadness, amusement, awe, contentment, and excitement.

The *FilmStim* database is a set of emotional film excerpts developed by Schaefer et al. (2010). The film excerpts were rated in multiple dimensions, and each film clip lasted 1–7 min. The films were rated in terms of emotional discreteness using self-reported levels of arousal, positive and negative affect (PANAS), positive and negative affect scores from the differential emotions scale (DES) and six emotional scores for anger, disgust, sadness, fear, amusement, and tenderness.

The database for emotion analysis using physiological signals *(DEAP)* (Koelstra et al. 2012) is a multimodal dataset composed of 40 music videos with a 1-min duration. These videos were rated in arousal, valence, dominance, like/dislike, and familiarity. This database also recorded multiple physiological signals (EEG, EDA, BVP, RESP, SKT, EOG) and the face recording video of the participants.

The emotional movie database *(EMDB)* (Carvalho et al. 2012) is a validated database composed of 52 film clips without auditory content, divided into the following categories: erotic, horror, social positive, social negative, scenery and objects. The film clips were rated in arousal, valence, and dominance, and the physiological signals (HR and EDA) were collected.

The *MAHNOB-HCI* (Soleymani et al. 2012) is a multimodal database composed of 20 emotional film clips with a duration of 35–117 s, rated in terms of arousal, valence, dominance, emotional label (neutral, anxiety, amusement, sadness, joy, disgust, anger, surprise, and fear), and predictability. It also contains recording from facial video, audio, eye gaze, and multiple physiological signals (EDA, ECG, RESP, SKT, EEG).

Several works used pictures and film clips to elicit emotional states. Both these methods have the ability to elicit emotions in most subjects, although films have recently been used to elicit emotions instead of pictures since films are more dynamic, ecologically valid and more similar to real-life experiments when compared to static images (Uhrig et al. 2016).

2.5 Physiological signals

Physiological signals can be objective sources of information about emotions and may be produced by the activity of the autonomous nervous system (ANS). Since natural bodily reactions might be harder to control, physiological signals are more adequate for emotional recognition (Wioleta 2013). For example, changes in the heart rate, respiration rate, and blood pressure are responses to the activation of the sympathetic nerves of the ANS. Hence, when someone is positively or negatively excited, such changes are inevitable and detectable by physiological signals (Jerritta et al. 2011). Namely, some emotions need more than one physiological signal to be properly explained. Thus, a multi-modal approach needs to be taken into account (Egger et al. 2019).

Electrocardiography *(ECG)* measures the heart's electrical activity, resulting in the contraction and relaxation of the cardiac muscle, which allows obtaining information regarding the Heart Rate (HR) and Heart Rate Variability (HRV). HR is directly related to the activity of the ANS, as the parasympathetic nervous system (PNS) decreases HR, and the sympathetic nervous system (SNS) increases HR. This relationship between the HR and the ANS may inform about different emotions, as HRV is a useful measure of stress and mental effort (Jerritta et al. 2011).

The human skin is innervated by numerous efferent fibres, including sympathetic fibres. The eccrine sweat glands in the skin produce sweat that changes the skin's electrical properties, thus making it possible to measure the Electrodermal Activity *(EDA)*. EDA directly reflects the activity of the SNS because there is no innervation of the sweat glands by the PNS (Boucsein 2012). Thus it enables us to infer the level of arousal and the activation of the SNS (Bota et al. 2019).

Electromyography *(EMG)* measures the muscles' electrical activity where there is an increase of amplitude in the EMG signal during muscle contraction (Bota et al. 2019). Recognizing emotions through facial expressions can be achieved by placing EMG electrodes on the face to measure the activity of the muscles, e.g., the Zygomaticus major and Corrugator supercilii muscles, as the muscle activity increases during emotions with negative valence (Gouizi et al. 2011).

Respiration *(RESP)* signals measures thoracic displacement to monitor the respiration pattern. A respiration cycle is composed of an inhalation period followed by an exhalation period and can be monitored to measure how deep and fast a person breathes (Schmidt et al. 2019). For example, slower respiration frequencies are associated with relaxed or depressive states, while faster frequencies may indicate happiness or anger (Egger et al. 2019).

Skin temperature *(SKT)* is related to the vasodilatation of peripheral blood vessels induced by increased activity of the SNS (Gouizi et al. 2011). This sensor can assess if someone is relaxed (SKT is higher) or stressed (SKT is lower).

The limbic system is responsible for the emotional response. This system comprises deep brain structures such as the amygdala and the hypothalamus (Lindquist et al. 2015). Although it is not possible to measure deep brain signals through electroencephalography and functional near-infrared spectroscopy, these two sensors allow us to detect brain activity at the scalp.

Electroencephalography *(EEG)* measures the sum of the electrical currents incoming from neurons' excitation in the cerebral cortex during their activation (Bota et al. 2019). These signals are a direct response to the activity of the central nervous system (CNS) (Jerritta et al. 2011). Despite being a highly time-consuming assessment to perform (Egger et al. 2019), many EEG studies have been performed throughout the years to study the relationship between emotions, brain regions, and EEG frequency bands. Previous studies have shown that emotions are related to increased brain activity in the frontal, parietal and central regions of the brain (Crawford et al. 1996). For example, Hu et al. (2017) showed that high arousal stimuli decreased the alpha power in the parietal-occipital region, while high valence stimuli exhibited high alpha power in the frontal region.

Functional near-infrared spectroscopy *(fNIRS)* measures the concentration changes of oxygenated haemoglobin (HbO) and deoxygenated haemoglobin (HbR) through blood flow changes and brain activations caused by neuron firings (Naseer and Hong 2015). The brain's prefrontal region is known to be involved in emotion processing. Thus fNIRS provides information regarding valence and arousal (Moghimi et al. 2012).

2.6 Emotion detection

Over the last years, several research studies have been performed on the topic of emotion detection and recognition using physiological signals (Jerritta et al. 2011).

Most of these works focused on a user-dependent approach to recognising emotions, which highly depends on each subject, and more recently on user-independent approaches, with unknown data from different users being used as testing data, to obtain the accuracy of the models (Jerritta et al. 2011).

These works also vary in terms of the emotional elicitation stimuli used, as well as the physiological signals acquired to recognize emotional states. Thus, in this section, a review of the previous works will be described from a user-dependent and user-independent point of view.

Kim and André (2008) acquired four physiological signals (EMG, ECG, EDA and RESP) to classify musical emotions (positive/high arousal, negative/high arousal, negative/low arousal, and positive/low arousal) using an extended linear discriminant analysis (pLDA), without any laboratory setting. The recognition accuracy obtained was 95% for user-dependent classification and 70% for user-independent classification. Liu et al. (2005) measured ECG and EMG to classify the following emotional states: anxiety, boredom, engagement, frustration and anger. These emotional states were elicited using computer-based cognitive tasks, like anagrams and pong tasks. The average correct classification across all participants (user-dependent) and all affective states was 75.12% with K-nearest neighbor, 83.5% with regression trees, and 74.03% with Bayesian networks.

Several studies used the IAPS database to elicit emotional states. Rigas et al. (2007) classified happiness, disgust and fear using EMG, ECG, EDA and RESP. The accuracy obtained was 62.70% with K-nearest neighbor and 62.41% with random forest, both for user-independent models. Haag et al. (2004) used a combination of multiple physiological sensors to measure valence and arousal. Using a neural network classifier, the classification accuracy for valence was 89.7% and 63.76% for arousal, with user-dependent models. Maaoui and Pruski (2010) used blood volume pulse (BVP), EMG, SKT, EDA and RESP sensors to classify amusement, contentment, disgust, fear, sad and Neutral. They obtained an accuracy of 90% with support vector machine and 92% with fisher linear discriminant analysis, both from user-dependent models. Finally, Gu et al. (2010) used ECG, BVP, EDA, EMG and RESP to classify several discrete emotions using the K-nearest neighbor classifier. The accuracy was 50% for user-independent and 90.7% for user-dependent models.

On the other hand, film clips have recently been used to elicit emotions instead of pictures. Li and Chen (2006) acquired ECG, SKT, EDA and RESP while eliciting the emotional states of Fear, Joy and Neutral by using one film clip for each of these emotional states. Then, a Canonical Correlation Analysis was performed to obtain an accuracy of 85.3% from a user-dependent point of view. Nasoz et al. (2004) also used film clips to elicit several discrete emotions combined with EDA and heart rate. The accuracy obtained was 71% with K-Nearest Neighbor, 74% with discriminant function analysis and 83% with marquardt back propagation, all user-dependent. Ramzan et al. (2016) also used audio-video film clips to classify arousal and valence. A combination of ECG and EEG signals were used to classify valence and arousal with an average accuracy of 71.6% and 54.0%, respectively.

From these studies, we can conclude that a combination of physiological signals leads to a higher classification accuracy of emotional states. Furthermore, from the comparison of the accuracy using user-dependent and user-independent models, we can obtain higher accuracy from user-dependent point of view.

3 Materials and methodology

3.1 Sample

Participants were recruited from a convenience sample of volunteer subjects, all university students and workers, fluent in English. Fourteen healthy adults (9 females, 5 males) volunteered to participate in this study, with an average age of 31 ± 7.53—from 21 to 50 years old. All volunteers were considered healthy, as none reported suffering from any psychological disorder or to take regular medication. Written informed consent was provided before collecting any type of data. The study and all the ethical procedures were approved by the ethics committee of the University of Madeira.

3.2 Materials

3.2.1 Hardware

The hardware used for this study comprises the following elements: An HTC Vive HMD, a Desktop, a Laptop, an EEG headset and a wearable device.

The HTC Vive is one of the main HMDs, with a 1080x1200 per eye pixel resolution and a 110° field of view (FOV). It provides head orientation and position tracking through the two base stations' laser tracking system.

The Vive was connected to a computer with Windows 10, an Intel i7-6700 3.40 GHz processor with 16 GB of RAM, and an NVIDIA Quadro P6000 graphics card.

The laptop was used to acquire all the physiological signals from the EEG headset and the wearable device. It was a laptop with windows 10, an Intel i7-1165G7 2.80 Ghz processor, 16 GB of RAM, and an Intel Iris Xe graphics card.

The wearable device was a biosignalsplux 8-channel device developed by PLUX (PLUX Wireless Biosignals, S.A.). This device was developed for research purposes, allowing wireless and high-quality signal acquisition. It enables the acquisition and recording of eight physiological signals simultaneously, up to a sampling rate of 4000 Hz. For this study the following signals were acquired with a sampling rate of 1000 Hz: ECG, EDA, RESP, SKT and fNIRS. The fNIRS was acquired with a 24-bit resolution, while the remaining signals were acquired with a 16-bit resolution.

Finally, the EEG headset used was the g.Nautilus wearable EEG headset (g.tec medical engineering, GmbH Austria). This headset was designed for research applications and comes with flexible cables to configure the electrodes' position using gelled electrodes with 32 EEG channels.

