This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

Flow Experience Detection and Analysis for Game Users by Wearable Devices-based Physiological Responses Capture

Xiaozhen Ye, Huansheng Ning, *Senior Member, IEEE,* Per Backlund, and Jianguo Ding, *Senior Member, IEEE,*

***Abstract*—Relevant research has shown the potential to understand the game user experience (GUX) more accurately and reliably by measuring the user's psychophysiological responses. However, current studies are still very scarce and limited in scope and depth. Besides, the low detection accuracy and the common use of professional physiological signal apparatus make it difficult to be applied in practice. This paper analyzes the GUX, particularly flow experience, based on users' physiological responses, including the galvanic skin response (GSR) and heart rate (HR) signals, captured by low-cost wearable devices. Based on the collected datasets regarding two test games and the mixed dataset, several classification models were constructed to detect the flow state automatically. Hereinto, two strategies were proposed and applied to improve classification performance. The results demonstrated that the flow experience of game users could be effectively classified from other experiences. The best accuracies of two-way classification and three-way classification under the support of the proposed strategies were over 90% and 80%, respectively. Specifically, the comparison test with the existing results showed that Strategy1 could significantly reduce the negative interference of individual differences in physiological signals and improve the classification accuracy. In addition, the results of the mixed dataset identified the potential of a general classification model of flow experience.**

***Index Terms*—Flow, games, game user experience, wearable devices, physiological responses.**

I. Introduction

THE prosperity of the game industry resulting from the development of computer technologies and the large-scale game consumers across the world have contributed to making digital games one of the most popular entertainment forms in recent years. It further promotes the development of scientific research of games. Hereinto, game user experience (GUX) relevant research attracted increasing attention as a representative paradigm of human-computer interaction (HCI) in games. On the one hand, it helps to improve the game itself concerning game design, development, evaluation, as well as users' understanding. On the other hand, the advancement of eliciting rich and meaningful experiences makes the game an excellent test-bed for various areas such as education, training, artificial intelligence, and HCI. Further research of GUX will contribute to the investigation and development of various cross-domains as well.

X. Ye and H. Ning are with the School of Computer and Communication Engineering, University of Science and Technology Beijing, 100083, Beijing, China. E-mail: ninghuansheng@ustb.edu.cn

P. Backlund and J. Ding are with University of Skövde, Sweden.

Generally, GUX is understood as the subjective relationship between users and games, which refers to the perception and responses of a user during the gameplay [1]. There is no formal taxonomy or commonly accepted criterion to define GUX at present since it is usually diversiform and complicated. In this case, a variety of terms and concepts are often used to describe the user's subjective cognitive or emotional experiences during gameplay, such as flow, boredom, anxiety, frustration, immersion, engagement, fun, stressful, valence-arousal, etc. They provide measurable criteria for reference, despite the limited description. Flow experience is regarded as an essential factor to maintain the participant's lasting involvement in a certain activity, thereby attracting a large number of studies in many fields. Different frameworks, models, methodologies, and experiments related to flow experience have been proposed, aiming at a better understanding of UX in HCI. Despite the increasing interest in flow experience in game research, existing studies are mostly based on post-game questionnaires, interviews, or online comments. They have considerable limitations in providing objective and precise information in terms of the user's cognition and emotion. In contrast, the physiological responses of users are generally involuntary, which guarantees the authenticity of information without interfering with the user's personality traits such as understanding of questions, answering style, emotion biases, short-term memory, etc. [2]. The measurement of physiological responses will undoubtedly open a new way for GUX evaluation, making more precise and real-time detection and analysis of GUX possible without disturbing the gameplay process.

However, physiological responses-based GUX researches are currently still scarce and do not yet form a unified field. Some occasional works with different scientific backgrounds and motivations may have achieved some success, but most of them are confined to the hypothetical verification or feasibility testing. Experiments with more samples as well as further exploration and analysis are urgently needed to establish a more precise and universal evaluation mechanism for GUX. In addition, the common use of professional physiological signal apparatus makes it difficult to apply in practice. It usually not only requires a typically high cost but also results in tedious preparation for an experiment as well as an uncomfortable experience for users because of the replacement of the electrodes.

