Multimedia Tools and Applications (2023) 82:19189–19220
https://doi.org/10.1007/s11042-022-13845-5

An automated approach to estimate player experience in game events from psychophysiological data

Elton Sarmanho Siqueira[1] · Marcos Cordeiro Fleury[2] · Marcus Vinicius Lamar[2] · Anders Drachen[1] · Carla Denise Castanho[2] · Ricardo Pezzuol Jacobi[2]

Received: 4 December 2020 / Revised: 18 April 2022 / Accepted: 6 September 2022 /
Published online: 15 October 2022

Abstract

Games User Research (GUR) is a relevant field of research that exploits knowledge on human-computer interaction, game design, and psychology, with a focus on improving the player experience (PX) and the quality of the game. Games form an environment of rich interactions which can lead to a variety of experiences for the player. Researchers employ new ways to assess PX over time with some degree of precision, while avoiding the interruption of gameplay. A possible way of attaining great PX evaluation can be using psychophysiological data. It is a source that can provide relevant details about the emotional states and a potential information in the context of GUR. This paper presents a process for classifying PX in games based on psychophysiological data acquired from the user during the gameplay. Biosensors and a webcam were employed to capture three signals: Galvanic Skin Response (GSR), Blood Volume Pulse (BVP) and Facial Expression. Our artificial neural network was trained with a dataset formed by psychophysiological data and human-annotated emotional expressions derived from assessment and judgment of players' face and behavior with the help of an emotion annotation tool. Four classes of emotions, derived from the most significant game events, are considered for classification: Anger, Calm, Happiness and Sadness. The experimental results indicate that the proposed method leads to good human emotion recognition, and an accuracy score of 64%. The automatic assessment of player experience was compared with a traditional evaluation based on self-report, corroborating the effectiveness of the method.

Keywords Player experience · Psychophysiological data · Games · Biometric sensors · Machine learning · Emotion classification

1 Introduction

The underlying goal of digital games is to provide entertainment for the user through a set of experiences linked to its dynamicity. For example, horror games create situations to

✉ Elton Sarmanho Siqueira
eltonsarmanho@gmail.com

Extended author information available on the last page of the article.

invoke fear experiences to the player [23, 85]. It is the task of the game designer to build game sequences that contribute to the player experiences such as fun, fear, sadness and excitement, among others. For this reason, it is essential that teams of developers can assess whether these experiences have been achieved or not. Research groups in universities and game companies have a growing interest in the emotional and affective aspects of user experience (UX), especially in the digital games field. In particular, user experience in the games industry context is known as *player experience* (PX), which focuses on the qualitative aspects of player interaction with games, considering fun and difficulty, among other factors [61]. In the past, Game User Research (GUR) was often done informally within the game industry, e.g., there were not appropriate specialists for data analysis process, the method of selecting the testers had no specific criterion, and there was no formal instruction guide to apply the playtest. Nowadays, GUR is a formal and structured process with its own set of techniques and methodologies, also composed of experts in the subject, and always finding new methods to improve the quality of the interaction between player and game [20, 23].

Current approaches [2, 14] to assess player experience are largely based on procedures that have been incorporated from other fields (such as psychology, sociology, and neuroscience), adapted to the GUR domain [53]. In particular, there is a large amount of literature on psychophysiology-based emotional experience detection [47, 60, 76, 90] which support researchers and game development teams in applying playtests in a consistent and reliable way, in order to increase the quality of player experience evaluation.

At the same time, the variety of study fields and scientific protocols [47, 60, 76, 90], along with the different ways in which game development teams apply playtests, can affect the process of evaluating the player experience [53].

Traditional evaluation methods have been adopted with a reasonable success rate for evaluating player's experience, and include both subjective and objective techniques [73, 88]. The most common procedures are subjective self-reports, including questionnaires, interviews, and objective reports from observational video analysis. However, these approaches are based only on the subjective responses of the player and hardly capture the player's experiences in real time during gameplay [52]. The player may not remember the emotions and feelings that he/she felt at certain moments in the game session and may forget some relevant information for the game designers.

The limitations of traditional methods fostered the development of new alternatives based on psychophysiological data to estimate the human affective state. Since the physiological state of the human body is influenced by emotions, monitoring changes in this state allow us to infer the emotions that probably caused them. Some physiological signals may be directly obtained from sensors placed over the skin, while others can be inferred by processing its effects. For instance, a smile can be detected by monitoring facial muscles or by image processing, which is an indication of happiness. The sounds produced by a player may also be processed to infer emotional states.

The use of psychophysiological metrics has some potential advantages:

- it allows a spontaneous assessment of the player's experience during the gameplay without breaking the player's immersion, since it is performed in a less invasive manner than other measures, such as interview or direct observation (some people may view theses methods as a difficult or invasive situation) [41, 54];
- if applied successfully, it may allow for more objective measurement of the physiological state of the player during a game session, based on the sensors data [18, 28];

- the psychophysiological signals are involuntary and of continuous nature and, hence, are useful to detect the real experience of the player

Psychophysiological metrics provide some advantages over self-report for better evaluation of player experience [41, 55]. As an example, a well known signal used in affective computing is produced by the Galvanic Skin Response (GSR), which is related to changes in the sweat gland activity that can be measured through the skin conductivity, and is related to the intensity of our emotional state, also known as emotional arousal.

Several works in the literature focused on analysing physiological data to better understand the PX and also to advance in the search to automate the process. The imprecise and subjective nature of the problem is such that no deterministic algorithms are known to solve it exactly, and it is even hard to find heuristics that produces good results. Some works used Fuzzy Logic [49, 64, 78, 81] to cope with the lack of precision in data. Other approaches have adopted linear or partial nonlinear methods (such Support Vector Machine, K-Nearest Neighbor and Principal Component Analysis) to solve the multi-label emotion classification problem [12, 17, 40], but these methods did not work properly on the nonlinear psychophysiological dataset and they did not consider all important aspects of features interactions. In this context, Artificial Neural Networks appear as the preferred alternative in literature [14, 47, 70, 71], being successfully applied to the emotion classification problem under different constraints.

The motivation of this research is to reliably infer player experience in video games using automated, noninvasive methods, to improve and refine game development. More specifically, we aim to investigate new alternatives for automatic player experience assessment. In this work we accomplish the classification of physiological signals with a neural network trained with a dataset annotated by psychologists, aiming to respond the following research question: How can a custom human-annotated dataset impact the evaluation of the Player Experience? We want to discover if the results obtained by psychophysiological methods can be close or equivalent to those of traditional methods. If both approaches produce equivalent results, we can assume that it is possible to automate the process of player experience evaluation.

Figure 1 shows a block diagram of our classification model. It receives three input signals: Galvanic Skin Response (GSR), Blood Volume Pulse (BVP) and Facial Expression. The model is trained with a custom dataset formed by features derived from psychophysiological data (predictor variable) combined with human-annotated emotional labels (predicted variable). The labels were generated from assessment and judgment of players' face and behavior using an annotation tool, including a total of 50 videos, 20 hours, and 3700 game events. To improve the consistency of the human evaluation, each video was annotated by 3 different raters.

In terms of the methodologies used to evaluate player experience, this work is aligned to interests of game researchers and professionals. Thus, our approach considered a number of fundamental issues:

- the experiment must ensure systematic repetition of sessions and reduce external interference;
- noises (or artifacts) should be excluded to improve accuracy;
- all data should be normalized to eliminate inter-subject variations;
- a preprocessing step is performed through a set of algorithms and obtain relevant features to reduce non-informative data;

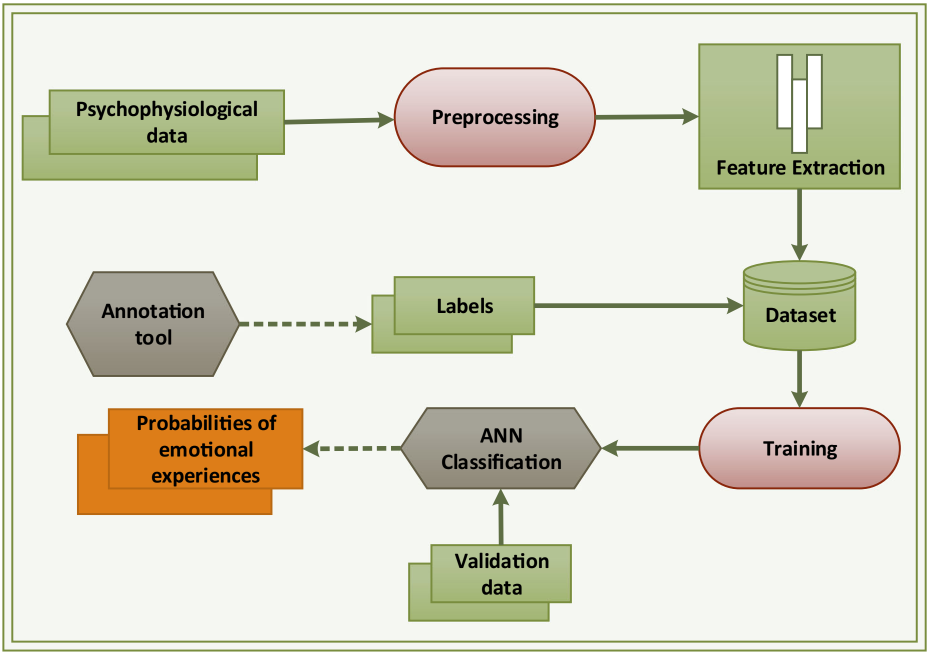


Fig. 1 Overview of our classification pipeline

– the classification model should be tested with a cross validation method to obtain an accurate evaluation of model performance.

This paper is organized as follows. The second section presents some fundamental concepts for this work. The state of art and related works are discussed in section three. The details about the experiments are given in the fourth section, while the fifth section focuses on the creation of the artificial neural network. Section six describes the model validation and the obtained results. At last, the discussion and conclusions sections are given.

2 Fundamental concepts

2.1 Game user research measures

Game User Research (GUR) is a research area that studies how users interact with games, in order to measure real experience of players. Initially, GUR incorporated studies from other fields (Psychology, Statistics, Neuroscience, and so on), which were adapted to develop appropriate tools for better player experience assessment [23]. Classical tools such as *questionnaires* and *interviews* are often used in the context of the game industry and can be applied before, during, or after a playtest session. These two procedures focus on gathering data about player behavior regarding elements contributing to an understanding about the player experience [9, 73]. However, questionnaires and interviews have limitations in the data collection process, such as the difficulty in reporting the player behavior at specific moments in the game, or the inhibition of actual gaming experiences (in this case, the players may not feel comfortable when someone is watching or questioning them) [63]. In order to mitigate these problems, researchers have been creating alternative methods such as *stimulated recall* (video recordings of the players gameplay session to obtain an improvement on visual memory) and *experience graphs* (technique used to support player's memory, where developers ask them to draw a curve showing their experience with game) [69].

However, such methods still depend on the user's participation after the game session. An alternative approach is based on *Psychophysiology*. It consists of a set of methods employed to derive psychological states from physiological measurements, where the most common physiological metrics analyzed are: eye tracking, facial recognition, skin temperature, electroencephalography (EEG), electromyography (EMG), electrocardiogram (ECG), photoplethysmography (PPG) and galvanic skin response (GSR) [41, 57, 60, 77]. The employment of physiological measures to recognize and understand physiological reactions is common in several neuroscience studies [40, 77].

Physiological signals are captured through specialized sensors that are placed on the human body and connected to a computational system for processing. Different technologies are used according to the kind of signal that is being monitored. An important factor to be observed when selecting a device is its degree of interference in the game flow. Figure 2 shows the E4 wristband which is a non-invasive wearable equipment and it is used to capture physiological signals that are strongly related to emotions [27]. E4 wristband consists of a photoplethysmography (PPG) sensor, Electrodermal Activity (EDA) Sensor, and infrared thermopile. From these sensors we can measure: the heart rate (HR), heart rate variability (HRV), inter-beat interval (IBI), blood volume pulse (BVP) and galvanic skin response (GSR).