3.2.2 Software

The emotional stimuli experiment and the emotional self-report scales were designed using the PsychoPy2 software (Peirce et al. 2019). PsychoPy is a free cross-platform that allows creating and running experiments in behavioural sciences, like neuroscience and psychology.

The Virtual Desktop software[1] was used to cast the desktop running the PsychoPy experiment to the HMD. It allows duplicating the desktop using different personalized options, such as the screen size, screen distance and screen curve. On

[1] https://www.vrdesktop.net/ [accessed February 15, 2022].

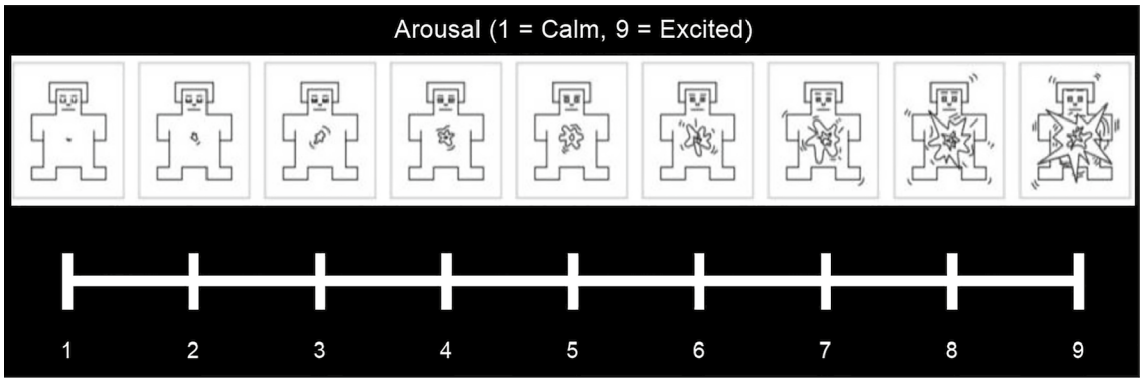

Fig. 2 Arousal scale (1 = calm, 9 = excited)

the immersive setup, the following values were setup for this software: Screen Size = 230°, Screen Distance = 1.50 m, Screen curve = 100%, while on the non-immersive setup, the values used were: screen size = 70°, screen distance = 1.50 m, screen Curve = 0%.

The acquisition of the physiological signals was performed using the OpenSignals software[2] and the OpenVibe software (Renard et al. 2010). The OpenSignals allows real-time visualization of the physiological signals acquired with all PLUX devices, like the biosignalsplux wearable device, while the OpenVIbe allows to acquire and visualize the EEG data.

All the signals acquired during this experiment from different devices were synchronized using the lab streaming layer software (Christian et al. 2019). This synchronization includes all the physiological signals from the biosignalsplux, the EEG acquired with OpenVibe and the emotional film clips onset and offset timestamps from the PsychoPy software.

These synchronized signals were saved in a.xdf file, and then only the segments of the signals corresponding to the visualization of a film clip (time between onset and offset) were processed using Python.

3.3 Emotional database

The emotional database used for this study was the EMDB film clips database, which is validated for Portuguese subjects, our target population (Carvalho et al. 2012). The EMDB database comprises the following emotional categories: erotic, horror, social positive, social negative, scenery and objects. For this experiment, the emotional category "Objects" was removed since there were only two film clips in this category, which are neutral videos in terms of valence and arousal. In comparison, there are ten film clips in each of the remaining categories, resulting in 50 film clips used for this experiment, each with an approximate duration of 40 s.

3.4 Emotional self-report scales

3.4.1 Self-assesssment manikin (SAM)

The self-assessment manikin (SAM) scale (Bradley and Lang 1994) was implemented using the PsychoPy software to provide a digital version of the original scale (see Figs. 2, 3 and 4). The subjective ratings' score for arousal, valence, and dominance was obtained for every film clip shown. All the scales were presented in a 9-point Likert Scale, with the arousal scale ranging from 1 = Calm to 9 = Excited (Fig. 2), the valence scale from 1 = Unpleasant to 9 = Pleasant (Fig. 3), and the dominance scale from 1 = Controlled to 9 = In Control (Fig. 4).

3.4.2 Ad-hoc appraisal scale

The ad-hoc self-reported appraisal scale was designed to study further emotional dimensions that could potentially add to the arousal, valence, and dominance dimensions of emotions.

Physiological signals related to emotional reactions are organized around the personal meaning attributed to the situation through an appraisal process (Scherer et al. 2001). An association between the response of physiological signals from the ANS and appraisals have been demonstrated repeatedly. Thus, the opposite idea can be studied, that physiological signals can be indicators of specific appraisals.

For this experiment, we presented the following appraisal items on a 7-point Likert scale (1 = Not at all to 7 = completely agree): anticipatory effort, perceived obstacles, novelty, unexpectedness, intrinsic goal, perceived control, and control/stress.

[2] https://www.pluxbiosignals.com/collections/opensignals [accessed March 03, 2022].

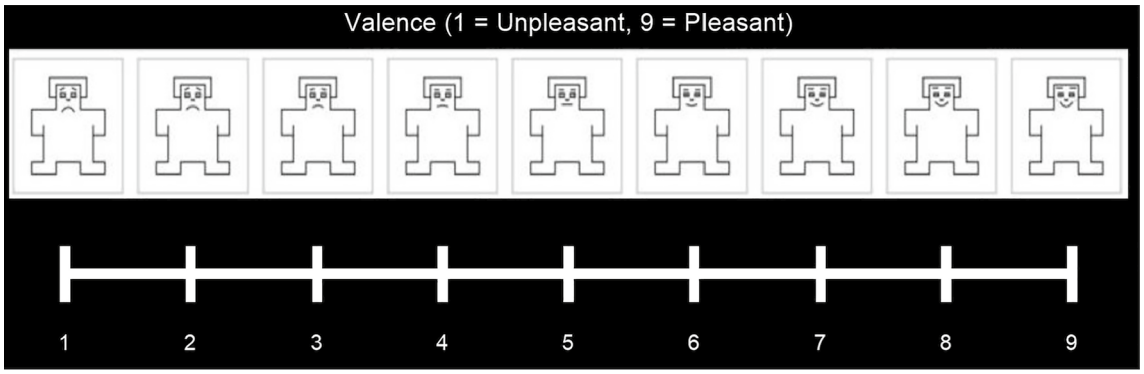

Fig. 3 Valence scale (1 = unpleasant, 9 = pleasant)

The anticipatory effort is related to the appraisal of a situation as objectively relaxing or requiring effort. Thus, the higher the effort, the higher the activation of the activity of the Corrugator muscle (EMG) (Smith 1989). The perceived obstacles appraisal is related to the presence vs absence of resistance between the person and their goal, leading to a higher heart rate (ECG) (Smith and Ellsworth 1985). The novelty and unexpectedness items are related to how new and unexpected the event is to the user, increasing the skin conductance levels (EDA), decreasing the heart rate (ECG) and pausing the respiration rhythm (Reisenzein et al. 2019). The intrinsic goal is the appraisal of a stimulus as a goal congruent or incongruent (the relevance of the stimulus to the user). It influences the skin conductance, the forehead and finger temperature (SKT) (Aue et al. 2007). Finally, the perceived control and control/stress is the extent to which one feels in power of an event that occurs since we can appraise situations as challenges, leading to an increase in heart rate and a decrease in the skin conductance levels (Denson et al. 2009). All the items presented to the users for the appraisal scale are shown in Figs. 5 and 6.

3.5 Experimental procedure

This experiment was divided into two sessions on consecutive days to present the EMDB database with the two degrees of immersion (immersive and non-immersive), one degree of immersion setup per session. Thus, the EMDB database was split in half—25 film clips (5 per category) and the order of the emotional categories were randomized for every session and participant. The film clip order within each category was also randomized; e.g., if the first selected category were "Horror", the user would be presented with five consecutive horror film clips in random order, and then the following category would be presented. Each session had an approximate duration of 1h30min due to the time spent configuring the gelled electrodes for the EEG wearable device.

Participants were randomly assigned to the two sets resulting from the split of the database and randomly assigned to the immersive and non-immersive setup set on the Virtual Desktop software.

In each session, participants were asked to rest in a chair as the physiological sensors were placed. Cardiac signals were acquired by placing three gelled electrodes on the participant's chest in the V2 configuration of Einthoven's triangle system. Two gelled electrodes were placed on the middle phalanxes of the second and third fingers of the participant's non-dominant hand to measure EDA. The RESP signal was acquired using a piezoelectric sensor that measures displacement variations caused by the volume changes in the abdomen or thorax during a respiration cycle. The wearable chest belt was placed on the thorax of the user. SKT was acquired with a sensor placed on the distal phalanx of the 4th finger of the non-dominant hand to measure the peripheral body temperature. The fNIRS sensor was placed approximately near the AF7-AF9 region in the 10–20 EEG system to measure the activity of the prefrontal cortex of the brain (blue mark in Fig. 7). The EEG configuration map is shown in red circles in Fig. 7. First, the impedance of all 32 electrodes was checked using the g.NEEDaccess software (g.tec medical engineering, GmbH Austria). Then, the HTC Vive headset strap was placed over the EEG cap after all the electrodes had been placed without moving the EEG cap (Fig. 8). Finally, the EEG signals were visually inspected to assess their quality.