In this paper, we present a study that detects game users' flow experience by capturing their physiological responses

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

based on low-cost wearable devices. An experiment that acquires the user's galvanic skin response (GSR) and heart rate (HR) signals under different game experiences is designed and conducted to generate associated datasets. Two datasets were established for two test games, and we analyzed them independently to investigate the flow experience in different games. Further, a mixed dataset was constructed to explore the potential of a general classification model for game users' flow experience. Considering the negative effects of individual differences and the variability of experience states overtime on classification accuracy, we proposed and applied two data processing strategies to improve the classification performance.

The remainder of this paper is organized as follows. Section II introduces the related theoretical background of flow experience and the related work with respect to psychophysiological-based GUX detection and measuring. Section III presents the detail of the data acquisition experiment. Next, the data processing methods and extracted features are described in Section IV. In Section V, the classification results of three datasets and the comparison test results are presented. Finally, a summary of this paper is presented in section VI.

II. Theoretical Background

A. Flow Experience

"Flow", as a term originated from psychological theory, was first proposed by Csikszentmihalyi. It refers to a mental state characterized by perceiving oneself to be entirely absorbed in some activity such as sports activity, musical performance, literary writing, artistic creation, etc., and achieving an enjoyable experience [3], [4]. Flow experience in games can be viewed as a kind of optimal experience with the game user's increasing motivation and engagement. Although there is still no well-accepted common definition of flow experience, the flow experience usually contains the following features: total involvement, high concentration (i.e. enough attention), and positive excitement. Many synonyms may share some aspects with flow experience. Some of them are even used as a substitution, such as presence, engagement, and immersion, while differences usually appear due to the different application scenarios or expression intension.

For the sake of qualitative and quantitative measurements, flow experience is usually identified in several dimensions. For example, Jackson et al. [5]–[8] divided the flow experience into nine dimensions based on the flow characterization descriptions of Csikszentmihalyi, to better understand the feelings of athletes when experiencing flow in various sports or physical activities. They further made a qualitative investigation of the flow state and established a flow state scale (FSS) based on the nine-dimensional flow description. Specifically, the nine dimensions of flow include: 1) clear goals, 2) timely and unambiguous feedback, 3) the balance between task challenge and participant's skill, 4) autotelic experience (i.e. an intrinsically rewarding experience), 5) sufficient concentration on current tasks, 6) sense of control, 7) the integration of self-awareness and action, 8) loss of self-consciousness, and 9) indistinct sense of time.

Supported by the flow theory of Csikszentmihalyi [3], [4] and currently existing evidence, the challenge-skill balance is no doubt a controllable antecedent of flow experience. In other words, flow experience generally occurs when participants' required skill level almost equals the challenge of the activity. Hence, many researchers suggested characterizing flow experience by adjusting the level difference between the task challenge and the participant's skill. On the contrary, sorely lacked or overwhelming challenges will prevent the participant from achieving the flow experience. As shown in Figure 1 [9], a high-level challenge beyond the participant's skill tends to result in anxiety experience while a too low challenge level easily leads to boredom experience. We call the area, where task challenge keeps balance with participant's skill, the "flow channel".

Figure 1 shows how flow experience works on the challenge-skill axis, but the challenge-skill balance is not the only necessary condition for flow. Moreover, the flow state in different periods also presents different manifestations. As Brown et al. pointed out that the flow experience in a continuous activity at least includes three progressive degrees: engagement, engrossment, and total immersion [10]. Engagement is the first stage where participants overcome the barriers of access and investment, and then put their time, effort, and attention to the activity. As engagement deepens, participants enter the stage of engrossment with a higher emotional attachment and less awareness of their surroundings. When overcoming all barriers of empathy and atmosphere, participants will lose self-awareness entirely as if their consciousness has shifted into the ongoing activity. This is the final stage of total immersion. To some extent, the three stages reveal that flow experience is a changeable measure, varying with the time and energy that participants put into the ongoing activity. However, the boundaries of the three stages are blurry, which offers less help for the quantitative evaluation of the different forms of flow experience. Also, the three stages' division lacks the consideration of participants' emotions, like joy and excitement.