The EDA sensor provides a signal related to the excitation level of a person when exposed to a stimulus (e.g., a user playing a digital game). It can be measured through the Galvanic Skin Response (GSR)[1] [77], whose value dependents on the sweat glands (eccrine glands) activity. The idea is that changes in the user's emotional states affect sweat production, which leads to changes in the skin conductance. Some authors have shown that GSR is directly correlated with arousal, reflecting emotional responses and cognitive activity [8, 10, 77].

The PPG sensor provides the heart rate (HR) and heart rate variability (HRV) from the Blood Volume Pulse (BVP). As described in the literature [37], blood pressure is a measure of the pressure created to push blood over the arteries, while BVP relates to the amount of blood flowing over the periphery of an individual and a period of time. Since the BVP signal is related to the cardiac activity, it is used to extract the distances between consecutive beats (IBI, Inter Beat Intervals) and then analyze the HRV, related to the regulatory activity of the Autonomic Nervous System (ANS).

Some works consider that facial expressions may be regarded as a psychophysiological measure [79, 84]. Besides, it can be used to infer affective states [84]. Figure 3 presents an example of facial expression analysis. This approach is non-intrusive compared to some other physiological approaches (some authors show that ECG can be intrusive because the electrodes placed on the player's chest can generate discomfort during playtest [84]). According to some studies [71, 79, 84, 85], facial expression analysis can provide relevant results about the player's experience, allowing the collection of data in non-laboratory settings with some degree of precision.

The measures referred in this section are part of what is commonly known as *gameplay metrics* [21]. They show relevant information about the behavior and interactions of the

[1] In older terminology as "skin conductance response".

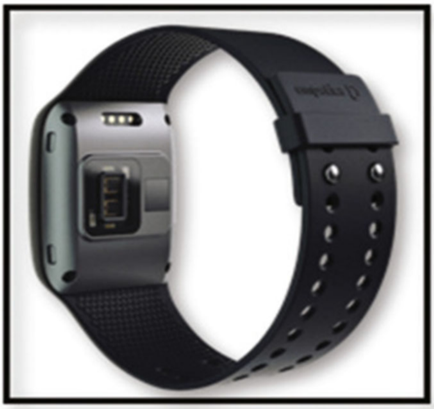


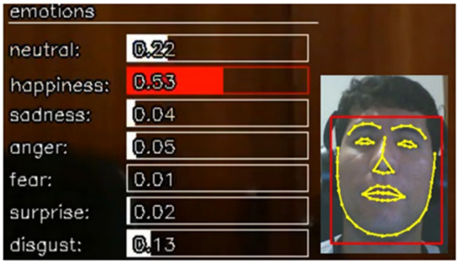
Fig. 2 E4 wristband [1]

players with the game. However, the amount of data provided by sensors in a game session can be huge, which leads to difficulties in analyzing and detecting patterns of player's behavior. Some recent studies [24, 56, 86, 87] show that the visualization techniques of playtesting data can be adopted to support researchers and developers in this task. Graphical representations of gameplay data can help designers and researchers to search for visual patterns of player's behavior over time.

2.2 Emotion models

Emotion, as a research topic, involves several distinct research fields. Although it has been studied for decades, there is no consensus about its definition, basic structure, characteristics and investigation methods [43, 75]. Despite this, there is no doubt about its importance and influence on perceptual, cognitive and behavioral processes [11, 19]. The complexity of emotion as a research topic led to distinct theories and investigation methods [43], and the existence of different emotional experiences resulted in a variety of taxonomy attempts [39].

As a consequence, it is not consensual how to precisely measure and classify emotions and, ultimately, to use this knowledge in the evaluation of systems and products. For example, one of the difficulties in classifying emotion is the use of different meanings and labels by distinct languages from many countries [26, 78, 85]. Two significant models based on dimensional approaches were proposed in the last five decades: the discrete emotion model (proposed by Ekman [66]) and dimensional emotion model (proposed by Russell [68]). The first model emerged from a study conducted by Ekman where a set of participants changed their facial muscles to produce an expression that was coherent with their physiological



emotions

neutral: 0.22
happiness: 0.53
sadness: 0.04
anger: 0.05
fear: 0.01
surprise: 0.02
disgust: 0.13

Fig. 3 A facial expression analysis system

experiences. This study points out six basic universal emotions (happiness, surprise, sadness, anger, fear and disgust) that were defined by author. According to Harmon-Jones et al. [32] - that have mentioned more than thirty discrete emotions - the most commonly listed are those of anger, disgust, fear, anxiety, sadness, happiness, relaxation, and desire. The second model, known as Russell's Circumplex Model of Affect, suggests the use of a two-dimensional distribution of the emotion types organized by their valence and arousal values. The valence concept refers to the evaluation of the positive or negative effect of the emotion, also represented as pleasure or displeasure. The arousal concept can be comprehended as the intensity of the emotion, varying from a relaxation (or sleepiness) condition to an arousal (or tension) state [45, 78]. Both have often been applied to help understand specific emotions and are currently used in conjunction with methods to infer psychological data. [2, 50, 74].

The discrete emotion model was adopted by the present research with the emotion class labeling scheme by evaluators who use participant data, such as facial videos, to determine their emotional state through their interaction during the game session. [76]. This scheme facilitate use of Multi-label emotion classification using machine learning, as seen in studies [2, 16].

2.3 Neural network

According to Haykin [33], artificial neural networks (ANN) are mathematical models inspired by the mechanism of the human neurological system, with the capability to learn non-linear functions. They are used to provide outputs for an input parameter set without requiring to know specific models or functions. The ANN architecture consists of a simple element known as *Perceptron* (as shown in Fig. 4).

A perceptron is a single layer neural network. More complex problems can be processed by increasing the number of layers in the network, resulting in a structure known as *Multi-Layer Perceptron* (MLP). The basic MLP has at least three layers: the input layer, the hidden layer, and the output layer. Each layer produces an output used as an input for the next layer. The output layer provides the result according to the function learned by the ANN [33, 36].

3 Related work

In game user research, psychophysiological data have been intensively studied by academics and game development teams as an important metric for studies on player interaction with games. Studies by Kivikangas [41], Nacke [63] and Soares [77] present a literature review

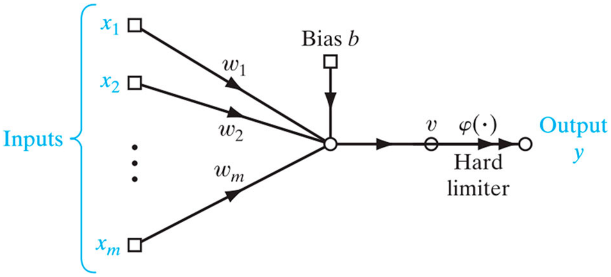


Fig. 4 Overview of Perceptron [33]

on physiological data telemetry in the game user research domain. They show advantages and limitations about collecting, treating and visualizing these data.

There is a significant body of work on the quantification of emotional states during a game session. In particular, the work of Mandryk and Atkins [49] discusses a procedure that uses a complex fuzzy logic system that processes ECG, EMG, and GSR signals and produces arousal and valence as the system output. Another relevant work was developed by McMahan et al. [55] which evaluates many aspects of PX, considering isolate game events (e.g., death of a character) using the Emotiv EEG device. Isbister and Schaffer [38] describe the use of Electromyography on the face to measure the emotional state of valence in a game session. In this same line of research, we find important statistical studies, such as: i) Yannakakis et al. [91] that correlated physiological data and subjective data of emotional components of the player's experience; ii) Tognetti et al. [80] who used physiological data to recognize user enjoyment in a car racing game; iii) Drachen et al. [22] that showed a study on the correlations of Galvanic Skin Response and Heart Rate with the player's experience in a First-Person Shooter game.

The authors in [34, 50, 64] explored the correlation between physiological changes (using GSR, ECG and EMG) in a game under different settings, and they discovered some important points, such as players who experience different emotional states while playing the same game with different settings.

The study done by Susana et al. [78], aims to infer players' emotions from an emotion model using some artificial intelligence techniques, e.g., Fuzzy Logic and Adaptive Neuro-Fuzzy System (ANFIS). Regarding the emotion model, the authors used Russell's affective grid to collect data from participants, where each of them was asked to mark an "X" where they felt that their emotions were best represented in the grid. Certainly, this procedure is a quick way to assess the emotional state considering the dimensions of the Arousal-Valence (AV) space. Although this procedure is relatively easy to conduct (Affect Grid), it relies on player's ability to recall and explain their experience, which can lead to problems in understanding the players' gaming experience [57, 60].

In the work done by Chang et al. [16], the authors created an emotion recognition system considering facial expression and three physiological signals (GSR, hear rate and skin temperature). Based on those signals two learning vector quantization (LVQ) neural networks were applied to classify four emotions: love, joy, surprise and fear. A predefined set of specific emotion induction events through videos was executed to obtain facial expressions and physiological data for classification from six subjects. However, this study uses a set of psychophysiological data derived from induced events using a small sample of participants. Considering that the variability among individuals can impact data quality, in our work we propose to collect and treat psychophysiological data using a larger sample of participants.

In [70], the authors contribute to the development of an automatic game stream emotion annotation system that combines neural network analysis of facial expressions, video transcript sentiment, voice emotion, and low-level audio features. Specifically, the authors used videos of users' faces to obtain data to assess emotional valence. For the classification process, a human-annotated emotional expression dataset was built by two different human annotators with 70% inter-rater agreement between them. Also, the authors selected the five most intense events of each video for analysis and reached good precision in detecting clearly positive emotions like amusement and excitement, but had a less precise result with subtle emotions like puzzlement. However, the study done by Kramer et al. [42] shows that increasing the number of raters led to an increase in the resulting reliability value, which is important to ensure the quality of the dataset. For this reason, the present research used eighteen evaluators, divided into groups of three, that annotated the videos with labels that indicate the estimated emotion of the player, based on his/her facial expression.

In another work, Mirza-Babaei et al. [58] present a tool that combines EMG sensor with video observation and gamelogs. This tool visualizes players' emotional reactions over some pre-defined game events (for example, events that game designers or developers are specially interested in reviewing or analyzing). Selected game events are shown to the playtesters to insert some additional information.

In [90] Yang et al. evaluate psychophysiological data related to the game events together with a database that contains biometrics data (GSR, ECG, EMG, respiration and temperature), facial and screening recordings. The authors perform player's self-reported event-related emotion evaluation during the game. They used some machine learning techniques to emotion detection and recognition with different segmentation lengths of time. The SVM (Support-Vector Machine) technique achieved the best average accuracy with 65% using 10 seconds as the time segmentation lenght. Similarly, Granato et al. [29] show a system to collect physiological data (electrocardiogram, electromyography, galvanic skin response, and respiration rate), store, visualize, and synchronize them, along with a tool to self-assess the players' emotions. The tool uses values of arousal and valence to create a dataset based on the emotional response of players. Considering the Mean Square Error (MSE) index between the average value acquired by the tool and the survey answers by players, the study reached 0.26 for arousal, while 0.28 for valence. Maier et al.[47] propose a specific method to automatically measure the player's flow using physiological data (GSR and BVP) from a wrist-worn device. The method is based on a convolutional neural network (CNN). The validation of the model was done using cross-evaluation and compared to existing methods for flow classification, based on self-report. The model allows to estimate whether the player is on the flow with 67.5% of accuracy, and could identify the player affective state for a small set of options (flow, bored or stressed) with 49.5% of accuracy.

Chanel et al. [15] conducted an experiment to discover three discrete emotional states (boredom, engagement and anxiety) from twenty players during playing Tetris game using some physiological signals (EEG, Temperature, GSR, HR and BVP). They achieved 59% classification accuracy using Fast Correlation Based Filter for Feature Selection and QDA (Quadratic Discriminant Analysis) classifier of peripheral features. Chanel and Lopes [14] using deep neural networking with one physiological signal (EDA) obtained emotion recognition accuracy of 73.2% using Tetris game. Similarly, AlZoubi et al. [2] selected a survival game for detecting six discrete emotions (happy, fear, neutral, anxiety, excited and angry) from 12 players. They recorded EMG, EDA, BVP, RESP and ECG for testing the effectiveness of individual physiological channel features and a fusion feature from all physiological channels. AlZoubi et al. (2021) reported that ECG was the best channel with an accuracy of 63% using XGBoost classifier. Using a fusion feature, XGBoost achieved accuracy with 66.15%.