Then, participants looked at a fixation cross at the centre of the screen for 3 min to record the baseline physiological signals. After that, a practice trial was performed to provide instructions on completing the subjective rating scales. Finally, the actual experiment starts by

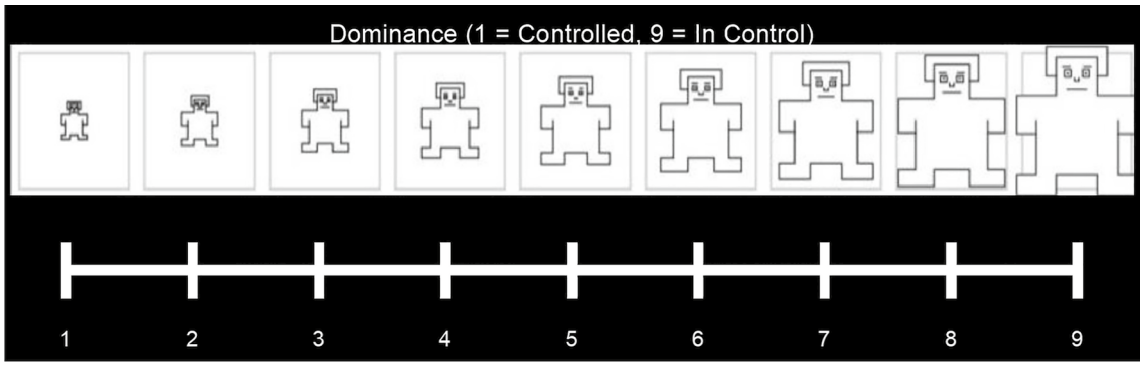

Fig. 4 Dominance scale (1 = controlled, 9 = in control)

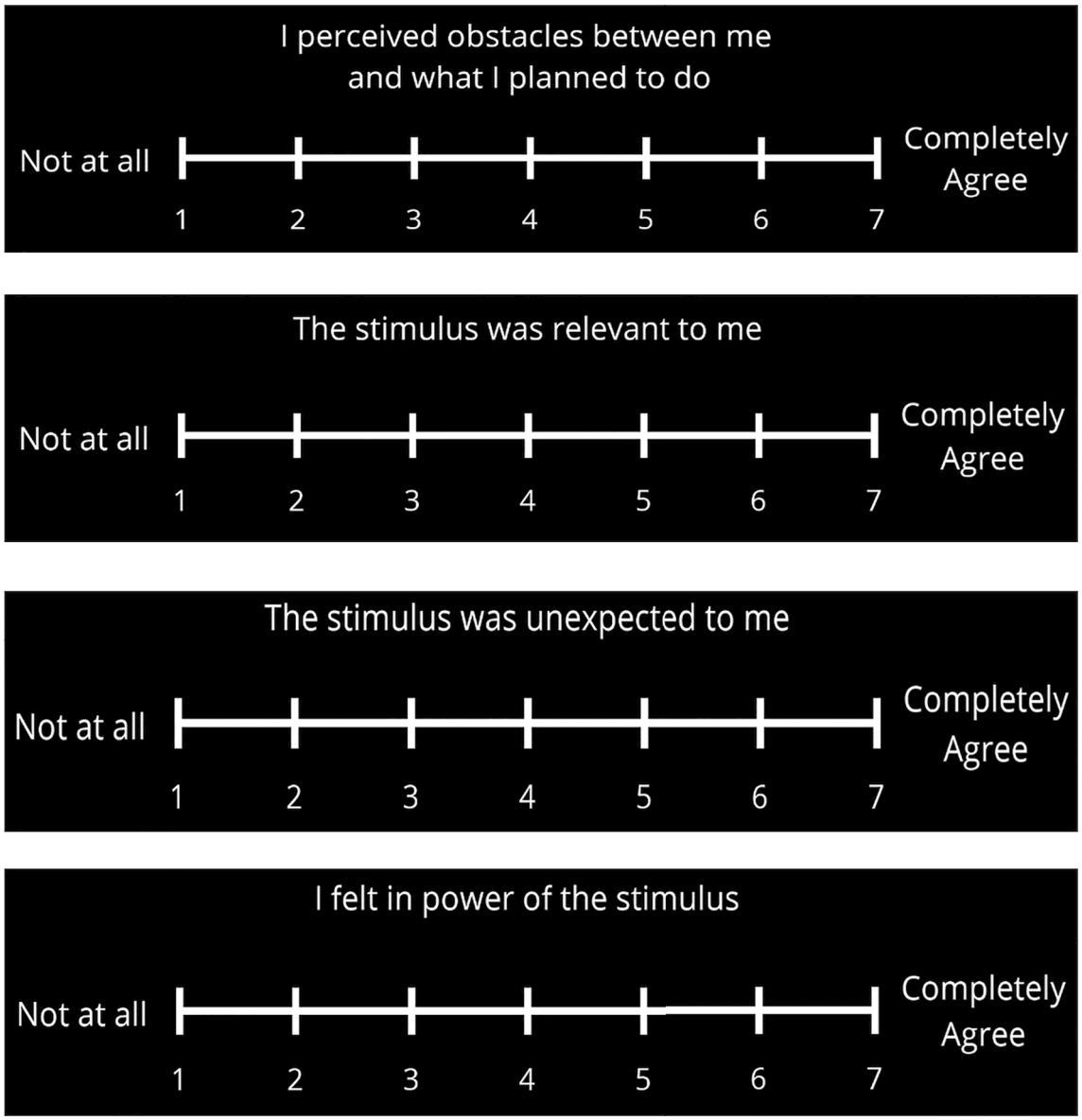

Fig. 5 Ad-hoc appraisal scale

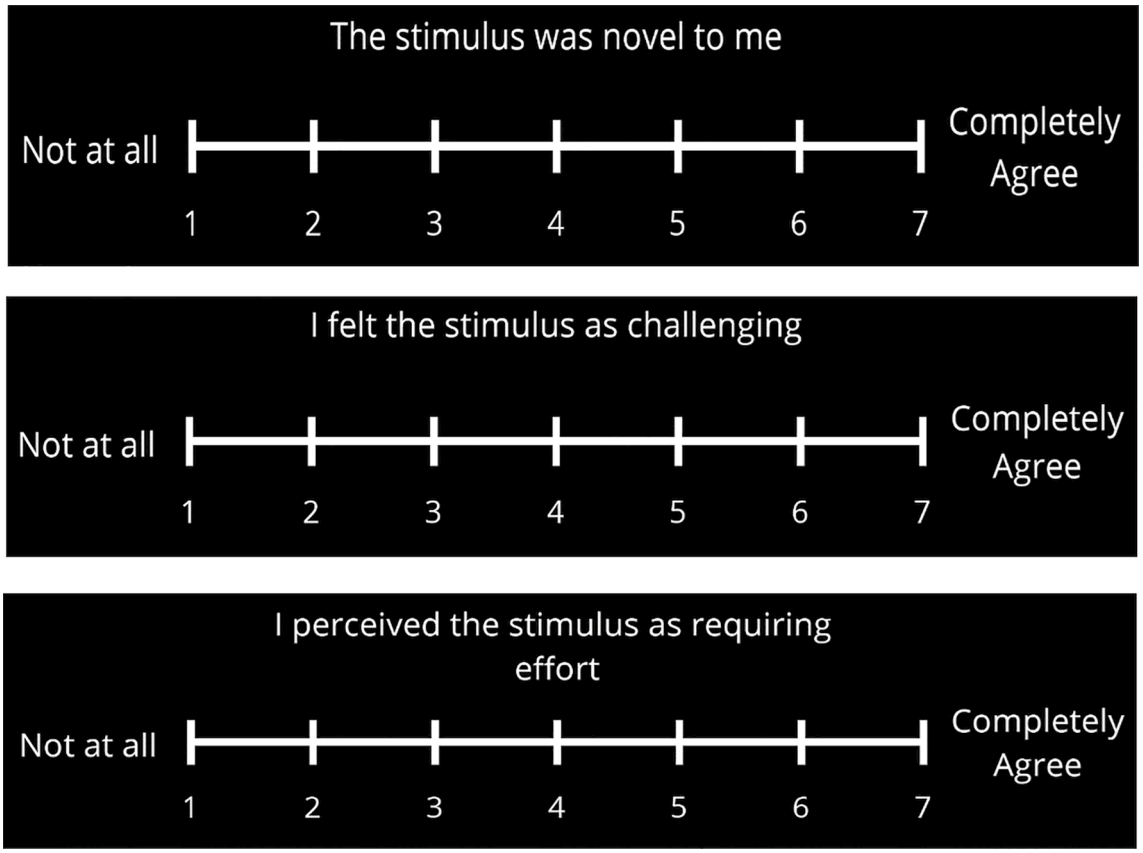

Fig. 6 Continuation of the ad-hoc appraisal scale

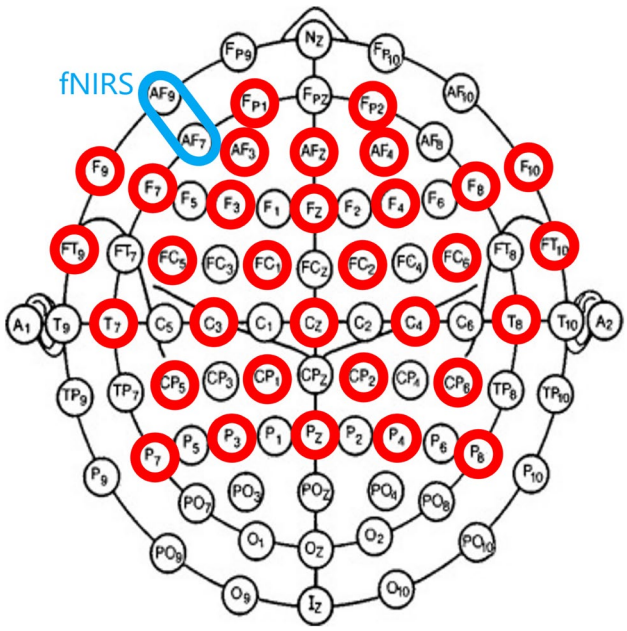

Fig. 7 EEG 32-lead electrode configuration (red circles) and fNIRS location (blue circle) used in this experiment

visualizing the emotional film clip. Then, participants had to answer the SAM and ad-hoc Appraisal scale after the end of each film clip, plus an additional interval of 10 s between each affective trial. Although the minimum spacing between each film clip was 10 s, this spacing and the time participants had to answer the SAM and Appraisal scale was enough for the emotional response to fully dissipate before presenting the following emotional stimulus. This was performed for every film clip, regardless of the emotional category, with no pause between consecutive emotional categories. Figure 9 shows a schematic of this procedure.

3.6 Signals processing

3.6.1 EEG

The EEG signals were acquired at a sampling frequency of 250 Hz. For every participant, the impedance levels for each electrode were checked using the g.NEEDacess software to reduce the noise in the input signals.

The raw EEG signals for each electrode were filtered using an independent component analysis (ICA). Then, an 8th

order Butterworth band-pass filter at frequencies between 1 and 50 Hz was applied to remove any noise artifact. Finally, the power of the following EEG bands was computed for each electrode: Alpha (8–14 Hz), Beta (14–30 Hz), Gamma (30–49 Hz) and Theta (4–8 Hz).

3.6.2 ECG

The ECG signal (Fig. 10) was acquired to obtain information regarding the user's heart rate (HR) and heart rate variability (HRV). Then, the Pan and Tompkins' algorithm was implemented to detect the R peak in the QRS complex of the ECG signal (Pan and Tompkins 1985). This algorithm starts by filtering the raw signals with a 5 to 15 Hz band-pass 2nd order Butterworth filter. Then it computes the derivative of the filtered ECG signal, followed by the squaring and the integration of the signal with a time window of 80 ms. Finally, the R peaks are detected using an adaptive threshold. After obtaining the R peaks, it is possible to compute the peak-to-peak interval series, denominated as the RR-interval time series, which gives us the participant's heart rate ($HR = 60/\Delta RR$). Then, from the RR-interval time series, we computed a total of 23 ultra-short-term HRV metrics in the time-domain, frequency-domain, and non-linear features. According to the task force of the European Society of Cardiology and the North American Society of Pacing Electrophysiology (Task Force of the European Society of Cardiology 1996), a minimum time window of 2 to 5 min is necessary to obtain reliable HRV spectral components, in opposition to the 40-s time window used in this work. However, according to recent studies, it is possible to compute ultra-short-term HRV measurements in mobile settings (Salahuddin et al. 2007; Schaaff and Adam 2013). Salahuddin et al. (2007) found significant differences between baseline and Stroop test using HRV measurements within a 50-s window. Similarly, Schaaff and Adam (2013) used a 30-s time window for frequency-domain and a 15-sec window for time-domain features to measure emotional arousal. Therefore, while it might not be reliable from a clinical standpoint, ultra-short-term HRV metrics may still contain valuable information as studies demonstrated their value in the context of emotions classification. All the 23 features calculated for the ECG have been enumerated in Table 1.