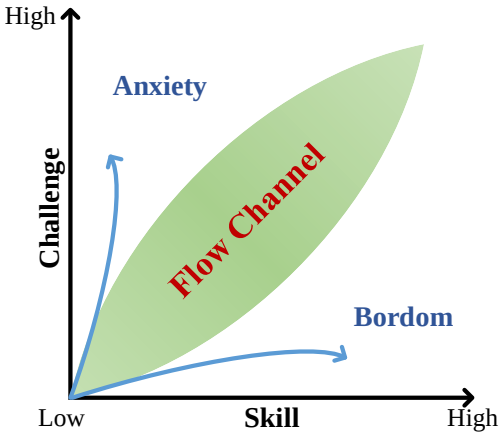


Fig. 1. The relation between flow experience and the challenge-skill balance

In most cases, flow experience enormously helps participants to gain better performance in the ongoing activities, thereby benefiting motivation promotion and some skill devel-

opment. For games, especially serious games, flow experience contributes to keeping users engaged in the game and better achieving some non-entertainment purposes such as training, learning, and so on [11], [12]. Therefore, several efforts attempt to catch the flow experience of game users or investigate the factors that impact the evocation of flow experience. Nevertheless, the most popular psychological questionnaire-based method has many limitations. On the one hand, the self-reported result is usually overall and rough impression of the gameplay. It lacks objectivity and veracity and does not support changeable flow state detection in different time spans. On the other hand, the activity process is more or less interrupted, which is also unfavorable to the real-time analysis. So far, very few studies (e.g., [13]–[16]) consider evaluating flow experience by measuring the user's physiological responses during the gameplay due to the difficulty of physiological-relevant data capture and analysis as well as the relatively significant investment. The limited works mostly tend to verify the method feasibility. Consequently, more extensive and deeper studies are still in need for a thorough understanding of the flow experience of game users.

B. Related Work

During the past decades, a number of research efforts of affective computing have provided insight for psychophysiological-based emotional experience detection and analysis. Most of the works aroused various physiological responses by materials like pictures, short auditory, video clips, etc. Similar studies conducted in GUX community are still limited, even though games often serve as an ideal test-bed of emotional induction because of its abundant content.

Inspired by the emotion detection based on the valence-arousal model, Mandryk et al. [17] attempted to develop an evaluation methodology for the quantification of the emotional experience in games. To induce different emotions in subjects, an NHL 2003 by EA Sports video game was used, and three game conditions were set by assigning the subjects different opponents, including a friend, a stranger, and a computer. Several physiological responses of subjects during interacting with games were recorded by surface electrodes, including GSR, electrocardiography (EKG), EMG of the face, and HR. In addition, the subject's other behavior reactions, such as code gestures, facial expressions, and verbalizations, were also recorded as auxiliary information for evaluation. The analysis results based on a fuzzy logic model showed that the variance of the subject's valance and arousal during gameplay could be expressed by the physiological responses. Further, they transformed the two-dimensional emotion to five discrete emotion states (i.e. boredom, frustration, challenge, fun, and excitement) based on adapted Affect Grid of six levels and identified the feasibility of quantifying the various emotional states of game users.

The study of Mandryk [17] opens the possibility of measuring GUX by physiological responses. After that, several extensive attempts are devoted to the inherent relation between the game user's cognition/emotion and different physiological responses. For example, Giakoumis et al. [18] tried to automatically recognize the boredom experience during playing video-games by detecting the user's ECG (Electrocardiograph) and GSR. Data from 19 subjects were collected through an experiment in which the boredom state of users was induced by repetitive 3D Labyrinth gameplay. Results showed that the maximum classification accuracy by combining conventional physiological features with moment-based features reached around 94.17%. Similarly, to investigate the cause of stress (including intrinsic factors or extrinsic stimuli), Ohmoto et al. [19] attempted to detect the stress state of users while playing a virtual exercise game by physiological indicators including SCR and ECG. Results revealed that both indicators were effective in detecting the stress state and further distinguishing intrinsic and extrinsic stress.

As a common and essential psychological phenomenon in gameplay, flow experience is also an attractive focus on GUX research. According to the intrinsic relation between multiple physiological responses (e.g. EEG, ECG, EMG, GSR, and eye-tracking) and emotional expressions, Nacke et al. [13] conducted an experiment to investigate the different representations of boredom, immersion, and flow experience in gameplay (here the flow and immersion were distinguished as two different gameplay experiences) by utilizing a first-person shooter game called Half-Life 2. It is worth mentioning that three types of experiences were produced by setting the game design criteria such as the perfection of the game environment, the number of available weapons, richness of the game narrative, etc. However, this work just compared the differences in physiological signals between the three game conditions while lacking more-in-deep analysis.