Both authors Kharat [81] and Unluturk [82] conducted investigations about the analysis of facial expressions. In particular, the former author employs Fuzzy logic while the latter uses Neural Networks. Also, a relevant contribution to the analysis of facial expressions was developed by Vieira [84]. He explored the feasibility of evaluating different emotions from the computational analysis of facial images captured with a low cost device (web cam). The data was collected through videos recorded from the faces of the participants while they played three different games. It combined two techniques for the detection of emotions, the Viola-Jones face detection algorithm and the Active Appearance Model algorithm for tracking the facial landmarks.

Our work aims to contribute to the automation of player experience classification. The task is a challenge since it involves several fields of research, including human biology and behavior. Among the physiological metrics used to estimate the affective state of the player, the blood volume pressure (BVP) and the galvanic skin response (GSR) are widely used in the evaluation of the arousal level of an individual [44, 51]. Fewer options are available to find the valence, i.e., the positive or negative aspect of the emotion. Face expressions are, in this case, the most relevant signal for that purpose and is commonly probed with electromyography (EMC) or image processing. Considering that immersion in the game may be affected by the degree of physical interference of the sensors, analysing the face expressions with image processing seems a suitable approach, which can be performed with normal webcams. BVP and GSR can also be non-intrusively measured with a wristband, with little interference on the player mechanics. Some studies ([14, 47, 84]) about emotion recognition succeeded using only a data source such as facial expression or physiological signal. However, other studies ([2, 90]) reported that multiple data sources succeeded in improving the performance of emotion recognition.

Mapping biometric measures to emotions is a complex task, because there is not a one to one correspondence between physiological signals and emotions. One emotion may trigger several biometric signals and, on the other hand, one signal may be affected by different emotions. One common approach is to adopt a two phases process. First, physiological signals are used to obtain Arousal and Valence estimations. Second, these data are mapped to discrete emotions in the bidimensional space (e.g. Russell's circumplex model of emotion [68]). It is debatable to argue that a two-step process can achieve better results, as each step introduces approximations and combining them will possibly accumulate an estimation error. In our approach we adopt a single processing step by mapping psychophysiological signals directly into emotional states. Considering that there is no algorithm that solves this mapping exactly, we resorted to machine learning, using neural networks for that end. To improve the results we opted to build a dataset that includes the evaluation of facial expressions and the psychophysiological signals of the players. Using an annotation tool, several psychology students evaluated the mood of the players based on videos, producing information that contributes to build a dataset used to train the neural network. Validation was then performed with a different set of players using the classification model. The next sections detail this process.

4 The experimental setup

In order to build and evaluate our classification model, we designed an experimental setup based on a racing game called TORCS (The Open Racing Car Simulator) [5] (Fig. 5). It was selected for the following reasons:

- it is a game with a rich gameplay mechanics in which the players experiment emotionally different conditions;
- it is an open source project, so it is possible to implement new functionalities;
- the game is simple, even for a player having little familiarity with the genre, thus the game experience can be held as uniform as possible among participants involved in the research.

The player experience is correlated to his/her ability to overcome challenges introduced by the game. A game that is too easy to play can became quickly boring, while a game

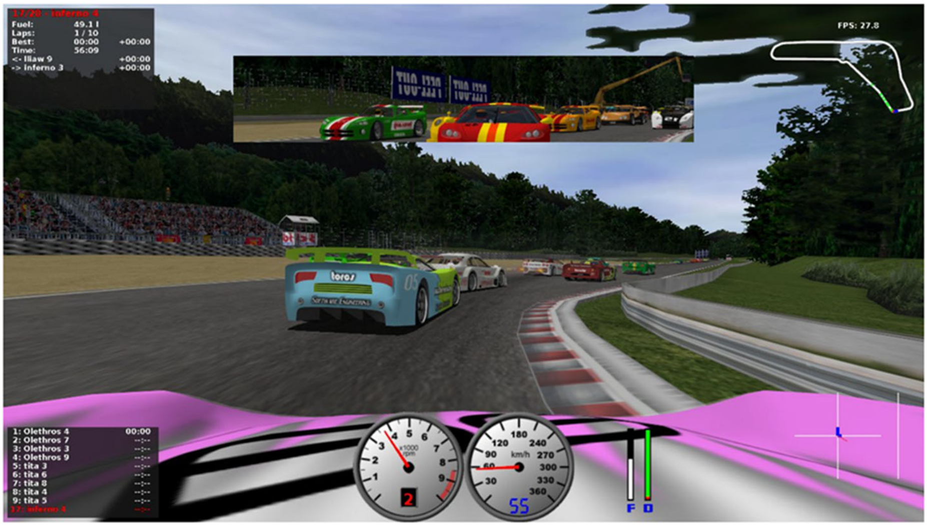


Fig. 5 Screenshot of TORCS [5]

that is too difficult turns out to be very frustrating. In order to examine how game settings can influence the players' experience, we conducted a within-subject study in which participants played TORCS under some configuration conditions at random. For that, we added two configuration parameters on the race: *drift* (settings: Enable or Disable) [2] and *gearbox* (settings: Manual or Automatic). Using an automatic gearbox makes driving easier. On the other hand, a manual gearbox allows greater control over the car. Drifting also introduces a new alternative in driving that may be explored by the players. Both parameters allow to provide a good realistic car handling model [31].

All game sessions were held at the University of Brasília - Department of Computer Science, on working days, lasting approximately 30 minutes. After a brief introduction including information about the playtest, the participant was invited to sign a consent form with the purpose of the experiment, its rights and how the collected data would be processed and stored. Before starting the experiment, participants completed a questionnaire, used to gather information about their experience with racing games.

The experiments were divided into three phases, as indicated in Fig. 6. In the first phase, psychophysiological data was collected from a set of participants together with videos of the players and the game session. For the second phase of the study eighteen volunteers were recruited and divided into groups of three to analyze and judge the videos of the participants' faces. Based on the raters responses and the psychophysiological data, we apply a set of procedures to create our dataset (details are given in section *Dataset*) and thereafter we build and train our artificial neural network (ANN) model. In the third phase, we collected data from a set of new participants and compared the self-report results generated from an annotation tool with ANN estimated results in order to validate the research.

[2] Drifting implies traveling through tight corners in over steering, the rear wheels without traction, and the front wheels pointing in the opposite direction to the turn [30]

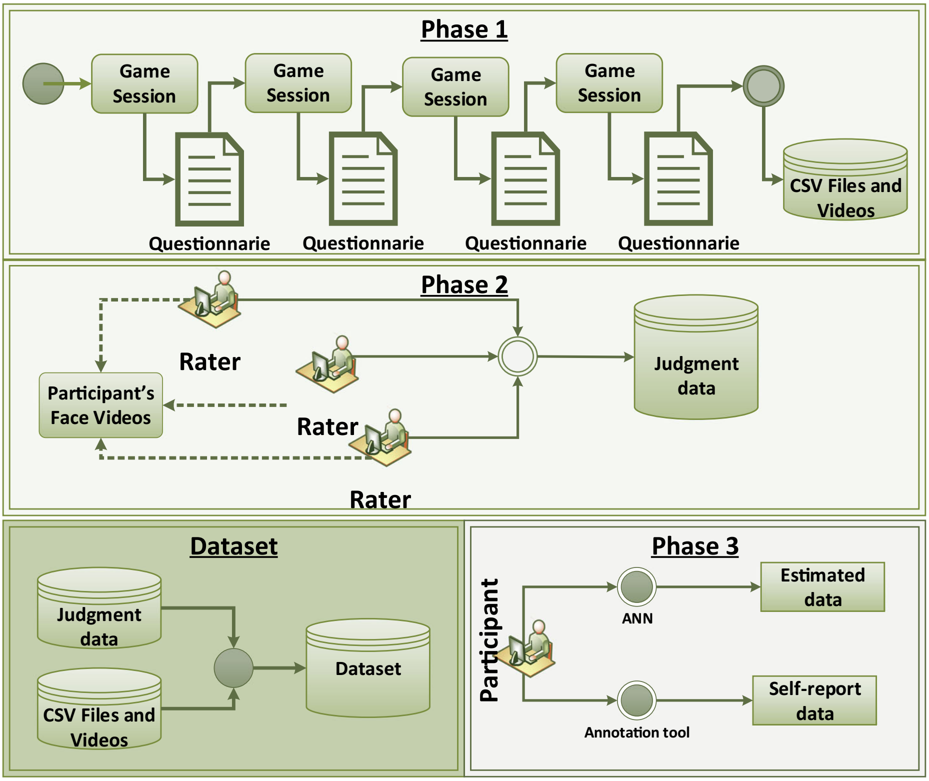


Fig. 6 The phases of the experiment

4.1 Participants

The participants in the experiments were young (age ranging from 18 to 32 years, average age of 23.25 years and standard deviation of 3.36) volunteers recruited from a university mailing list and social networks (Facebook). In addition, participants' characteristics ranged from college students to game developers. Specifically, in the second stage, sixteen psychology students and two psychologists [3] were recruited to analyze and judge the videos of the participants' faces. To better understand the distribution of the participants in the phases, see Table 1.

4.2 Procedure

At first, participants sat in a chair while the E4 wristband was placed on their non-dominant hand. Participants were asked to play a warm-up session that's about 1-2 minutes. Finishing this session, they were then instructed to remain calm for four minutes while a baseline for physiological measurements was being recorded. After the rest period, participants were

[3] The psychologists that participated in this research work in the area of social psychology and they have experience in the construction of personality measures and statistical analysis of behavior.

Table 1 Number of participants in the the experiments

Phases of the experiments	Total	Completed	Males/Female
1	35	30	20/10
2	18	18	4/14
3	20	20	18/2

instructed on the experiment. They played each game session for approximately five minutes. To reduce the potential of carryover effects affecting the data collection, we asked the participants to avoid getting up from the chair. Then they played four game sessions with different settings, as showed in Table 2.

Regardless of the player's skills, the second parameter of the race configuration (gearbox) could be considered indeed challenging, and more attractive for the player (for some participants, the first parameter made no difference during the game session). Race parameters such as type of track and number of opponents were selected to allow the game to have two difficulty levels (easy and difficult). Only one car model was used in all game sessions. It is important to note that the order of the game sessions was randomized for each participant in order to reduce the learning effects of the game.

The E4 bracelet was placed on the participants to measure peripheral physiological activity (GSR and BVP). The participants were asked to wear a headphone to ensure a good game involvement through race sounds. After this step, videos (Video 1: player's face and Video 2: game screen) and physiological signals acquisition were started while the player was playing. The participants were instructed to reduce movements during gameplay to reduce the introduction of noise in the acquired images, which may interfere in the facial expression analysis. The experimental protocol was carried out by an automatic script on the computer that started each game session and synchronized all data sources. After each session, participants' subjective experience was assessed through a game experience questionnaire. The data for each game session was exported to a CSV file.

Psychophysiological data (BVP, GSR and facial expressions) were recorded during the game sessions for all participants (phase 1 and 3 of the experiment). Galvanic Skin Responses and Blood Volume Pulse were measured using a Empatica E4 (Fig. 2). In addition, the face of the participants and the computer screen were video-recorded for later analysis of facial expressions (see Fig. 7). Although informative, the early experiments had some logistical problems: incorrect use of the sensor and some participants (five people in phase 1 and none in phase 3) stopped the experiment. Consequently, we excluded those with noisy data from data analysis. The total time of a single experiment was about 30 minutes, divided in 20 minutes (4 sessions x 5 min) of racing and about 10 minutes of setup questionnaire answering. On successful completion of experiment we can observe the entire evolution of the player's behavior during the game session, as can be seen in Fig. 8. Only

Table 2 Experiment setup

Session	Drift Skill	Gearbox
1	Disable	Automatic
2	Enable	Automatic
3	Enable	Manual
4	Disable	Manual

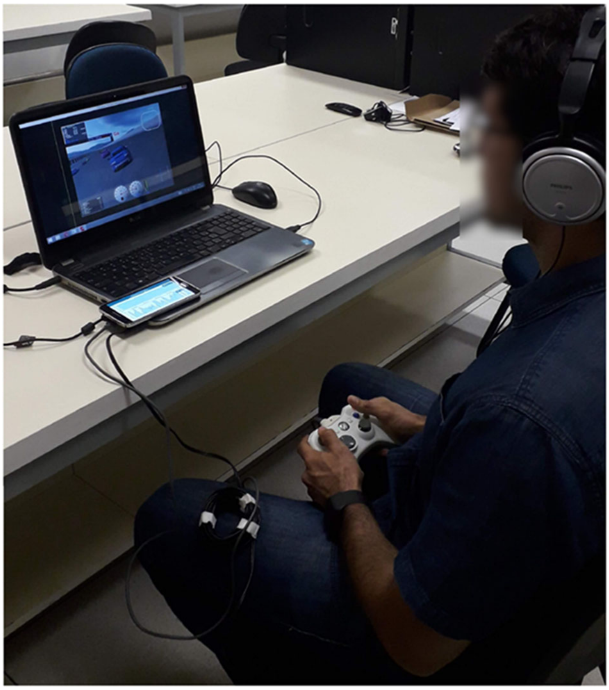


Fig. 7 A participant taking part in a game session

in phase 3, the participants self-reported their emotional experiences through an annotation tool (Fig. 9), a task which took about 25 minutes to complete.