3.6.3 RESP

The RESP signal (Fig. 10) was processed using the neurokit2 toolkit (Makowski et al. 2021) and analyzed in the time-domain, including respiration rate variability metrics, frequency-domain and non-linear features. A total of 23 features were computed (see Table 1).

3.6.4 fNIRS

The fNIRS (Fig. 10) functions by taking advantage of the different optical properties of the oxygenated haemoglobin (HbO) and the deoxygenated haemoglobin (HbR), which, in the red and near-infrared spectrum (wavelength of 650 and 900 nm), have different absorption coefficients. The data was filtered with a 2nd order 0.05−0.4 Hz band-pass Butterworth filter. The data was then processed using the modified Beer–Lambert law to convert the radiation intensity to changes in oxygen concentration (HBO and HBR), following the algorithms presented by Varandas et al. (2022). A total of 12 features were extracted (see Table 1).

3.6.5 EDA

The EDA signal (Fig. 11) was acquired to obtain information regarding the skin conductance levels to measure participants' arousal. First, the raw signal was filtered using a 2nd order Butterworth low-pass filter at 1 Hz. Then, using the neurokit2 toolkit (Makowski et al. 2021), the phasic component (skin conductance response (SCR)) and the tonic component (skin conductance level (SCL)) were computed. Frequency-domain features were also obtained after performing a 3-step down-sampling with a factor of 10. Then, the frequency components corresponding to the frequency bands of the HRV were computed. A total of 26 features were extracted (see Table 1).

3.6.6 SKT

The SKT signal (Fig. 11) was filtered using a 2nd order Butterworth low-pass filter at 0.1 Hz, and then 4 time-domain features were computed (see Table 1).

3.7 Machine learning

In this study, only supervised machine learning methods were applied, thus, the focus will be on these methods. Supervised learning is a technique in which the algorithm learns from previously labelled data to predict never seen before data. It allows the models to correctly classify the input data into the target label faster than unsupervised learning algorithms. Thus, we considered the following supervised classifiers: K-nearest neighbor (kNN) (Pedregosa et al. 2011), support vector machines (SVM) (Vapnik 1998), random forest (RF) (Donges 2018), and artificial neural network (NN) (Jain et al. 1996). These classification models have been frequently used in previous studies regarding emotion recognition, as we enumerated in Sect. 2.6, thus,

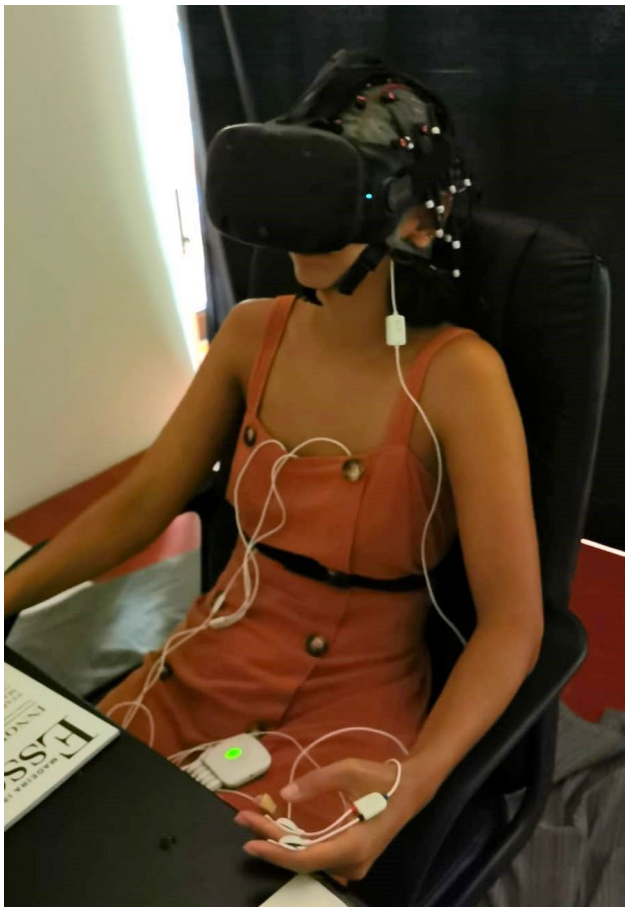

Fig. 8 An example of a participant wearing all the physiological sensors and the HTC Vive headset

it will be possible to compare our results with the results of other studies.

Machine learning methods are usually expensive to compute with a large set of features, and the data is associated with a high noise level. To overcome this problem, it is necessary to perform feature selection algorithms on the input data of the classification models to select a smaller subset of features that maximizes the relevance to the target output, minimizing redundancy and leading to a better learning performance of the classification models (Chen et al. 2020). Recursive feature elimination (RFE) is an example of a feature selection method in which the final set of features is selected by recursively considering smaller sets of features and eliminating the least informative feature in each iteration (Guyon et al. 2002).

Training machine learning models is a systematic process that maximizes the available data to train the model to obtain the best performance model. Before starting the training phase, it is common to normalize the input data. This normalization is performed using standardization (Shanker et al. 1996), which subtracts the statistical mean from each value and divides the results by the standard deviation. Equation 1 shows this normalization, where X is the original sample of values, $\overline{X}$ is the statistical mean of the values, $STD(X)$ is the standard deviation of X, and X' is the standardized new sample of values.

$$X' = \frac{X - \overline{X}}{STD(X)} \tag{1}$$

Splitting the original dataset into a training and testing set is necessary to start the training phase. This split is fundamental to get an accurate estimate of the model's performance when introduced to new data it has never seen before. The most used method to split the original data into training and testing data is to use cross-validation methods, such as K-fold and leave one out cross-validation (Refaeilzadeh et al. 2016; Wong 2015).

The model's performance was assessed using commonly used metrics for machine learning classification algorithms: accuracy, precision, recall, and F1-score (Eqs. 2–5), where TP is true positive, TN is true negative, FP is false positive, and FN is false negative (Mohammad and Nasir 2015).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

The area under the receiver operating characteristic curve (AUC-ROC) is also a performance metric for classification

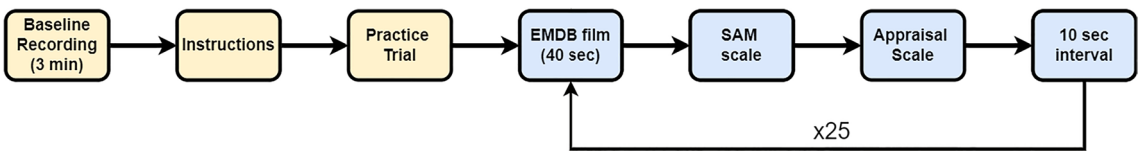

Fig. 9 Experimental procedure of this experiment for each session

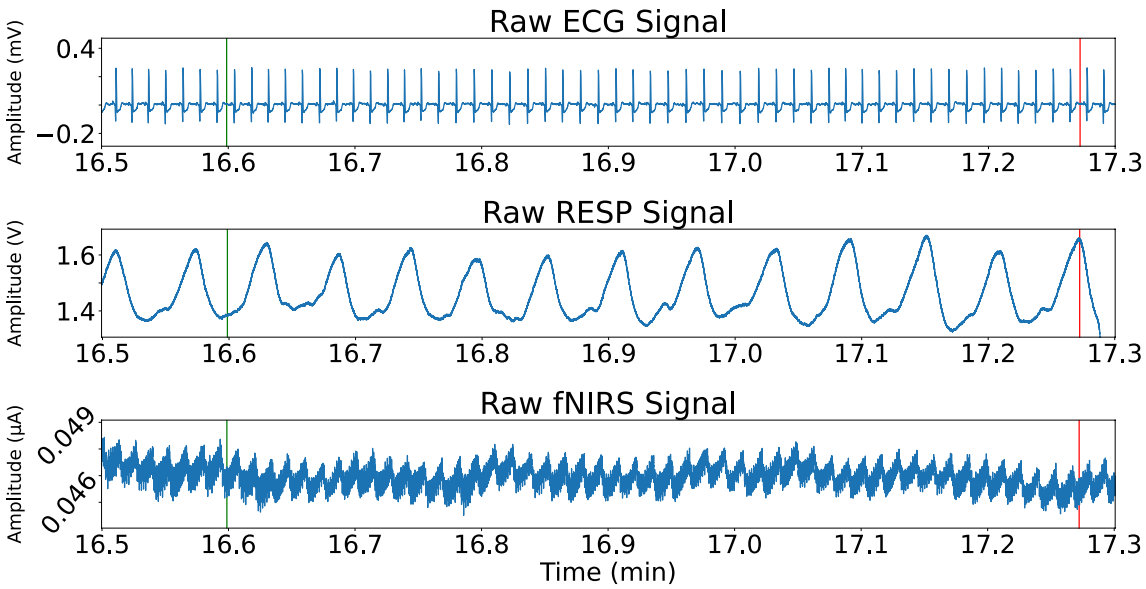

Fig. 10 Raw ECG, RESP and fNIRS signals acquired, respectively, on channel 1 in millivolts (mV), channel 3 in volts (V) and on the digital channel in micro-Amperes (μA). The green vertical line represents the onset time of the film, and the red line represents the offset time of the film

algorithms. It measures how much a model can distinguish between different classes. Thus, the higher the value of AUC, the better the model is at distinguishing between different classes (Hanley and McNeil 1982).

4 Results

In this experiment, we presented the emotional film clips in two different degrees of immersion—Immersive vs non-Immersive—to compare the effect of immersion in the film clips shown on the HMD on the subjective ratings of the SAM scale, thus, if it had any effect on the reported levels of arousal, valence and dominance, using non-parametric statistical tests. Then, we developed classification models to predict the emotional category of the film clips from the EMDB using the physiological signals features and the self-reported ratings.

4.1 Immersion

Regarding the two degrees of immersion (immersive and non-immersive), we obtained the data from the self-reported levels of arousal, valence and dominance and tested whether there were differences in these reported levels between the two degrees of immersion.

Since the data in the SAM scale is obtained through a 9-point Likert scale, the data is ordinal. Thus, non-parametric statistical tests need to be used. As mentioned in Sect. 3.5, the EMDB database was split in half. This resulted in two groups: Group A—the first five film clips of each category, and Group B—the remaining five film clips of each category. Then, participants were randomly assigned to view the film clips from Group A in one degree of immersion and clips from Group B in the other degree of immersion, and vice-versa. Thus, some participants viewed the film clips from Group A in the first session in the immersive setup and the film clips from Group B in the second session in the non-immersive setup. The remaining participants viewed the film clips from Group A in the non-immersive setup and film clips from group B in the immersive setup. Therefore, for each group, the data regarding arousal, valence and dominance was evaluated using the Mann–Whitney U statistical test to compare the respective mean ranks.

The results from the Mann–Whitney U test for the median of both groups are reported in Table 2. From these results, we only found a significant difference regarding *"Dominance"* in the category "Horror" for Group A and B ($U = 8.00$; p-value $= 0.043$; $r = 0.55$; Power $= 0.15$) and "social negative" for Group B ($U = 8.00$, p-value $= 0.043$; $r = 0.55$; Power $= 0.15$). No significant difference was found in the levels of arousal and valence in both groups.

4.2 Classification

Classification models were trained to classify the subjective ratings from the SAM (arousal, valence and dominance), ad-hoc appraisal scale (anticipatory effort, perceived obstacles, novelty, unexpectedness, intrinsic goal, perceived control,

Table 1 Physiological features extracted from the ECG, EDA, RESP, SKT fNIRS and EEG. A total of 216 features were computed

Sensor (# of features)	Features
ECG (23)	Average HR, Minimum HR, Maximum HR, Standard Deviation HR, Average RR-interval, Minimum RR-interval, Maximum RR-interval, SDNN, RMSSD, NN50, pNN50, NN20, pNN20, VLF (0.0033−0.04 Hz) power, LF (0.04−0.15 Hz) power, HF (0.15−0.40 Hz) power, Total Power, LF (n.u), HF (n.u.), LF/HF, SD1, SD2, SD2/SD1
EDA (26)	Average SCR, Maximum SCR, Minimum SCR, Standard Deviation SCR, Average SCL, Maximum SCL, Minimum SCL, Standard Deviation SCL, Average SCR amplitude, Maximum SCR amplitude, Minimum SCR amplitude, Standard Deviation SCR amplitude, Average Rise Time, Maximum Rise Time, Minimum Rise Time, Standard Deviation Rise Time, Average Recovery Time, Maximum Recovery Time, Minimum Recovery Time, Standard Deviation Recovery Time, VLF (0.0033−0.04 Hz) power, LF (0.04−0.15 Hz) power, HF (0.15−0.40 Hz) power, Total Power, LF (n.u), HF (n.u.)
RESP (23)	Average RESP Rate, Maximum RESP Rate, Minimum RESP Rate, Standard Deviation RESP Rate, Average RESP amplitude, Maximum RESP amplitude, Minimum RESP amplitude, Standard Deviation RESP amplitude, Median Breath-to-Breath interval, Median Absolute Deviation BB, MCVBB, SDBB, SDSD BB, RMSSD, SDSD, CVBB, CVSD, BB50, pBB50, BB20, pBB20, HF (0.15−0.40 Hz) power, SD1
SKT (4)	Average Skin Temperature, Maximum Skin Temperature, Minimum Skin Temperature, Standard Deviation Skin Temperature
fNIRS (12)	Average ΔHbO, Minimum ΔHbO, Maximum ΔHbO, Standard Deviation ΔHbO, Average ΔHbR, Minimum ΔHbR, Maximum ΔHbR, Standard Deviation ΔHbR, Average ΔHbTotal, Minimum ΔHbTotal, Maximum ΔHbTotal, Standard Deviation ΔHbTotal
EEG (128)	Alpha (8–14 Hz) band power, Beta (14–30 Hz) band power, Gamma (30–49 Hz) band power, Theta (4–8 Hz) band power for the 32 electrodes

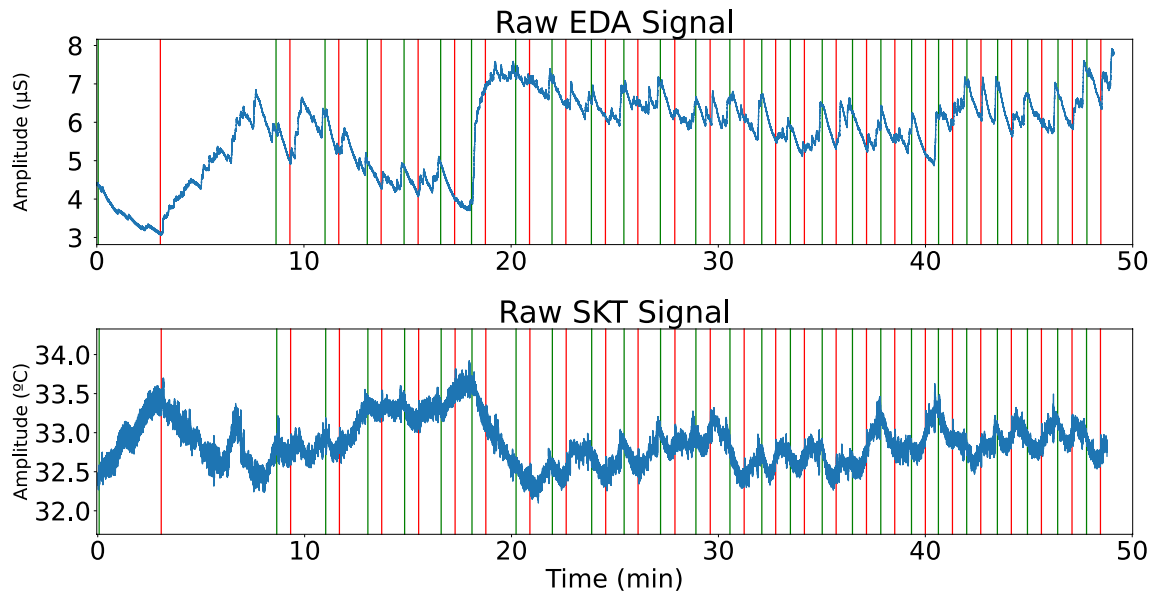

Fig. 11 Raw EDA and SKT signals acquired, respectively, on channels 2 in micro-Siemens (μS) and 4 in Celsius degrees (°C). The green vertical line represents the onset time of the film, and the red line represents the offset time of the film

and control/stress), and the features extracted from the physiological signals into the five target emotional categories of the EMDB (erotic, horror, social positive, social negative and scenery), according to the corresponding labels of the EMDB database. Since no significant difference was found for the levels of arousal and valence in both groups of the previous section (Sect. 4.1), all the film clips will be used as input data to the classification models, regardless of the degree of immersion in which they were presented to the participants. The results will be reported from a user-independent and a user-dependent point-of-view.

4.2.1 Subjective ratings

From a user-independent point-of-view, the following classifiers were used to evaluate which one had a better performance: k-nearest neighbor using the five closest neighbors (k = 5), SVM (C = 1, a radial basis function (RBF) kernel, and random seed of 42), a random forest classifier with 100 decision trees (Gini criterion and random seed of 42), and a multi-layer perceptron classifier with a random seed of 42.

The data was split into training and testing sets using the cross-validation method from scikit-learn (Pedregosa et al. 2011), leave one subject out (LOSO). Thus, the samples from one user are used as a testing sample, as this process is repeated for every user available. This results in 14 different subsets of data. The final accuracy score is the average accuracy across all users.

Then, a recursive feature elimination (RFE) algorithm was applied as a feature selection method to reduce the number of input features provided to the classification models. This feature selection was applied only to the training set to avoid data leakage and overfitting the model performance, using a RF classifier as an estimator of the features' importance, removing the least important feature on each algorithm iteration and mitigating wrapper bias (Samala et al. 2020). The final number of input features used to evaluate the model's performance was reduced to half (5 features).

The results obtained for the user-independent classifiers are shown in Table 3. The metrics used are Accuracy, Precision, F1-Score and ROC-AUC with a weighted average parameter. The ROC-AUC was computed using a one-vs-rest configuration—which computes the AUC of each class against the rest.

All the results are in percentage. The classifier with the best performance in classifying the emotional categories using the subjective ratings is the SVM, with an average accuracy of 54.0 ± 17.2%.

From a user-dependent point-of-view, the classifier with the best performance was the RF classifier with 100 decision trees (Gini criterion, random seed of 42). This classifier was used for each individual to classify the emotional categories.

The data of each user was split into training and testing sets using a Stratified K-Fold method from scikit-learn (Pedregosa et al. 2011) with ten folds (k=10), preserving the percentage of samples for each target class on each fold. This results in 10 different subsets of data from that specific user. For each user, the model's accuracy is the average accuracy between the 10 subsets of data from that user.

Then, the RFE algorithm was also applied to the training set as a feature selection method, using the same classifier (random forest) as an estimator, mitigating wrapper bias (Samala et al. 2020). This process entails 140 subsets of data for our sample (14 x 10), with different features being selected as the most important for each user. Table 4 shows the results obtained for each user.

The average accuracy of all the users was 69.3 ± 11.4%, with a minimum and maximum accuracy of 56% and 90%, respectively.

4.2.2 Physiological signals

Regarding the physiological signals, 216 features were extracted from the following sensors: EEG, ECG, EDA, RESP, fNIRS and SKT.

From a user-independent point-of-view, the same procedure performed in Sect. 4.2.1 was applied to these features, but using a random seed of 0, instead of a random seed of 42. After completing the RFE feature selection algorithm, 108 features were used as input features for the classification models.

The results are shown in Table 5. The classifier with the best performance in classifying the emotional categories using the physiological features is the RF classifier, with an average accuracy of 24.9 ± 4.0%.

From a user-dependent point-of-view, the classifier with the best performance was also the RF classifier with 100 decision trees (Gini criterion, random seed of 0). This classifier was used for each individual to classify the emotional categories.

The data was split into training and testing sets using the same procedure as in Sect. 4.2.1. Then, the RFE algorithm was also applied as a feature selection method to reduce the number of input features to half (108 input features). The results obtained for each user are shown in Table 6.

The average accuracy of all the users was 71.0 ± 8.0%, with a minimum and maximum accuracy of 58% and 86%, respectively.