5 DATASET

This section describes the second phase of the experiment. It focuses on the preparation of our dataset. We created an emotion expression annotation tool that allows for judging emotional expressions according to a video of the participant's face (see Fig. 9 (A)). In addition, this tool allows to view the gameplay video and record game events (Fig. 9 (C)) of a game session. We have chosen to develop an intuitive tool based on the studies done in [25] and we extend Ekman's basic emotions [66], because they are too limited for the range of players' emotional expressions.

During this phase the evaluators viewed the recording of match and annotated the emotions triggered by significant events during the game. To make the annotation easier, we offered evaluators a list of emotions[4] and events. The elements included in each annotation are:

- Emotions: anger, mad, pissed off, rage, grossed out, revulsion, sickened, nausea, terror, scared, fear, panic, worry, anxiety, dread, nervous, lonely, grief, sad, empty, wanting,

[4] We have included a set of discrete emotions as used in [32] to provide more options for the evaluators.

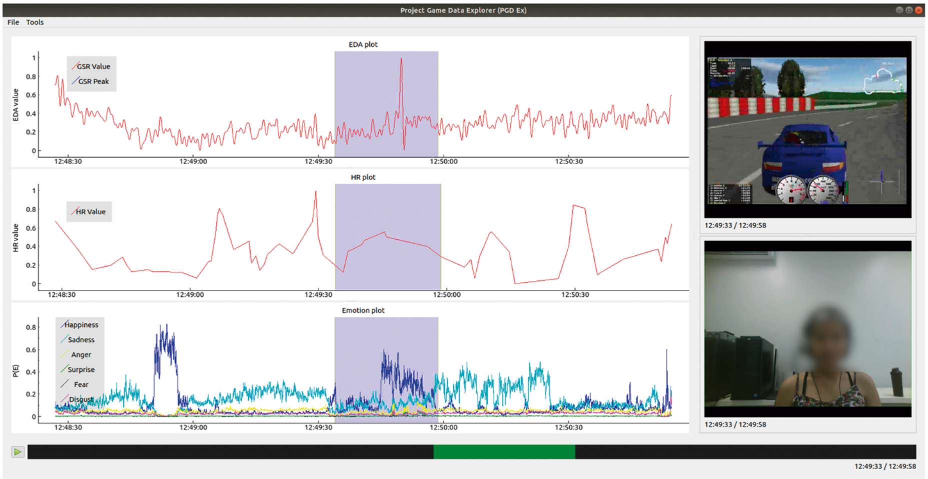


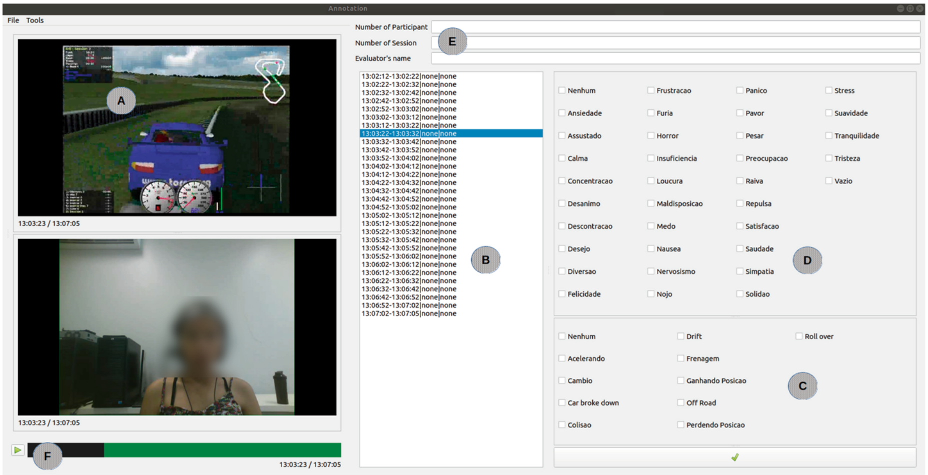
Fig. 8 The 2D visualization of psychophysiological data over game session time for a player

craving, longing, desire, calm, relaxation, chilled out, easygoing, happy, enjoyment, satisfaction, liking

- Events: roll over, off road, collision, speeding up , gearbox, car broke down, drifting, braking, overtaking, overtaken

After reviewing all videos, the following game events were selected for analysis:

- Gearbox: the gearbox can be set to manual or automatic mode. The events related to the automatic gearbox are to engage reverse gear or to engage forward gear. In manual



```
13:02:12-13:02:22|none|none
13:02:22-13:02:32|none|none
13:02:32-13:02:42|none|none
13:02:42-13:02:52|none|none
13:02:52-13:03:02|none|none
13:03:02-13:03:12|none|none
13:03:12-13:03:22|none|none
13:03:22-13:03:32|none|none
13:03:32-13:03:42|none|none
13:03:42-13:03:52|none|none
13:03:52-13:04:02|none|none
13:04:02-13:04:12|none|none
13:04:12-13:04:22|none|none
13:04:22-13:04:32|none|none
13:04:32-13:04:42|none|none
13:04:42-13:04:52|none|none
13:04:52-13:05:02|none|none
13:05:02-13:05:12|none|none
13:05:12-13:05:22|none|none
13:05:22-13:05:32|none|none
13:05:32-13:05:42|none|none
13:05:42-13:05:52|none|none
13:05:52-13:06:02|none|none
13:06:02-13:06:12|none|none
13:06:12-13:06:22|none|none
13:06:22-13:06:32|none|none
13:06:32-13:06:42|none|none
13:06:42-13:06:52|none|none
13:06:52-13:07:02|none|none
13:07:02-13:07:05|none|none
```

Fig. 9 The User Interface of the Annotation Tool. (A): The video display (participant face or gameplay video). (B): Time window. (C): The game event options (in Portuguese) for the session. (D): The emotion options (in Portuguese) for the session. (E): Text input to identify the evaluator, session and participant. (F): Controls for precise playback with a mouse

mode, the selected events are to use reverse gear, slow down with engine braking, or shift to overtake.
- Collision: events includes collision with game objects (vehicle, wall, tree, among others)
- Drifting: the event occurs when the car is sliding sideways while making a turn. It is generally a pleasant experience, but requires driver's control and skill.

The structure of this dataset corresponds to a set of input features associated with output labels. Table 3 illustrates it for a few inputs. The data is acquired continuously over time. To convert continuous time to discrete samples, the data was divided into time windows of ten seconds, each window constituting of a single data point. The input features for each window comprise various features of signals from the face, GSR and BVP (described in next section). The output label (column called "Emotional Experience") of a window is either "NONE" or the most frequently occurring emotion.

The experiments were analyzed with two levels of granularity:

- First level with 35 emotion classes;
- Second level with 8 emotion classes grouped into Anger, Disgust, Fear, Anxiety, Sadness, Desire, Calm and Happiness as described in [32].

The game sessions were examined by the evaluators for both levels of granularity. The inter-rater agreements were calculated with the Fleiss' kappa coefficient. Only game sessions with the coefficient greater than 0.6 were selected for analysing and processing. As expected, using less emotion classes resulted in larger inter-rater agreement, as can be seen in Table 4. The raters agreement reached 81% for session 1, with 8 emotions classes. The four sessions average agreement in this case was 69.5%. With 35 emotion classes it is very hard to achieve good agreement rates. The number of samples on happiness, calm, sadness, and anger was 210, 180, 170, and 100, respectively. Disgust, Fear, Anxiety and Desire were removed due their very low occurrence. Our dataset, then, used only four emotional experience labels with a moderate degree of unbalance.

5.1 Preprocessing and feature extraction

The sampling of the three analysed signals (BVP, GSR and facial expressions) produces a large amount of data. Having fifty participants (phase 1 and 3 of the research) playing the game for 1000 minutes, there was 800 minutes of gameplay/psychophysiological recordings (excluding erros), 600 minutes of recordings of the self-reports (with interviews, questionnaire with closed questions about the session, and self-evaluation using the annotation tool). The overview of the preprocessing process (reduction and extraction) involved for each type of data can be seen in Fig. 10.

As can be seen in the diagram in Fig. 10, all data is preprocessed using normalization, filtering and segmentation techniques. Normalization is needed since physiological signals are subject of high variability among individuals. Filtering reduces noise from the signals and the segmentation step breaks data into time windows. Afterward, the most significant features are extracted and reduced using the Principal Component Analysis (PCA) (in case of GSR and BVP) and Factor Analysis methods (in the case of facial expressions). The data from the first experiment, as well as the data from the third experiment, were treated in the same way. In addition, the psychophysiological features were are calculated in each data window using the time domain values. This window size has been found to be long enough to omit disturbances and to allow for the discovery of meaningful changes in human

Table 3 An example of a sample showing the entry for a 10s interval of dataset, with game event, emotion labels by raters, emotional experience and feature vector

Player	Session	Interval Initial	Interval Final	Game Event	Rater 1	Rater 2	Rater 3	Emotional Experience	Features
1	1	1567638246.95	1567638256.95	Colision	Anger	Anger	Anger	Anger	....
1	1	1567638256.95	1567638266.95	Drifting	Calm	Calm	Joy	Calm	....
....	....	....	....	....	....	....	....	....	....

Table 4 Inter-rater agreement and with respect to different levels of granularity

Inter-rater agreement (Mean)				Number of Classes
Session One	Session Two	Session Three	Session Four	
0.33	0.22	0.29	0.17	35
0.81	0.69	0.63	0.65	8

behavior [92]. The preprocessing of GSR, BVP and facial expressions is performed as follows.

- **Galvanic Skin Response**: We use a convolution filter to remove noise. In addition, the GSR intensity has high individual variability, making it impossible to directly compare the subjects. Thus, we used the most common procedure to normalize the GSR intensity (1 [62]).

$$Signal_{normalized} = \frac{Signal_t - Signal_{\min}}{Signal_{\max} - Signal_{\min}} \tag{1}$$

  We extracted the phasic response from GSR signal, related with the sympathetic nervous system (SNS) [3]. The GSR features are defined based on time domain, thus we extracted and analysed both time-based and statistical features (Table 5). For more details about GSR features, see [4].
- **Blood Volume Pulse**: We used the Reverse Combinatorial Optimization (RCO) algorithm [7] to identify corrupt signals and abnormal outliers, which are then removed. The normalization of BVP signals is computed in the same way as for GSR signals. In addition, statistical and time domain features of BVP are extracted for analysis (Table 5 lists BVP features extracted in this study). More details about them can be found in [48].
- **Facial Expressions**: The analysis of facial expressions was performed with the help of a tool developed by Viera [84]. A face tracking mechanism based on the Viola/Jones algorithm collects face images from the video of the players and a variation of the Active Appearance Model (AAM) algorithm is used to track facial landmarks. A SVM then provides estimations of prototypical emotions (*Sadness*, *Fear*, *Surprise*, *Anger*, *Happiness* and *Disgust*), which are converted to two basic signals representing the positive and negative aspects of the detected emotion. For more details about the facial expression analyzer refer to Vieira [84].

  Figure 11 shows the data distribution of prototypical emotions collected in all game sessions. We found that some emotions have the same behavior between sessions.