Table 2 Results from the Mann–Whitney U test to compare the effect of the degrees of immersion in terms of arousal, valence and dominance, for the median of both groups

	Category	Group A[a]			Group B[b]		
		Non immersive	Immersive	*p*-value	Non immersive	Immersive	*p*-value
Arousal	Erotic	6.0	5.4	0.662	5.0	4.9	0.573
	Horror	6.9	5.8	0.345	5.7	6.1	0.852
	Social negative	4.8	3.0	0.108	2.5	4.1	0.345
	Scenery	3.1	3.0	0.755	3.2	3.1	0.755
	Social positive	3.7	3.1	0.755	2.9	3.2	0.950
Valence	Erotic	5.5	6.5	0.345	6.5	5.0	0.345
	Horror	1.3	2.0	0.345	1.8	1.4	0.573
	Social negative	3.7	3.4	0.662	3.9	4.1	0.491
	Scenery	7.1	6.1	0.282	5.9	6.7	0.491
	Social positive	6.9	6.6	1.00	6.5	5.9	0.662
Dominance	Erotic	3.3	5.2	0.081	5.0	4.6	0.573
	Horror	2.3	4.8	0.043*	4.7	2.5	0.043*
	Social negative	4.3	4.9	0.345	5.0	3.8	0.043*
	Scenery	4.1	5.0	0.282	5.3	3.8	0.081
	Social positive	4.2	5.1	0.228	5.0	5.3	0.662

[a]Group A—Film clips number: 1000–1004, 2000–2004, 3000–3004, 4000–4004, 5000–5004

[b]Group B—Film clips number: 1005–1009, 2005–2009, 3005–3009, 4005–4009, 5005–5009

*p-value < 0.05

Table 3 Results for the user-independent classification of the subjective ratings

	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)	ROC-AUC (%)
KNN	45.9 ± 13.3	51.3 ± 11.3	42.4 ± 15.0	45.9 ± 13.3	75.7 ± 7.9
SVM	54.0 ± 17.2	64.3 ± 13.5	49.2 ± 20.0	54.0 ± 17.2	84.9 ± 8.9
Random forest	45.4 ± 16.0	51.3 ± 16.2	41.2 ± 18.2	45.1 ± 16.0	79.1 ± 8.9
NN	52.7 ± 19.3	60.2 ± 16.7	48.6 ± 21.9	52.7 ± 19.3	83.3 ± 8.9

5 Discussion

The main goal of this work was to study and validate a machine learning pipeline (data collection, data preprocessing, feature extraction and feature selection) to predict emotional states from a combination of physiological signals while participants were subjected to validated emotional stimuli. Then, this pipeline will be transferred to a real-time prediction system to predict emotional states in real time. Thus, selecting the appropriate emotional stimuli database and the presentation format of these stimuli was necessary to elicit the emotional response required to be measurable through physiological signals.

Regarding the degree of immersion of the emotional stimuli, when analyzing the impact of the immersion, we only found a significant difference for *"Dominance"* in the categories "horror" and "social negative", as participants reported lower levels of control—being controlled by the stimulus - in the immersive setup compared to non-Immersive. However, previous studies have shown that dominance does not provide consistent results across several studies and that arousal and valence are considered the most important metrics to measure basic emotions (Betella and Verschure 2016). Thus, we assume that there is no difference in the degree of immersion between the Immersive vs non-Immersive setup, as all the film clips were used as input data to the classification models.

In this study, we compared the performance of both user-independent and user-dependent classification models for the subjective ratings and physiological signals to classify the five emotional classes of the EMDB database. Most of the previous work in emotion recognition state that emotions are highly subjective, so generalized models for the classification of emotions are still a challenge. Thus, classification models should be developed from user-dependent models rather than generalized models for all subjects (Varandas et al. 2022).

Therefore, in this work, we did not focus on feature and model utility. Since our procedure will be applied in a real-time experiment, continuously monitoring the participant's emotional state is needed. Thus, we had to consider computing ultra-short frequency features for HRV, EDA and RESP signals. Although there are no clear guidelines on which features are comparable and reliable as the standard 5-min recordings (Pecchia et al. 2018), several studies apply ultra-short features.

For instance, regarding HRV spectral features, Salahuddin et al. (2007) showed that it was possible to monitor mental stress with wearable sensors, using a 30-sec window. Similarly, Schaaff and Adam (2013) also computed HRV frequency features using a 30-sec window, to classify emotional arousal. Bernardes et al. (2022) also found out that the LF and HF band computed in a 40-sec window had a correlation higher than 50% compared to a 2-min window, indicating a moderate to strong positive correlation.

Regarding the performed frequency analysis for EDA, we based our approach on the proposed by Posada-Quintero et al. (2016) to assess the dynamics of the autonomous nervous system based on the spectral analysis and frequency bands of HRV. Although the skin conductance level (SCL) is a slow-changing signal and may not be captured within the 40-second window, the skin conductance response (SCR) occurs as the sympathetic response to a stimulus, about 1–5 s after the presentation of a stimulus (Posada-Quintero and Chon 2020), should be measurable. Indeed, Posada et al. found that there was an increase in spectral power of the EDA signals between the frequencies 0.08–0.24 Hz for several stressors. Therefore, we believe that the spectral analysis of EDA signals is an indicator of the sympathetic-parasympathetic balance (Posada-Quintero and Chon 2020), so we included these features in our pipeline.

Acknowledging the challenges of using these ultra-short features, we analyzed the number of occurrences of the spectral features computed for the classification models. These results are shown in Tables 7 and 8. From these tables, it is possible to conclude that the VLF power band for HRV and EDA was never selected for either user-dependent or user-independent pipelines. Additionally, regarding the LF and HF bands for HRV and EDA, we can see that for EDA, these features had a high rate of selection, evidencing that features in this range of frequencies for EDA are reliable and associated with the activation of the sympathetic tone. For HRV, these features had a lower selection rate when compared to

Table 4 Results for the user-dependent classification of the subjective ratings

Users	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)	ROC-AUC (%)
1	76.0	85.7	70.3	76.0	93.5
2	56.0	74.3	46.7	56.0	83.0
3	58.0	79.5	49.1	58.0	83.3
4	70.0	81.0	64.0	70.0	91.3
6	64.0	75.0	56.7	64.0	86.8
7	58.0	75.0	51.7	58.0	89.5
8	90.0	95.0	86.7	90.0	98.5
9	76.0	85.0	70.0	76.0	93.3
10	58.0	74.3	50.0	58.0	93.0
11	78.0	86.7	73.0	78.0	95.3
12	86.0	92.0	82.0	86.0	98.3
13	78.0	89.3	72.0	78.0	96.0
14	64.0	79.7	57.0	64.0	89.8
15	58.0	76.0	51.0	58.0	84.8
Average	69.3 ± 11.4	82.0 ± 7.0	62.9 ± 13.0	69.3 ± 11.4	91.1 ± 5.2

Table 5 Results for the user-independent classification of the physiological signals

	Accuracy (%)	Precision (%)	F1-Score (%)	Recall (%)	ROC-AUC (%)
KNN	22.1 ± 6.7	30.4 ± 16.1	18.0 ± 7.1	22.1 ± 6.7	51.6 ± 7.7
SVM	21.0 ± 4.1	34.8 ± 20.0	14.9 ± 6.1	21.0 ± 4.1	54.5 ± 5.4
Random forest	24.9 ± 4.0	28.7 ± 8.1	21.9 ± 5.8	24.9 ± 4.0	55.5 ± 5.3
NN	23.9 ± 4.5	29.3 ± 16.6	19.5 ± 5.8	23.9 ± 4.5	54.6 ± 4.8

the EDA frequency features. Both these results are similar to the results obtained by Finseth et al. (2023) and suggests that our feature selection process only selected reliable features.

Regarding related work using film clips for emotion recognition, the classification accuracy ranges from 71.0 to 85.3% using user-dependent models (Jerritta et al. 2011). Thus, these accuracy values were the reference values to compare the performance of our models.

In Sect. 4.2.1, the subjective ratings from the SAM and the ad-hoc Appraisal scale were used to predict the emotional categories of the EMDB. Using user-independent classification models (Table 3), we obtained an accuracy of 54.0 ± 17.2% using an SVM classifier. On the other hand, using a user-dependent classifier for each participant (Table 4), the average accuracy was 69.3 ± 11.4%, ranging from 56% to 90% accuracy using a Random Forest classifier. The average ROC-AUC obtained was 91.1 ± 5.2%, which indicates that our models have a high ability to distinguish between classes.

In Sect. 4.2.2, the features extracted from the physiological signals were used with the same goal—to predict the emotional category from the EMDB database. Again, the accuracy was lower in a user-independent classification than in a user-dependent classification. The accuracy of the user-independent was 24.9 ± 4.0% (Table 5) for a Random Forest Classifier. On the other hand, the average accuracy for user-dependent classification was 71.0 ± 8.0%, from 58 to 86% (Table 6). The average ROC-AUC was 91.0 ± 3.9%. which indicates that our models have a high ability to distinguish between classes.

Compared to previous works that used film databases in non-immersive environments, we note that results related to user-independent models achieved lower accuracy scores, while user-dependent results are similar. This points to the hypothesis that immersive environments lead to higher variability of emotional responses, thus hardening the task of generalization across subjects.

Regarding the number of classified emotions, it is also clear that the higher the number of target classes, the lower the accuracy scores results. Namely, the work of Li and Chen (2006) achieved an accuracy score of 85.3% classifying 2 emotions, while Nasoz et al. (2004) achieved an accuracy score of 71% for 6 target emotions. Thus, our work that classifies 5 different target classes achieved results similar to those that classify a higher number of emotions in the literature using user-dependent models.

On the other hand, we see that different stimuli lead to different results in emotion detection. For example, previous works that used IAPS, a database composed of pictures, achieved results as high as 92% in user-dependent scenarios,

Table 6 Results for the user-dependent classification of the physiological signals

User	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)	ROC-AUC (%)
1	74.0	84.7	67.7	74.0	95.3
2	70.0	84.7	61.0	70.0	92.8
3	58.0	72.0	50.0	58.0	86.3
4	66.0	79.7	59.0	66.0	89.5
6	62.0	82.0	53.7	62.0	88.0
7	80.0	89.7	74.3	80.0	93.3
8	70.0	75.0	66.7	70.0	88.3
9	70.0	80.0	63.3	70.0	90.8
10	70.0	82.0	64.7	70.0	89.0
11	86.0	92.0	82.0	86.0	98.5
12	64.0	76.0	56.0	64.0	92.8
13	68.0	80.3	62.7	68.0	87.0
14	72.0	80.7	67.0	72.0	86.8
15	84.0	92.0	78.7	84.0	96.3
Average	71.0 ± 8.0	82.2 ± 6.0	64.8 ± 9.1	71.0 ± 8.0	91.0 ± 3.9

which is higher than our work and all works that employed film-based databases. Despite higher results, since pictures are less realistic than films, their use may be less adequate for real-world applications.

The results for subjective ratings and physiological signals support the hypothesis that the recognition of emotions is subjective for each individual. The accuracy scores for subjective ratings and physiological signals were lower for user-independent classification models compared to user-dependent, as expected from Sect. 3.7. Specifically, the ROC-AUC for user-independent models using physiological signals showed that the model could not distinguish and separate the data into the different emotional categories, while user-dependent models could clearly distinguish the data between different emotional classes. Thus, individuals react differently to the same emotional stimuli.