  We applied ANOVA (Analysis of Variance) on each emotion, to determine if there are statistically significant differences among the means of the sessions. As a result, "Fear" ($F(3, 717161) = 0.06, p\text{-}value = 0.98$), as well as "Surprise" ($F(3, 717161) = 0.015$, $p\text{-}value = 0.99$) have the same average across all sessions. As shown in Fig. 11

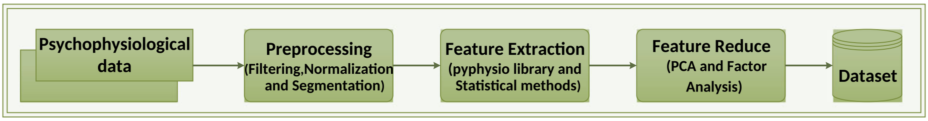

Psychophysiological data
Preprocessing (Filtering,Normalization and Segmentation)
Feature Extraction (pyphysio library and Statistical methods)
Feature Reduce (PCA and Factor Analysis)
Dataset

Fig. 10 Block diagram representing the data acquisition and processing chain

Table 5 Features Extracted from the psychophysiological signals

Signal	Features
GSR	Mean, Range, Standard deviation, Mean peaks amplitude, mean slope,
	Maximal peak amplitude, Peak slope, Peak duration, Number of peaks
BVP	RRmean, RRSTD, RMSSD, pNN50, pNN25,
	pNN10, triang, TINN, SD1, SD2, SD12, DFAa1, DFAa2
Facial Expression (Positive and Negative)	Maximum, Minimum, Mean, Range, Standard deviation

through the Violin plot[5], both have a small data density. Thus, for analysis purposes, we removed “Fear” and “Surprise” from this experiment.

In addition, to reduce the number of dimensions (or features) in this set of prototypic emotions without losing much information, we have chosen to use **Factor Analysis** (FA) which is a fast, robust and accurate method (for details on FA, see [13]). The process of dimension reduction established the Kaiser criterion (KC) and Kaiser-Meyer-Olkin (KMO) Test as the main approaches. The first is based on the most significant proportion of explained variance by the factor that will be selected, since the “eigenvalue”[6] is an excellent indicator for determining the number of factors. In general, an $eigenvalue > 1$ will be considered as selection criterion for the features. Figure 12 shows the visual representation of factors' eigenvalues and the quantity of factors is two. While the second (KMO) measures the suitability of data for factor analysis. It defines the adequacy for each observed variable, checking the quantity of variance among all the observed variable. Thus, KMO values range between [0, 1], and values less than 0.6 are considered inadequate. In this study we found a KMO of 0.75.

Based on FA results, we combined “Anger”, “Disgust”, and “Sadness” as a **Negative Emotion**, and the prototypical emotion “Happiness” is called a **Positive Emotion**. Finally, statistical features of the factors are extracted for analysis (see Table 5).

After the features extraction, some of them have almost the same values in the dataset. These features are not helpful to classify the player experience. In the feature extraction process described above, some features from the signals are strongly related. Some researchers [65, 83] have applied feature selection algorithms as PCA to filter out these redundant and strongly related features. Using PCA on the dataset we removed 13 features of a total of 32 features. The dataset was constructed by these features and emotional experience labels (annotation process).

6 The Ann architecture

Our ANN consists of a Multi-Layer Perceptron with 183 neurons distributed in four layers (two of them are hidden layers) and the number of input neurons corresponds to the number

[5] The Violin plot performs a similar function as histograms and box plots. It presents a distribution of quantitative data on several levels of one or more variables such that those distributions can be compared.

[6] It expresses explained variance by each factor from the total variance [13].

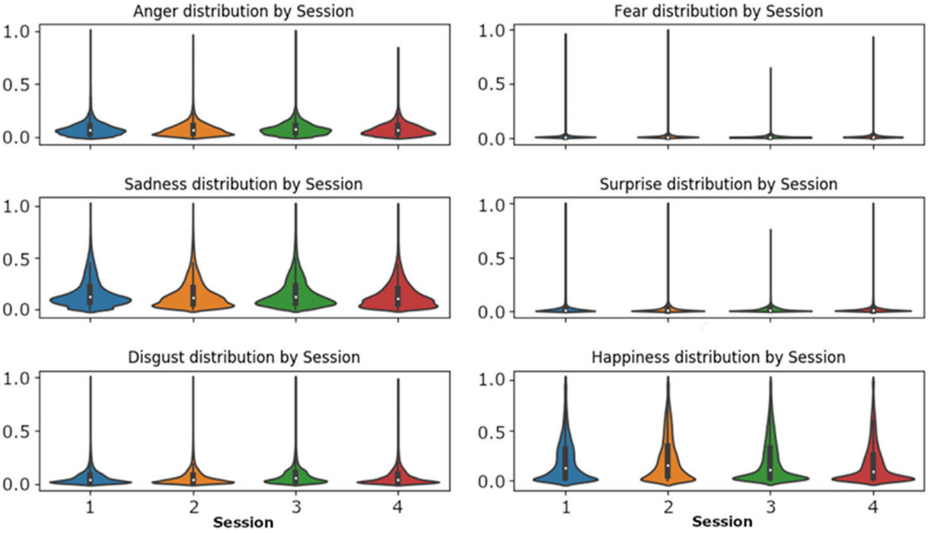


Fig. 11 Prototypic Emotion Distribution by Session using the Violin Plot

of variables. In order to have a correct estimation of performance our model, we applied 10-fold cross validation process. The ANN architecture is described as follows:

- During the experiment, different numbers of hidden layer as well as different amounts and distributions of neurons were tested. The best results were achieved with two hidden layers with 80 neurons in each hidden layer. Table 6 shows the different ANNs analyzed in this work.
- Training and test sets are necessary for building an ANN. The training set is employed for the ANN model development and the test set is used for evaluating it. Based on the

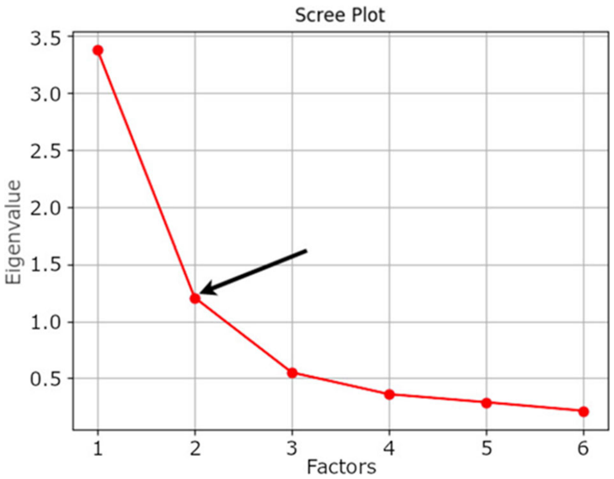


Fig. 12 The scree plot supports the choice of the number of factors

Table 6 The different ANNs tested using 10-fold cross validation

Number of hidden layers	Learning rate	Neurons in each hidden	Overall accuracy
1	0.1	10	39%
2	0.1	10-10	43%
2	0.01	25-25	55%
2	0.001	35-35	57%
2	0.001	70-70	60%
2	0.001	80-80	64%
3	0.01	55-55-55	58%
3	0.001	70-70-70	61%

study described in [46], in our experiments we used 75% of the data for the training set and 25% for the test set.

- The learning procedure was implemented using the backpropagation algorithm. It requires two training parameters: (i) learning rate and (ii) momentum. In this study, the learning rate was set to 0.001 and the momentum factor was 0.9. The training stopped after 1000 learning epochs.

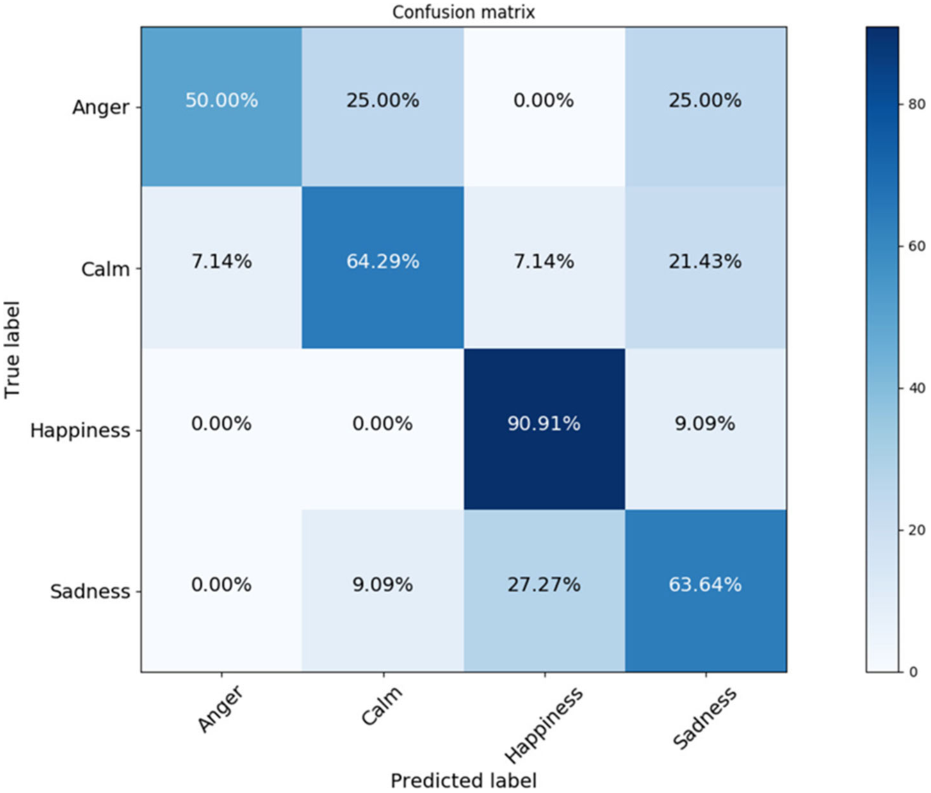


Fig. 13 The confusion matrix for the classification model

- The training function for weight optimization of the network was SGD (Stochastic Gradient Descent) and the activation function was the hyperbolic tangent (for the hidden layer) $\varphi(x) = tanh(x)$.
- In the output layer, the number of neurons is defined by the number of outputs of the system. In this case, four neurons were needed to provide estimates in terms of probabilities of the emotional experience associated to the player experience, measured as: calm, happiness, anger and sadness .
- In Fig. 13, we show the confusion matrix (for details on it, see [89]), it was used as a mechanism to validate the classification performance. For the test set, the lowest accuracy was evaluated for "Anger" with 50% while the highest accuracy was achieved for "Happiness" with a 90% of accuracy. Other metrics are reported: Precision (0.71), Recall (0.70) and F1-Score (0.69). Lastly, we used the chi-squared statistic to determine whether our classification model performs well, as result we have ($\chi^2 = 2.27$, $p\text{-}value = 0.13 > 0.05$). In particular, if $p\text{-}value > \alpha$, we can trust that our expectations match actual data well.

7 Model validation

To analyze the effectiveness of our model of classification, we used data gathered from the twenty participants not considered in the generation of the model. Data were preprocessed using the previously described methods. Self-report data was produced through the annotation tool (Fig. 9). The model and self-report results were compared to validate this research.

Our results show the emotional experience of the participants for the most relevant events of the game sessions. We compare the self-reports with the emotions estimated by the model to determine the degree of concordance between them. We defined the **Score** indicator as a percentage of the participants emotions for both the self-report and for the model estimation. We selected "Collision", "GearBox" and "Drifting" as the investigated events. To determine how accurately the estimated data (the answers produced by our model) resembled the data reported by the participants, we used the Spearman correlation coefficient (ρ)[7] to correlate the two data samples for each game event. We also examined the player experience questionnaires.

Figure 14 shows the results for "Collision" events. Some participants kept calm under this event, while others reported anger, indicating that the "Calm" scores, as well as the "Anger" ones, are in agreement with our model. The percentage of players that reported "Calm" was 30.46%, and the model estimation was 29.44%. For "Anger", the reported value was 17.88% while the estimated value was 16.56%. Results for "Sadness" were somewhat smaller, with reported and estimated values of 38.41% and 32.07%, respectively. However, the model identified too much occurrences of "Happiness", resulting in a low estimation accuracy for this emotion. In fact, some players clarified that ironic smiles were mistakenly classified as expressions of happiness in collisions, leading to an increase in the detection of "Happiness" in our model. This incorrect classification had a strong effect on the correlation coefficient ($\rho = 0.80, p\text{-}value = .2$). Indeed, excluding this extraneous behavior and recalculation ρ for the other emotions we reach a relatively high correlation score ($\rho = 0.99$, $p\text{-}value < .05$). The identification of ironic smiles could be an interesting research topic for future works.

[7] The Spearman correlation is a nonparametric measure of the monotonicity of the relationship between two samples. It is used for the data that are not normally distributed [59].

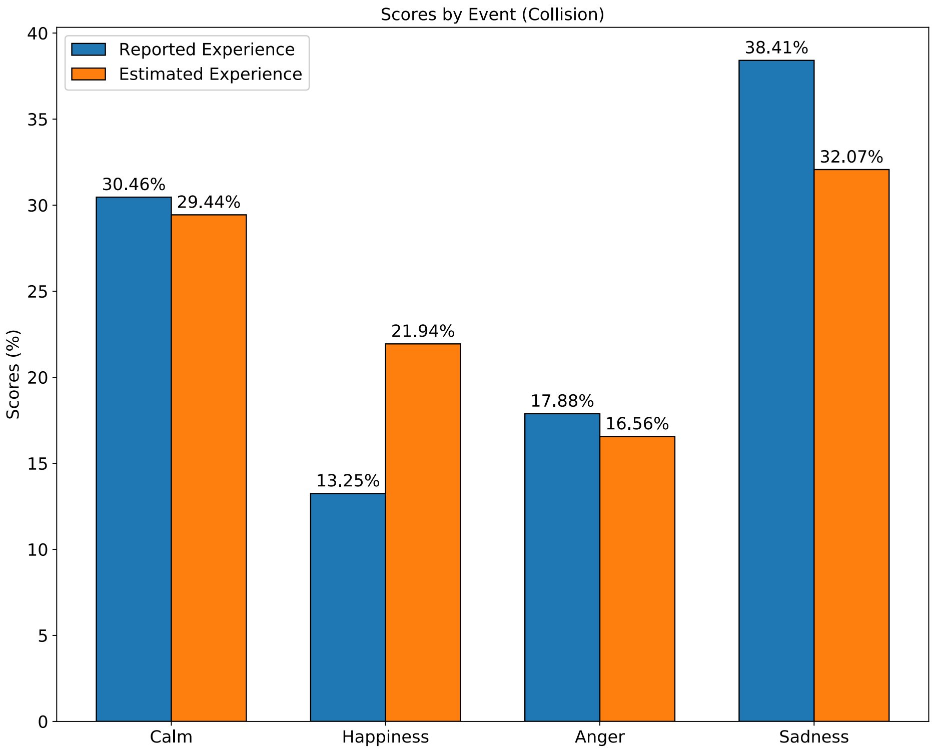


Fig. 14 Comparison of self-reported and measured player experience in events of “Collision”

Figure 15 shows the results from “Gearbox” events. There was a significant correlation for the reported experience and the estimate experience ($\rho = 0.94$, *p-value* $< .05$). We believe that the sessions stimulated opposite but clearly defined emotions. On one hand, the use of automatic gear and an easy road in sessions 1 and 2, with fewer curves, clearly delineated “Happiness” and “Calm” emotions. The model estimated that 23.92% of the players felt “Calm” and the self-report indicated 23.81%, while the values for “Happiness” were estimated as 20.82% against 19.05% for the self-report. The negative emotions were more associated to sessions 3 and 4, where the lack of ability to use “manual gear” has made the game experience of several participants stressful, increasing the “Anger” and “Sadness” scores. The model estimated that 27.73% and 27.54% of the players felt sadness and anger, respectively, with both values in good agreement with respect to self-reported data: 28.57% for both.

Figure 16 shows the results for “Drifting” events and the Spearman coefficient indicates a high correlation between the reported experience and the estimate experience ($\rho = 0.99$, *p-value* $< .05$). The sparse mentions of “Anger” and “Sadness” from reported data corroborate with the low estimates of our model for the same emotions. “Anger” and “Sadness” had rates of 12.33% and 14.05%, while reported data were 8.00% and 16.00%, respectively. “Anger” was a bit overestimated by the model, and the apparently large percentile difference with respect to the reported experience (8.00%) is due to the low quantity of participants with that emotion. In fact, the participants reported that when performing “Drifting” they

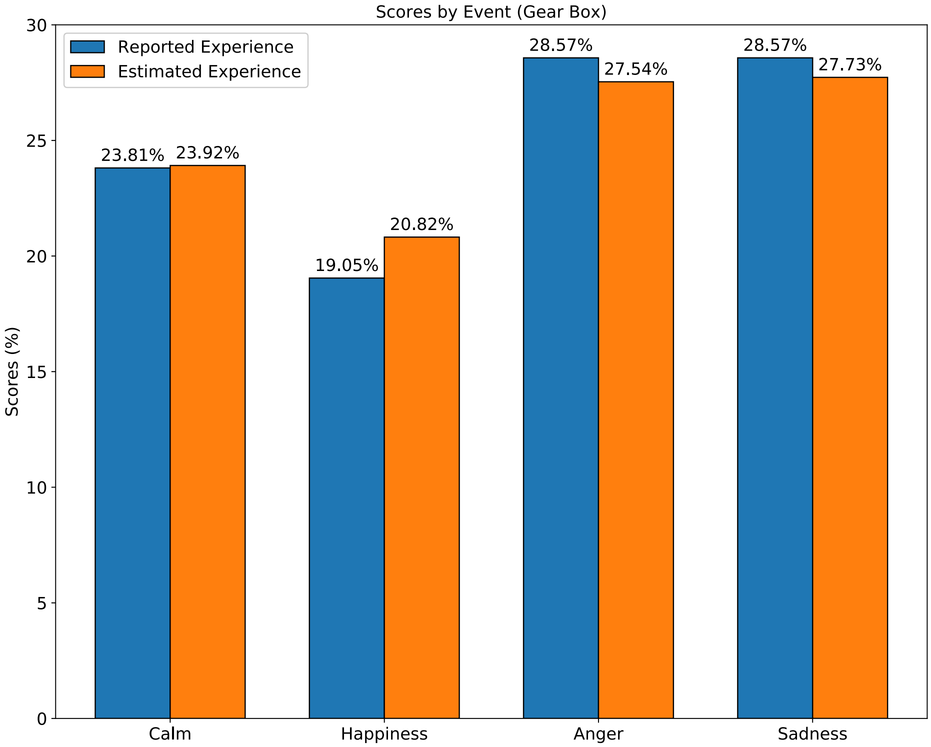


Fig. 15 Comparison of self-reported and measured player experience in events of “Gear Box”

felt a profound sense of satisfaction and increased confidence during gameplay. This is one of reasons for a high “Happiness” score, estimated by model as 49.42% and reported as 48.00% of the player experience. Finally, “Calm” was reported as 28.00% of the emotions, while the model estimated it as 23.9%.

8 Discussion

We conducted a detailed study of the state of the art and summarized into a survey-table (see Table 7), the most relevant results reported in the literature about the player experience evaluation through psychophysiological data obtained from biometrics sensors. Each row of Table 7 shows the main author along with the publication year, the set of psychophysiological signals used for that research, the emotion classes, the type of the method, and the results in terms of best recognition rate. Our classification model which is based on three type of psychophysiological data and trained with a human-annotated emotional expressions dataset provided an overall cross validation score of 64%, with a maximum accuracy of 90.91% for *Happiness*. The model validation with respect to the self-reported data resulted in emotions rates that were close to the reported ones. Some exceptions did occur, as the wrong interpretation of ironic smiles as expression of happiness. Some specific reactions could be related to cultural aspects and the classification must be further refined in order to take them into account.

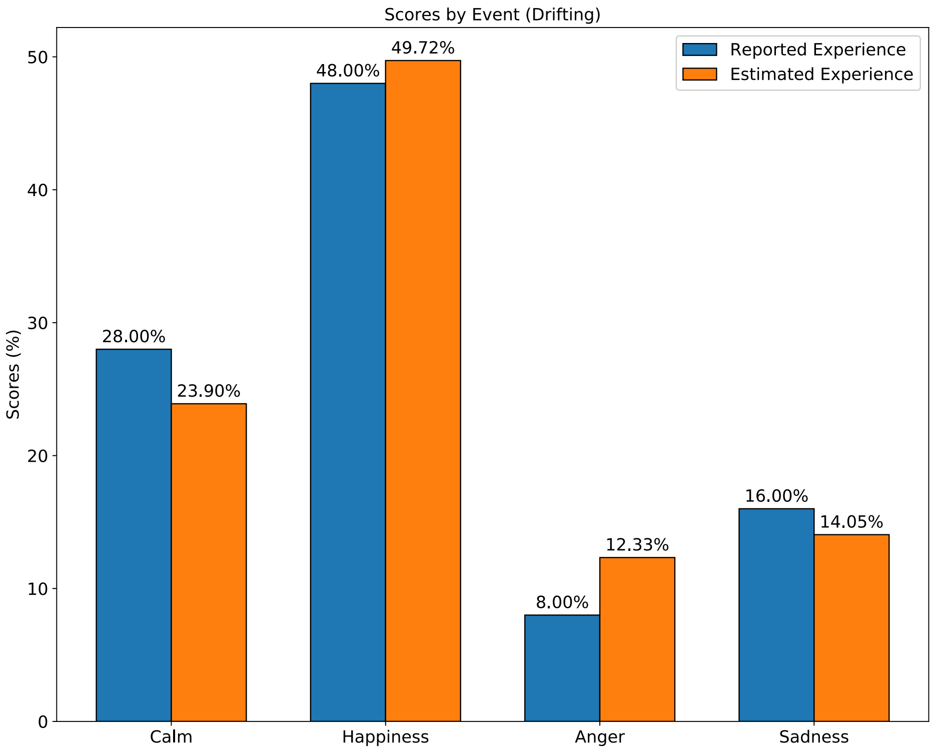


Fig. 16 Comparison of self-reported and measured player experience in events of "Drifting"

This paper described an experimental protocol from data collection to the creation of an artificial neural network. Player experience evaluation from psychophysiological data is not a new research topic, but our experiments evidenced that the combination of BVP, GSR and face expressions processed by an ANN trained with an expert annotated dataset can produce good estimations of emotional experience during gameplay. The adoption of a human-annotated dataset certainly contributed to the quality of the obtained results. The amount of improvement due to the dataset itself is difficult to quantify, however, as a qualitative analysis, the fact that humans know the kind of emotions produced by a racing game allows them to build a more customized classifier, improving the accuracy of the emotions estimations. On the other hand, classifiers obtained this way tend to be less portable among different game genres. To ensure a general validity of our classifier, reported results have been validated with a cross-validation method. A data analysis has been used to improve the accuracy of our classifier. Given the complexity of the dataset, 64% of accuracy over four classes is a considerable result. However, by analyzing the confusion matrix, we observe that there is an imbalance in the classification between classes that partially explains such performance. The our model's accuracy is close to that of works cited in Table 7 and it is able to generalize well on unseen data. As a result, it seems to have the potential to be applied as a general approach for evaluating the player experience, given that for each game genre a specific dataset is generated. Furthermore, we limited the number of emotional classes in order to the classification be feasible and increase the inter-rater agreement among evaluators.