Moreover, it is possible to conclude that the subjective ratings responses can be more generalized between every participant since the difference in accuracy between user-independent models (54.0 ± 17.2%) and user-dependent models (69.3 ± 11.4%) is considerably lower than the difference between these models using the physiological signals. This could be explained by the fact that participants were aware of the emotional content presented to them. Thus, the responses to the subjective scales for each emotional film clip are similar across subjects, while the physiological response to these stimuli is distinct due to the way each participant reacts to the emotional stimuli presented.

6 Limitations and future work

Some limitations can be pointed out in this work. Firstly, this work is the first exploratory step to develop a machine learning pipeline using physiological data to be applied in subsequent real-time systems to predict emotional content. Hence, this work aimed to validate the proposed pipeline, centering our approach on model creation and quantification through cross-validation. Given that our research intends to develop systems for real-time processing, we could not disregard the potential of ultra-short features. The choice of the 40-sec window relied on the length of the EMDB film clips, which limited the maximum window length for data acquisition. Opting for a larger window to obtain these features would mean incorporating information about the transitions between emotional stimuli into the classification model.

Acknowledging that some of these ultra-short features might be unreliable in such a short window size, using the recursive feature elimination to select the relevant features, it is expected that only reliable ultra-short frequency features in this window can be considered in the feature selection

Table 7 Number of frequency feature occurrences for the user-independent models

N = 14	HRV					EDA					RESP
	VLF	LF	HF	LF (n.u)	HF (n.u)	VLF	LF	HF	LF (n.u)	HF (n.u)	HF
Count	0	5	3	2	5	0	14	4	10	1	1
%	0%	36%	21%	14%	36%	0%	100%	29%	71%	7%	7%

Table 8 Number of frequency feature occurrences for the user-dependent models

N =140	HRV					EDA					RESP
	VLF	LF	HF	LF (n.u)	HF (n.u)	VLF	LF	HF	LF (n.u)	HF (n.u)	HF
Count	0	32	60	45	45	0	102	107	50	49	50
%	0%	23%	43%	32%	32%	0%	73%	76%	36%	35%	36%

algorithm. Nevertheless, to understand to what extent these were used in the models, we now report the number of occurrences of each frequency feature feeding the models (see Tables 7 and 8). In this analysis, we showed that VLF features are never selected for any model and identify the ones that have been considered and are relevant to distinguishing the emotional categories.

In future work, it is necessary to develop models to adapt the virtual environment in real-time according to the user's emotional state. One of the main limitations of the real-time system is the inability to retrain the model according to the changes in the physiological signals over time and repeated exposure to emotional stimuli (Finseth et al. 2023). Thus, this limitation needs to be addressed in future studies. Nevertheless, these models were trained on healthy individuals, which may not represent individuals with health disorders. However, since the outcome of this work points us to use user-dependent models, a new model will be developed for each user, regardless of whether they are healthy or not. This implies that the most informative sensors may not generalize to clinical populations. Thus, in future work, we will validate whether these sensors and methodology apply to individuals with health disorders, improving ecological validity in out-of-the-lab applications for healthy and clinical populations.

7 Conclusion

This work aimed to develop machine learning classification models to detect and recognize emotional states using a combination of physiological sensors through wearable devices, starting from videos.

The proposed methodology effectively elicited the target emotional states by presenting emotional film clips from the EMDB dataset in an immersive and non-immersive virtual environment.

The results showed that the degree of immersion of the emotional stimuli was not relevant since there were no significant differences in the reported levels of arousal and valence for each emotional category. In terms of the classification models, it was clear that the emotional response is highly subjective and it depends on each user. Thus, we obtained higher accuracy using user-dependent models compared to user-independent models. The accuracy for user-dependent models was 69.3 ± 11.4% and 71.0 ± 8.0% using subjective ratings and physiological signals, respectively.

Author contributions Conceptualization: RL, AC, SB; data curation: RL; formal analysis: RL; investigation: RL; methodology: RL; project administration: AC, HG, AG, SB; resources: RL, AC; software: RL, RV; supervision: HG, AG, SB; validation: AC, HG, AG, SB; visualization: RL; writing—original draft: RL; writing—review and editing: RV, AC, HG, AG, SB;

Funding Open access funding provided by FCT|FCCN (b-on). This work was funded by the FCT—Fundação para a Ciência e Tecnologia, through the Ph.D. Grants 2020.06024.BD and PD/BDE/150304/2019, supported by the NOVA Laboratory of Computer Science and Informatics (UIDB/04516/2020) and BRaNT project (PTDC/CCI-COM/30990/2017), by the ARDITI - Agência Regional para o Desenvolvimento da Investigação, Tecnologia e Inovação, through the project MACBIOIDI2 (MAC2/1.1b/352), and by PLUX Wireless Biosignals, S.A.

Availability of data and materials The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest Authors Rui Varandas and Hugo Gamboa are affiliated with PLUX Wireless Biosignals S.A., the company that produces the biosignalsplux acquisition device used in this work and also the OpenSignals software. The remaining authors have no Conflict of interest to declare that are relevant to the content of this article.

Ethics approval This work was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of University of Madeira (approved on the 17th of February of 2022).

Consent to participate A written informed consent was obtained from all the participants included in this work.

References

Achlioptas P, Ovsjanikov M, Haydarov K, et al (2021) ArtEmis: affective language for visual art. CoRR arXiv:2101.07396

Aue T, Flykt A, Scherer KR (2007) First evidence for differential and sequential efferent effects of stimulus relevance and goal conduciveness appraisal. Biol Psychol 74(3):347–357. https://doi.org/10.1016/j.biopsycho.2006.09.001

Berking M, Wupperman P (2012) Emotion regulation and mental health: Recent findings, current challenges, and future directions. Curr Opin Psychiatry 25(2):128–134. https://doi.org/10.1097/YCO.0b013e3283503669

Bernardes A, Couceiro R, Medeiros J et al (2022) How reliable are ultra-short-term HRV measurements during cognitively demanding tasks? Sensors. https://doi.org/10.3390/s22176528

Betella A, Verschure PF (2016) The affective slider: a digital self-assessment scale for the measurement of human emotions. PLoS ONE. https://doi.org/10.1371/journal.pone.0148037

Borrego A, Latorre J, Llorens R et al (2016) Feasibility of a walking virtual reality system for rehabilitation: Objective and subjective parameters. J NeuroEng Rehabil. https://doi.org/10.1186/s12984-016-0174-1

Bota PJ, Wang C, Fred AL et al (2019) A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. IEEE Access 7:140990–141020. https://doi.org/10.1109/ACCESS.2019.2944001

Boucsein W (2012) Electodermal activity, Second edn. https://doi.org/10.1007/978-1-4614-1126-0

Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Therapy Exp Psychiatry 25(1):49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Cacioppo JT, Tassinary LG (1990) Inferring psychological significance from physiological signals. Am Psychol 45(1):16–28. https://doi.org/10.1037//0003-066x.45.1.16

Carvalho S, Leite J, Galdo-Álvarez S et al (2012) The emotional movie database (EMDB): a self-report and psychophysiological study. Appl Psychophysiol Biofeedback 37(4):279–294. https://doi.org/10.1007/s10484-012-9201-6

Chanel G, Ansari-Asl K, Pun T (2007) Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In: Conference proceedings - IEEE international conference on systems, man and cybernetics 41(22):2662–2667. https://doi.org/10.1109/ICSMC.2007.4413638

Chen RC, Dewi C, Huang SW et al (2020) Selecting critical features for data classification based on machine learning methods. J Big Data. https://doi.org/10.1186/s40537-020-00327-4

Chen S, Zhang L, Jiang F et al (2020) Emotion recognition based on multiple physiological signals. Zhongguo yi liao qi xie za zhi Chin J Med Instrum 44(4):283–287. https://doi.org/10.3969/j.issn.1671-7104.2020.04.001

Kothe C, Medine D, Boulay C, et al (2019) LabStreamingLayer. https://github.com/sccn/labstreaminglayer

Crawford HJ, Clarke SW, Kitner-Triolo M (1996) Self-generated happy and sad emotions in low and highly hypnotizable persons during waking and hypnosis: Laterality and regional EEG activity differences. Int J Psychophysiol 24(3):239–266. https://doi.org/10.1016/S0167-8760(96)00067-0

Dan-Glauser ES, Scherer KR (2011) The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. Behav Res Methods 43(2):468–477. https://doi.org/10.3758/s13428-011-0064-1

Denson TF, Spanovic M, Miller N (2009) Cognitive appraisals and emotions predict cortisol and immune responses: a meta-analysis of acute laboratory social stressors and emotion inductions. Psychol Bull 135(6):823–853. https://doi.org/10.1037/a0016909

Donges N (2018) The random forest algorithm. https://machinelearning-blog.com/2018/02/06/the-random-forest-algorithm/https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Egger M, Ley M, Hanke S (2019) Emotion Recognition from Physiological Signal Analysis: A Review. Electron Notes Theor Comput Sci 343(May):35–55. https://doi.org/10.1016/j.entcs.2019.04.009

Ekman P (1992) An argument for basic emotions. Cogn Emot 6(3–4):169–200. https://doi.org/10.1080/02699939208411068

Finseth TT, Dorneich MC, Vardeman S et al (2023) Real-time personalized physiologically based stress detection for hazardous operations. IEEE Access 11:25431–25454

Garcia-Garcia JM, Penichet VM, Lozano MD (2017) Emotion detection: a technology review. In: ACM international conference proceeding series part F1311(October). https://doi.org/10.1145/3123818.3123852

Goncalves A, Borrego A, Latorre J, et al (2021) Evaluation of a low-cost virtual reality surround-screen projection system. IEEE Trans Vis Comput Graph, PP. https://doi.org/10.1109/TVCG.2021.3091485

Gouizi K, Bereksi Reguig F, Maaoui C (2011) Emotion recognition from physiological signals. J Med Eng Technol 35(6–7):300–307. https://doi.org/10.3109/03091902.2011.601784

Gu Y, Tan SL, Wong KJ, et al (2010) A biometric signature based system for improved emotion recognition using physiological responses from multiple subjects. In: 2010 8th IEEE international conference on industrial informatics, pp 61–66. https://doi.org/10.1109/INDIN.2010.5549464

Guyon I, Weston J, Barnhill S et al (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1):389–422. https://doi.org/10.1023/A:1012487302797

Haag A, Goronzy S, Schaich P et al (2004) Emotion Recognition using bio-sensors: first steps towards an automatic system. In: André E, Dybkjær L, Minker W et al (eds) Affective dialogue systems. Springer, Berlin, pp 36–48

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hu X, Yu J, Song M et al (2017) EEG correlates of ten positive emotions. Front Hum Neurosci. https://doi.org/10.3389/fnhum.2017.00026