Table 7 Performance of the player experience evaluation Methods Reported in the Literature

Author	Data	Emotion Classes	Method	Results(%)
Current work	Facial Expressions, BVP and GSR	Calm, Happiness, Anger and Sadness	ANN	64%
AlZoubi et al. (2021) [2]	BVP, ECG, EMG, EDA, Respiration and Facial Expressions	Happy, Fear, neutral, anxiety, excited and angry	XGBoost	66.16%
Chanel and Lopes (2020) [14]	GSR	Boredom, Flow and Anxiety	Deep Neural Networks	73.2%
Maier et al. (2019) [47]	BVP and GSR	Boredom, Flow and Stress	Deep Neural Networks	49%
Yang et al. (2018) [90]	GSR, ECG, EMG, Respiration, Temperature and Facial Expressions	Anger, Boredom, Fear, Frustration and Happiness	SVM	65%
Tognetti et al. (2010) [80]	BVP, ECG, GSR, Respiration and Temperature	Enjoyment	KNN classifier	70%
Mandryk and Atkins (2007) [49]	GSR, HR, EMG	Fun,Challenge, Boredom, Frustration, Excitement	Fuzzy Logic	64%,22%,8%, 9.7%,52%

The classifier proposed in this research could evaluate with good reliability the players' emotions when compared to the self-report. We believe that the classifier becomes more interesting for long game sessions, because it is more difficult for participants to remember all significant events in that case. On the other hand, for short game sessions it is easier for the players to provide useful feedback with self-report.

8.1 Limitations

During our experiments, we have found some technical limitations in the psychophysiological data collection and some restrictions that limit the achievement of best results. First, at the start of the research, Bitalino sensor [67] was used but it is sensitive to hand movements during the game sessions, therefore sudden movements resulted in high levels of noise. We then adopted Empatica E4 [35]. It is based on a wristband sensor, that is less sensitive to hand movements and provides a quite robust signal regularization system, reducing the noise related problems. Second limitation is that a high level of brightness in the environment can strongly impact the quality of the participants facial video capture, consequently decreasing the accuracy of the image processing system by introducing noise. The third limitation was an unbalanced set of data. One of the proposed solutions is to collect data from more players and work to generate a balanced affective dataset.

9 Conclusions

This study suggested an experimental protocol and ANN for the classification of the player experience from three psychophysiological data (GSR, BVP and facial expression) that can be applied in game applications. We collected physiological data from participants during playing a car racing game. Fifty players and eighteen judges participant with our experiments. The judges watched gameplay video of players for they assigned class labels (as describe in Section 5). Pyphysio [8] was used for preprocessing and feature extraction steps. We extracted 32 features (see Table 5) from three psychophysiological channels.

This paper reported on research on the effective use of psychophysiological data to classify player experience across different game events. The main contribution was the definition of an ANN architecture trained with an affective dataset that uses digital games as stimuli. Our model achieved an accuracy of 64% in detecting four emotions and the emotion "happiness" had the best result with 90% of correct classification. The results presented indicate that the strong individual differences of physiological responses affect the classifying, this same fact could be observed in AlZoubi et al. [2].

The validation based on the self-report of the participants indicated that the proposed model can classify four emotional experience states showing agreement in most of them regarding game events. Furthermore, the game events can provide more accurate information than a game session summary (which may not reflect actual experience), thus helping game developers to identify player experience issues.

The model could be improved by conducting an in-depth study of emotions models, e.g., using Russell's emotion model [72]. As future work, more attention can be devoted to processing new features extracted from different psychophysiological signals (e.g. temperature, respiration and eye tracker) and evaluating the dataset using different segmentation periods

[8] A library for physiological signal processing [6].

(15s, 20s, 25s and 30s). Also, other game genres (e.g., Horror games and Puzzle games, among others) will be studied, with the objective of analysing psychophysiological data to recognize and create a new affective dataset, including other emotions such as "fear", for instance. In the future, we might consider trying other classifiers such as support vector machines, decision trees and unsupervised learning techniques (clustering algorithms in order to classify player experience). Lastly, other improvements can be considered for this work, like the support for other input devices (e.g., EEG sensors)

Our work opens the perspective of introducing more sophisticated models of artificial intelligence for player experience evaluation. This research explores an alternative solution that allows us to evaluate player experience without using any traditional approach (for example, questionnaires and interviews). However, the traditional approach together with mixed-methods (using biosensors, facial expressions, among others) can make the evaluation process more robust and accurate.

The source code for our neural network and the data collected will be made available to other researchers at (https://github.com/eltonsarmanho/PGD-Ex). All videos collected in this study are strictly confidential. Only the researchers in charge have access to them.

Declarations The authors declare that they have no conflict of interest.

References

1. Alhassan S, Alrajhi W, Alhassan A, Almuhrij A (2017) Admemento: a prototype of activity reminder and assessment tools for patients with alzheimer's disease. In: Meiselwitz G. (ed) Social computing and social media. Applications and analytics, pp 32-43. Springer international publishing, Cham
2. AlZoubi O, AlMakhadmeh B, Bani Yassein M, Mardini W (2021) Detecting naturalistic expression of emotions using physiological signals while playing video games. J Ambient Intell Humanized Comput. https://doi.org/10.1007/s12652-021-03367-7
3. Benedek M, Kaernbach C (2010) A continuous measure of phasic electrodermal activity. J Neuroscience Methods 190(1):80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028. http://www.sciencedirect.com/science/article/pii/S0165027010002335
4. Benedek M, Kaernbach C (2010) Decomposition of skin conductance data by means of nonnegative deconvolution. Psychophysiology 47(4):647–658. https://doi.org/10.1111/j.1469-8986.2009.00972.x
5. Bernhard W, Eric E, Christophe G, Christos D, Remi C, Andrew S (2014) TORCS, the open racing car simulator. http://www.torcs.org. Accessed 29 Aug 2019
6. Bizzego A, Battisti A, Gabrieli G, Esposito G, Furlanello C (2019) Pyphysio: a physiological signal processing library for data science approaches in physiology. SoftwareX 10:100,287. https://doi.org/10.1016/j.softx.2019.100287. http://www.sciencedirect.com/science/article/pii/S2352711019301839
7. Bizzego A, Furlanello C (2017) Dbd-rco: derivative based detection and reverse combinatorial optimization to improve heart beat detection for wearable devices. https://doi.org/10.1101/118943. https://www.biorxiv.org/content/early/2017/03/21/118943
8. Boucsein W (2012) Electrodermal Activity. The Springer series in behavioral psychophysiology and medicine. Springer US, https://books.google.com.br/books?id=6N6rnOEZEEoC. Accessed 15 Jan 2019
9. Brockmyer JH, Fox CM, Curtiss KA, McBroom E, Burkhart KM, Pidruzny JN (2009) The development of the game engagement questionnaire: a measure of engagement in video game-playing. J Experimental Social Psycho 45(4):624–634. https://doi.org/10.1016/j.jesp.2009.02.016. http://www.sciencedirect.com/science/article/pii/S0022103109000444
10. Cacioppo J, Tassinary L, Berntson G (2016) Handbook of psychophysiology, fourth edn., https://doi.org/10.1017/9781107415782
11. Cacioppo JT, Gardner WL (1999) Emotion. Annu Rev Psychol 50:191–214
12. Cai J, Liu G, Hao M (2009) The research on emotion recognition from ecg signal. In: 2009 International conference on information technology and computer science, vol. 1, pp 497–500
13. Cattell R (1978) The scientific use of factor analysis in behavioral and life sciences. Plenum Press, https://books.google.com.br/books?id=JjoNAQAAMAAJ. Accessed 20 Jan 2019

14. Chanel G, Lopes P (2020) User evaluation of affective dynamic difficulty adjustment based on physiological deep learning. International conference on human-computer interaction, augmented cognition. Springer, Cham, https://archive-ouverte.unige.ch/unige:142293. Accessed 8 Feb 2022
15. Chanel G, Rebetez C, Bétrancourt M, Pun T (2011) Emotion assessment from physiological signals for adaptation of game difficulty. Trans Sys Man Cyber Part A 41(6):1052–1063. https://doi.org/10.1109/TSMCA.2011.2116000
16. Chang CY, Tsai JS, Wang CJ, Chung PC (2009) Emotion recognition with consideration of facial expression and physiological signals. 2009 IEEE symposium on computational intelligence in bioinformatics and computational biology, CIBCB 2009 - Proceedings pp 278–283. https://doi.org/10.1109/CIBCB.2009.4925739
17. Cheng B (2012) Emotion recognition from physiological signals using support vector machine. pp 49–52. https://doi.org/10.1007/978-3-642-03718-4_6
18. Cuthbert B, Schupp H, Bradley M, Birbaumer N, Lang P (2000) Brain potentials in affective picture processing: covariation with autonomic arousal and affective report. Bio Psychology 52(2):95–111. https://doi.org/10.1016/S0301-0511(99)00044-7
19. Davidson RJ (2003) Seven sins in the study of emotion: correctives from affective neuroscience. Brain Cogn 52(1):129–132
20. Dörner R, Göbel S, Effelsberg W, Wiemeyer J (2016) Player experience springer international publishing. https://books.google.com.br/books?id=nQ7pDAAAQBAJ. Accessed 3 July 2019
21. Drachen A, Connor S (2018) Game Analytics for Games User Research. pp 333–354. Oxford University Press, Oxford
22. Drachen A, Nacke L, Yannakakis G, Pedersen AL (2010) Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. In: Proceedings of the 5th ACM SIGGRAPH symposium on video games, sandbox '10, pp 49–54. ACM, New York, https://doi.org/10.1145/1836135.1836143
23. Drachen A, Seif El-Nasr M, Canossa A (2013) Game analytics – the basics. pp 13–40. Springer, London. https://doi.org/10.1007/978-1-4471-4769-5_2
24. Drenikow B, Mirza-Babaei P (2017) Vixen: interactive visualization of gameplay experiences. pp 1–10. https://doi.org/10.1145/3102071.3102089
25. Drioli C, Foresti GL (2015) The simple video coder: a free tool for efficiently coding social video data. pp 1563–1568. https://doi.org/10.3758/s13428-016-0787-0.The
26. El-Nasr MS, Drachen A, Canossa A (2013) Game Analytics, Maximizing the Value of Player Data. Springer, https://doi.org/10.1007/978-1-4471-4769-5
27. Empatica: real-time physiological signals - e4 wristband (2019) https://www.empatica.com/en-int/research/e4/. (November 10 Accessed 2019)
28. Fairclough SH, Venables L (2006) Prediction of subjective states from psychophysiology: a multivariate approach. Biol Psychology 71:100–110. https://doi.org/10.1016/j.biopsycho.2005.03.007
29. Granato M, Gadia D, Maggiorini D, Ripamonti LA (2018) Software and hardware setup for emotion recognition during video game fruition. In: Proceedings of the 4th EAI international conference on smart objects and technologies for social good, goodtechs '18, pp 19–24. Association for computing machinery, New York, https://doi.org/10.1145/3284869.3284895
30. Guardini P, Maninetti P (2013) Better game experience through game metrics: a rally videogame case study, pp. 325–361. Springer, London. https://doi.org/10.1007/978-1-4471-4769-5_16
31. Guardini P, Maninetti P (2013) Better game experience through game metrics: a rally videogame case study. In: El-Nasr MS, Drachen A, Canossa A (eds) Game analytics, maximizing the value of player data, pp 325–361. Springer, https://doi.org/10.1007/978-1-4471-4769-5_16
32. Harmon-Jones C, Bastian B, Harmon-Jones E (2016) The discrete emotions questionnaire: a new tool for measuring state self-reported emotions. PLoS One 11(8):1–25. https://doi.org/10.1371/journal.pone.0159915
33. Haykin S (2009) Neural networks and learning machines. Pearson international edition pearson. https://books.google.com.br/books?id=KCwWOAAACAAJ. Accessed 12 June 2019
34. Huynh S, Lee Y, Park T, Balan RK (2016) Japer: sensing gamers' emotions using physiological sensors. Proceedings of the 14th annual international conference on mobile systems, applications, and services companion p 104. https://doi.org/10.1145/2938559.2938576
35. Inc E (2018) Real-time physiological signals e4 eda/gsr sensor. https://www.empatica.com/research/e4/. (March 29 Accessed 2018)
36. Introduction to artificial neural networks (1995) Proceedings electronic technology directions to the year 2000, pp 36–62. https://doi.org/10.1109/ETD.1995.403491
37. Isbister K, Schaffer N (2008) Game usability: advancing the player experience. CRC Press. https://doi.org/10.1201/b14580

38. Isbister K, Schaffer N (2008) Using biometric measurement to help develop emotionally compelling games. pp 187–205. https://doi.org/10.1016/B978-0-12-374447-0.00013-5
39. Keltner D (2019) Toward a consensual taxonomy of emotions. Cogn Emot 33(1):14–19
40. Kim KH, Bang SW, Kim S (2004) Emotion recognition system using short-term monitoring of physiological signals. Med Biol Eng Comput 42(3):419–427. https://doi.org/10.1007/BF02344719
41. Kivikangas JM, Chanel G, Cowley B, Ekman I, Salminen M, Järvelä S, Ravaja N (2011) A review of the use of psychophysiological methods in game research. J Gaming Virtual Worlds 3:181–199
42. Kramer RSS, Mileva M, Ritchie KL (2018) Inter-rater agreement in trait judgements from faces. Plos One 13(8):1–17. https://doi.org/10.1371/journal.pone.0202655. https://doi.org/10.1371/journal.pone.0202655
43. Kuppens P (2019) Improving theory, measurement, and reality to advance the future of emotion research. Cogn Emot 33(1):20–23
44. Kushki A, Fairley J, Merja S, King G, Chau T (2011) Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites. Physiological Meas 32:1529–39. https://doi.org/10.1088/0967-3334/32/10/002
45. Lanata A, Valenza G, Scilingo EP (2012) The role of nonlinear dynamics in affective valence and arousal recognition. IEEE Trans Affect Comput 3:237–249. https://doi.org/10.1109/T-AFFC.2011.30
46. Larose D, Larose C (2015) Data mining and predictive analytics, 2nd edn. Wiley series on methods and applications in data mining. Wiley
47. Maier M, Marouane C, Elsner D (2019) Deepflow: detecting optimal user experience from physiological data using deep neural networks. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, pp 2108-2110. International foundation for autonomous agents and multiagent systems, Richland, SC
48. Malik M, Camm AJ, Bigger JT, Breithardt G, Cerutti S, Cohen RJ, Coumel P, Fallen EL, Kennedy HL, Kleiger RE, Lombardi F, Malliani A, Moss AJ, Rottman JN, Schmidt G, Schwartz PJ, Singer DH (1996) Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. European Heart J 17(3):354–381. https://doi.org/10.1093/oxfordjournals.eurheartj.a014868
49. Mandryk RL, Atkins MS (2007) A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. Int J Hum-Comput Stud 65(4):329–347. https://doi.org/10.1016/j.ijhcs.2006.11.011
50. Mandryk RL, Inkpen KM, Calvert TW (2006) Using psychophysiological techniques to measure user experience with entertainment technologies. Behaviour & information technology 25(2):141–158
51. Mandryk RL, Nacke L (2016) Biometrics in Gaming and Entertainment Technologies. chap. 6, pp 191–224. CRC Press, https://doi.org/10.1201/9781315317083-7
52. Marshall C, Rossman G (2014) Designing qualitative research SAGE publications. https://books.google.com.br/books?id=qTByBgAAQBAJ. Accessed 10 March 2019
53. McAllister G, White GR (2015) Video game development and user experience. Springer International Publishing, Cham, pp 11–35. https://doi.org/10.1007/978-3-319-15985-0_2
54. McGrath C, Palmgren PJ, Liljedahl M (2019) Twelve tips for conducting qualitative research interviews. Med Teach 41(9):1002–1006. https://doi.org/10.1080/0142159X.2018.1497149. PMID: 30261797
55. McMahan T, Parberry I, Parsons TD (2015) Modality specific assessment of video game player's experience using the emotiv. Entertainment Comput 7:1–6. https://doi.org/10.1016/j.entcom.2015.03.001. http://www.sciencedirect.com/science/article/pii/S1875952115000026
56. Medler B (2013) Visual game analytics. pp 403–433. Springer, London. https://doi.org/10.1007/978-1-4471-4769-5_18
57. Mirza-Babaei P (2014) Biometric storyboards: a games user research approach for improving qualitative evaluations of player experience. Ph.D. thesi University of Sussex. http://sro.sussex.ac.uk/47858/. Accessed 14 Aug 2019
58. Mirza-Babaei P, Nacke L, Gregory J, Collins N, Fitzpatrick G (2013) How does it play better? exploring user testing and biometric storyboards in games user research. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '13, pp 1499–1508. Association for computing machinery, New York, https://doi.org/10.1145/2470654.2466200
59. Myers J, Well A, Lorch R (2010) Research design and statistical analysis. Routledge. https://books.google.com.br/books?id=nbsOIJ_saUAC. Accessed 16 Feb 2019
60. Nacke L (2018) Introduction to biometric measures for Games User Research. In: Drachen A, Mirza-Babaei P, Nacke L (eds) Games user research, pp 281-299. Oxford University Press, New York
61. Nacke L, Drachen A (2011) Towards a framework of player experience research. In: Proceedings of the second international workshop on evaluating player experience in games at FDG 2011, Bordeaux, France
62. Nacke L (2013) An Introduction to Physiological Player Metrics for Evaluating Games. pp 585–619. Springer, London. https://doi.org/10.1007/978-1-4471-4769-5_26

63. Nacke L (2015) Games user research and physiological game evaluation. In: Bernhaupt R. (ed) Game user experience evaluation, chap. 4, pp 63–86. Springer international publishing
64. Orero JO, Levillain F, Damez-Fontaine M, Rifqi M, Bouchon-Meunier B (2010) Assessing Gameplay Emotions from Physiological signals - a Fuzzy Decision Trees Based Model. Kansei Eng Emotion Res Int Conf 2010(May):1684–1693
65. Park B, Jang E, Kim S, Huh C, Sohn J (2011) Feature selection on multi-physiological signals for emotion recognition. In: 2011 2nd International conference on engineering and industries (ICEI), pp 1–6
66. Paul E (2005) Basic Emotions, chap. 3, pp 45–60. Wiley-Blackwell. https://doi.org/10.1002/0470013494.ch3
67. Plux: Bitalino eda sensor datasheet (2016) http://bitalino.com/datasheets/EDA_Sensor_Datasheet.pdf. 20 Accessed November 2016
68. Posner J, Russell JA, Peterson BS (2005) The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol 17(3):715–734. https://doi.org/10.1017/S0954579405050340
69. Quandt T, Kröger S (2013) Multiplayer: The Ocial Aspects of Digital Gaming, 10001. Routledge, New York
70. Roohi S, Mekler ED, Tavast M, Blomqvist T, Hämäläinen P (2019) Recognizing emotional expression in game streams. In: Proceedings of the annual symposium on computer-human interaction in play, chi play '19, pp 301–311. Association for computing machinery, New York, https://doi.org/10.1145/3311350.3347197
71. Roohi S, Takatalo J, Kivikangas JM, Hämäläinen P (2018) Neural network based facial expression analysis of gameevents: a cautionary tale. In: Proceedings of the 2018 annual symposium on computer-human interaction in play, chi plAY '18, p. 429–437. Association for computing machinery, New York, https://doi.org/10.1145/3242671.3242701
72. Russell A, Weiss J, Mendelsohn AG (1989) Affect grid: a single-item scale of pleasure and arousal. J Pers Soc Psychol 57:493–502
73. Santhosh S, Vaden M (2013) Telemetry and Analytics Best Practices and Lessons Learned, pp 85–109. Springer, London. https://doi.org/10.1007/978-1-4471-4769-5_6
74. Sarmanho E, Santos T, Castanho C, Jacobi R (2018) Estimating player experience from arousal and valence using psychophysiological signals. SBGames 2018 - Computing Track. http://www.sbgames.org/sbgames2018/files/papers/ComputacaoFull/188391.pdf. Accessed 11 June 2019)
75. Scherer KR (2005) What are emotions? and how can they be measured? Soc Sci Inf 44(4):695–729
76. Siegert I, Böck R, Wendemuth A (2014) Inter-rater reliability for emotion annotation in human-computer interaction – comparison and methodological improvements. J Multimodal User Interfaces 8:17–28. https://doi.org/10.1007/s12193-013-0129-9
77. Soares RT, Sarmanho E, Miura M, Barros T, Jacobi R, Castanho C (2017) Biofeedback sensors in electronic games: a practical evaluation. In: 2017 16th Brazilian symposium on computer games and digital entertainment (SBGames). pp 56–65. https://doi.org/10.1109/SBGames.2017.00015
78. Susana MM, Lucía QMO, Jaime CM (2016) Dynamic analysis of emotions through artificial intelligence. Avances en Psicología Latinoamericana 34(2):205–232
79. Tan CT, Bakkes S, Pisan Y (2014) Inferring player experiences using facial expressions analysis. In: Proceedings of the 2014 conference on interactive entertainment, IE2014, pp. 7:1–7:8. ACM, New York, https://doi.org/10.1145/2677758.2677765
80. Tognetti S, Garbarino M, Bonanno AT, Matteucci M, Bonarini A (2010) Enjoyment recognition from physiological data in a car racing game. In: Proceedings of the 3rd international workshop on affective interaction in natural environments, AFFINE '10, pp 3–8. ACM, New York, https://doi.org/10.1145/1877826.1877830
81. UKharat G, Ul SVD (2008) Emotion recognition from facial expression using neural networks. In: 2008 Conference on human system interactions, pp 422–427. https://doi.org/10.1109/HSI.2008.4581476
82. Unluturk MS, Oguz K, Atay C (2009) Emotion detection with a fuzzy theory approach. In: Proceedings of the 10th WSEAS international conference on neural networks, Prague, Czech Republic
83. Valenza G, Lanata A, Scilingo EP (2012) The role of nonlinear dynamics in affective valence and arousal recognition. IEEE Trans Affective Comput 3(2):237–249. https://doi.org/10.1109/T-AFFC.2011.30
84. Vieira L (2017) Assessment of fun from the analysis of facial images. University of São Paulo, Ph.D. Thesis
85. Vieira L, Silva F (2016) Assessment of fun in interactive systems: a survey. Cogn Syst Res, vol 41
86. Wallner G, Halabi N, Mirza-Babaei P (2019) Aggregated visualization of playtesting data. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, pp 363:1–363:12. ACM, New York, https://doi.org/10.1145/3290605.3300593

87. Wallner G, Kriglstein S (2013) Visualization-based analysis of gameplay data – a review of literature. Entertainment Comput 4(3):143–155. https://doi.org/10.1016/j.entcom.2013.02.002. http://www.sciencedirect.com/science/article/pii/S1875952113000049
88. Weedon B (2013) Game metrics through questionnaires, pp 515–537. Springer, London. https://doi.org/10.1007/978-1-4471-4769-5_23
89. Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques, third edn., vol. 54. https://doi.org/10.1002/1521-3773(20010316)40:6:¡9823::AID-ANIE9823)¿3.3.CO;2-C http://www.cs.waikato.ac.nz/ ml/weka/book.html%5Cn http://www.amazon.com/Data-Mining-Practical-Techniques-Management/dp/0123748569
90. Yang W, Rifqi M, Marsala C, Pinna A (2018) Physiological-based emotion detection and recognition in a video game context. In: 2018 International joint conference on neural networks (IJCNN), pp 1–8. https://doi.org/10.1109/IJCNN.2018.8489125
91. Yannakakis GN, Hallam J, Lund HH (2008) Entertainment capture through heart rate activity in physical interactive playgrounds. User Model. User-Adapted Interaction 18(1):207–243. https://doi.org/10.1007/s11257-007-9036-7
92. Zalabarria U, Irigoyen E, Martínez R, Salazar-Ramirez A (2017) Detection of stress level and phases by advanced physiological signal processing based on fuzzy logic. In: Grańa M, López-Guede JM, Etxaniz O, Herrero Á, Quintián H, Corchado E (eds) International joint conference SOCO'16-CISIS'16-ICEUTE'16, pp 301-312. Springer International Publishing, Cham

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

Affiliations

Elton Sarmanho Siqueira[1] · Marcos Cordeiro Fleury[2] · Marcus Vinicius Lamar[2] · Anders Drachen[1] · Carla Denise Castanho[2] · Ricardo Pezzuol Jacobi[2]

Marcos Cordeiro Fleury
marcoscfleury@outlook.com

Marcus Vinicius Lamar
lamar@unb.br

Anders Drachen
anders.drachen@york.ac.uk

Carla Denise Castanho
carlacastanho@unb.br

Ricardo Pezzuol Jacobi
jacobi@unb.br

[1] Department of Computer Science, University of York, York, UK

[2] Department of Computer Science, University of Brasilia, Brasília, Brazil