Inwood E, Ferrari M (2018) Mechanisms of change in the relationship between self-compassion, emotion regulation, and mental health: a systematic review. Appl Psychol Health Well Being 10(2):215–235. https://doi.org/10.1111/aphw.12127

Jain AK, Mao J, Mohiuddin KM (1996) Artificial neural networks: a tutorial. Computer 29(3):31–44. https://doi.org/10.1109/2.485891

Jerritta S, Murugappan M, Nagarajan R et al (2011) Physiological signals based human emotion recognition: A review. In: Proceedings - 2011 IEEE 7th international colloquium on signal processing and its applications. CSPA 2011:410–415. https://doi.org/10.1109/CSPA.2011.5759912

Juvrud J, Gredebäck G, Åhs F et al (2018) The immersive virtual reality lab: possibilities for remote experimental manipulations of autonomic activity on a large scale. Front Neurosci. https://doi.org/10.3389/fnins.2018.00305

Kim J, André E (2008) Emotion recognition based on physiological changes in music listening. IEEE Trans Pattern Anal Mach Intell 30(12):2067–2083. https://doi.org/10.1109/TPAMI.2008.26

Koelstra S, Mühl C, Soleymani M et al (2012) DEAP: a database for emotion analysis; Using physiological signals. IEEE Trans Affect Comput 3(1):18–31. https://doi.org/10.1109/T-AFFC.2011.15

Lang PJ, Bradley MM, Cuthbert BN (2005) IAPS: Affective ratings of pictures and instruction manual. Emotion

Li L, Chen Jh (2006) Emotion Recognition Using Physiological Signals from Multiple Subjects. In: 2006 international conference on intelligent information hiding and multimedia, pp 355–358. https://doi.org/10.1109/IIH-MSP.2006.265016

Liapis A, Xenos M (2013) The physiological measurements as a critical indicator in users' experience evaluation. In: ACM international conference proceeding series. https://doi.org/10.1145/2491845.2491883

Lindquist KA, Kober H, Bliss-Moreau E et al (2015) The brain basis of emotion: A meta-analytic review. Behav Brain Sci. 35(3):121–143 https://doi.org/10.1017/S0140525X11000446.The, https://www.cambridge.org/core/product/identifier/S0140525X11000446/type/journal_article

Liu C, Rani P, Sarkar N (2005) An empirical study of machine learning techniques for affect recognition in human-robot interaction. In: 2005 IEEE/RSJ international conference on intelligent robots and systems, pp 2662–2667. https://doi.org/10.1109/IROS.2005.1545344

Maaoui C, Pruski A (2010) Emotion recognition through physiological signals for human-machine communication. In: Kordic V (ed) Cutting edge robotics 2010. IntechOpen, Rijeka, chap 20. https://doi.org/10.5772/10312,

Makowski D, Pham T, Lau ZJ et al (2021) NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav Res Methods 53(4):1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Marín-Morales J, Llinares C, Guixeres J, et al (2020) Emotion recognition in immersive virtual reality: from statistics to affective computing. Sensors 20(18). https://doi.org/10.3390/s20185163, https://www.mdpi.com/1424-8220/20/18/5163

Meehan M, Insko B, Whitton M et al (2002) Physiological measures of presence in stressful virtual environments. ACM Trans. Graph. 21(3):645–652. https://doi.org/10.1145/566654.566630

Mehrabian A, Russell JA (1974) An approach to environmental psychology

Moghimi S, Kushki A, Power S et al (2012) Automatic detection of a prefrontal cortical response to emotionally rated music using multi-channel near-infrared spectroscopy. J Neural Eng. https://doi.org/10.1088/1741-2560/9/2/026022

Mohammad H, Nasir MdS (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manage Process 5(2):01–11. https://doi.org/10.5121/ijdkp.2015.5201

Naseer N, Hong KS (2015) fNIRS-based brain-computer interfaces: a review. Front Human Neurosci 9(Jan):1–15. https://doi.org/10.3389/fnhum.2015.00003

Nasoz F, Alvarez K, Lisetti CL et al (2004) Emotion recognition from physiological signals using wireless sensors for presence technologies. Cogn Technol Work 6(1):4–14. https://doi.org/10.1007/s10111-003-0143-x

Pan J, Tompkins W (1985) A real-time QRS detection algorithm. Biomed Eng IEEE Trans 32(3):230–236

Pecchia L, Castaldo R, Montesinos L et al (2018) Are ultra-short heart rate variability features good surrogates of short-term ones? State-of-the-art review and recommendations. Healthc Technol Lett 5(3):94–100. https://doi.org/10.1049/htl.2017.0090

Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12(85):2825–2830. https://doi.org/10.1289/EHP4713

Peirce J, Gray JR, Simpson S et al (2019) PsychoPy2: Experiments in behavior made easy. Behav Res Methods 51(1):195–203. https://doi.org/10.3758/s13428-018-01193-y

Plutchik R (1982) A psychoevolutionary theory of emotions. Soc Sci Inf 21(4–5):529–553. https://doi.org/10.1177/053901882021004003

Posada-Quintero HF, Chon KH (2020) Innovations in electrodermal activity data collection and signal processing: a systematic review. Sensors (Switzerland). https://doi.org/10.3390/s20020479

Posada-Quintero HF, Florian JP, Orjuela-Cañón AD et al (2016) Power spectral density analysis of electrodermal activity for sympathetic function assessment. Ann Biomed Eng 44(10):3124–3135. https://doi.org/10.1007/s10439-016-1606-6

Ramzan N, Palke S, Cuntz T et al (2016) Emotion Recognition by Physiological Signals. Electronic Imaging 2016:1–6. https://doi.org/10.2352/ISSN.2470-1173.2016.16.HVEI-129

Refaeilzadeh P, Tang L, Liu H (2016) Cross-validation. Springer, New York, pp 1–7. https://doi.org/10.1007/978-1-4899-7993-3_565-2

Reisenzein R, Horstmann G, Schützwohl A (2019) The cognitive-evolutionary model of surprise: a review of the evidence. Top Cogn Sci 11(1):50–74. https://doi.org/10.1111/tops.12292

Renard Y, Lotte F, Gibert G et al (2010) OpenViBE: an open-source software platform to design, test, and use brain-computer interfaces in real and virtual environments. Presence Teleop Virt 19(1):35–53. https://doi.org/10.1162/pres.19.1.35

Rey B, Clemente M, Wrzesien M et al (2014) Assessing brain activations associated with emotional regulation during virtual reality mood induction procedures. Expert Syst Appl 42(3):1699–1709. https://doi.org/10.1016/j.eswa.2014.10.006

Rigas G, Katsis CD, Ganiatsas G et al (2007) A user independent, biosignal based, emotion recognition method. In: Conati C, McCoy K, Paliouras G (eds) User modeling 2007. Springer, Berlin, pp 314–318

Rose T, Nam CS, Chen KB (2018) Immersion of virtual reality for rehabilitation—review. Appl Ergon 69:153–161. https://doi.org/10.1016/j.apergo.2018.01.009

Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39(6):1161–1178. https://doi.org/10.1037/h0077714

Salahuddin L, Cho J, Jeong MG et al (2007) Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In: Annual international conference of the ieee engineering in medicine and biology society ieee engineering in medicine and biology society annual international conference 2007:4656–4659. https://doi.org/10.1109/IEMBS.2007.4353378

Samala RK, Chan HP, Hadjiiski L, et al (2020) Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks. In: Medical imaging 2020: computer-aided diagnosis, AA(University of Michigan), AB(University of Michigan), AC(University of Michigan), AD(University of Michigan), p 1131416, https://doi.org/10.1117/12.2549313, https://ui.adsabs.harvard.edu/abs/2020SPIE11314E..16S

Schaaff K, Adam MTP (2013) Measuring emotional arousal for online applications: Evaluation of ultra-short term heart rate variability measures. In: 2013 Humaine association conference on affective computing and intelligent interaction pp 362–368

Schaefer A, Nils F, Philippot P et al (2010) Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for

emotion researchers. Cogn Emot 24(7):1153–1172. https://doi.org/10.1080/02699930903274322
Scherer KR, Schorr A, Johnstone T (2001) Appraisal processes in emotion: theory, methods, research. Series in affective science, Oxford University Press. https://books.google.pt/books?id=IWLnBwAAQBAJ
Schmidt P, Reiss A, Dürichen R et al (2019) Wearable-based affect recognition-a review. Sensors (Basel, Switzerland). https://doi.org/10.3390/s19194079
Shanker MS, Hu MY, Hung MS (1996) Effect of data standardization on neural network training. Omega 24(4):385–397. https://doi.org/10.1016/0305-0483(96)00010-2
Sharma M, Mathew R (2020) Emotion recognition using physiological signals. Lecture Notes on Data Eng Commun Technol 49:389–396. https://doi.org/10.1007/978-3-030-43192-1_45
Smith C, Ellsworth P (1985) Patterns of cognitive appraisal in emotion. J Pers Soc Psychol 48:813–838. https://doi.org/10.1037//0022-3514.48.4.813
Smith CA (1989) Dimensions of appraisal and physiological response in emotion. J Pers Soc Psychol 56(3):339–353. https://doi.org/10.1037/0022-3514.56.3.339
Soleymani M, Lichtenauer J, Pun T et al (2012) A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput 3(1):42–55. https://doi.org/10.1109/T-AFFC.2011.25
Stasieñko A, Sarzyńska-Długosz I (2016) Virtual reality in neurorehabilitation. Postepy Rehabil 30(4):67–75. https://doi.org/10.1515/rehab-2015-0056
Task Force of the European Society of Cardiology (1996) Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task force of the European society of cardiology and the north American society of pacing and electrophysiology. Circulation 93(5):1043–1065
Trojan J, Diers M, Fuchs X et al (2014) An augmented reality home-training system based on the mirror training and imagery approach. Behav Res Methods 46(3):634–640. https://doi.org/10.3758/s13428-013-0412-4
Uhrig MK, Trautmann N, Baumgärtner U et al (2016) Emotion elicitation: a comparison of pictures and films. Front Psychol. https://doi.org/10.3389/fpsyg.2016.00180
Vapnik V (1998) The support vector method of function estimation. In: Nonlinear modeling. p 55–85. https://doi.org/10.1007/978-1-4615-5703-6_3
Varandas R, Lima R, Bermúdez I Badia S, et al (2022) Automatic cognitive fatigue detection using wearable fNIRS and machine learning. Sensors.https://doi.org/10.3390/s22114010. https://www.mdpi.com/1424-8220/22/11/4010
Wioleta S (2013) Using physiological signals for emotion recognition. In: 2013 6th international conference on human system interactions, HSI 2013 pp 556–561. https://doi.org/10.1109/HSI.2013.6577880
Wong TT (2015) Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recogn 48(9):2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009