Using the antecedent of flow theory that flow zone is usually achieved when a task challenge is balanced with user's skill, Chanel [14], [20], Katahira [15], and Harmat et al. [16] motivated flow experience and other distinguishable emotional experiences (e.g. boredom and anxiety/overload) by controlling game challenge. The Tetris game or mental arithmetic was chosen as inducement materials since it is well-known and easy to adjust the challenge level. The user's physiological responses were recorded by professional physiograph equipment, e.g. Biosemi Active or BIOPAC MP 150 System. Further, based on the recorded peripheral signals including GSR, HR, respiration (RSP), and skin temperature, Chanel et al. [14] classified boredom, engagement, and anxiety states by training three classifiers: LDA (linear discriminant analysis, QDA (quadratic discriminant analysis), and RBF SVM (support vector machine with radial basis function kernel), and achieved the best accuracy of 59%. In addition to the peripheral signals, multiple EEG signals were also recorded as a kind of supplement assessment indicators. Despite the premature and tentative exploration, these studies proved that flow experience in gameplay is related to the psychophysiological responses of users, and it can be distinguished from other emotional states through some physiological responses. Besides, flow experience is changeable as the user being familiar with the game.

The studies of [14]–[16], [20] undoubtedly offer the idea of capturing the physiological responses of game users and investigating their features when users are experiencing a flow state. However, an inevitable problem with professional

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

TABLE I
AN OVERVIEW OF TYPICAL EXISTING RELATED WORKS

Ref	Target emotion	Physiological measures	Devices	Test games	Sample size	Analysis methods	Results
[19]	stress	SCR, HR	PD[1]	a virtual exercise game	20	statistical analysis (t-test)	The objective physiological measures show statistical correlation with the subjective description of emotional experiences. It is possible to distinguish/characterize the experience states (e.g. stress, boredom, and flow) by multiple physiological responses.
[13]	valence and arousal	EEG, ECG, EMG, GSR	PD[1]	Half-Life 2	25 (only male)	statistical analysis (ANOVA)	
[15]	boredom, flow, overload	EEG	PD[1]	a mental arithmetic game	16	statistical analysis (ANOVA)	
[16]	flow	ECG, RSP, Cortical oxygenation	PD[1]	Tetris	77	statistical analysis (ANOVA with Tukey's HSD test as the post-hoc)	
[12]	flow	EEG	WD[2]	learning quiz game	20	statistical analysis (t-test)	
[11]	boredom, flow	EEG, HRV, GSR	WD[2]	modified Stroop game	20	statistical analysis (Markov chain, t-test)	
[17]	valence and arousal	GSR, EKG, EMG, HR	PD[1]	NHL 2003 by EA Sports	12	fuzzy logic	
[21]	boredom, flow	EEG	WD[2]	Tetris	8	machine learning (SVM)	Different game difficulties give rise to users' different emotional experiences, and users' physiological responses are effective in classifying some psychological states based on the designed classifiers.
[18]	boredom	ECG, GSR	WD[2]	a 3D Labyrinth game	19	machine learning (LDA)	
[14], [20]	boredom, engagement, anxiety	GSR, HR, RSP, EEG	PD[1]	Tetris	20	machine learning (LDA, QDA, SVM)	
our work	*boredom, flow, anxiety*	*GSR, HR*	*WD[1]*	*Whack-A-Mole, and obstacle avoidance game*	*67*	*machine learning (DT, LR, SVM, NB, RF)*	*Users' different cognitive or emotional experiences aroused by different game challenge settings can be classified by some discernible features of GSR and HR signals, even though differences may appear in various games. It is possible to construct a general GUX analysis model across different games. (See Section V).*

PD[1]: professional physiological apparatus with surface electrodes, such as BioSemi Active, BIOPAC MP 150 System, and Procomp5 Infiniti.
WD[2]: wearable devices, such as Mindfield® eSense, Neurosky, Contec, and NeuroSky Mindeset EEG headset.

physiograph is that it not only requires much time to prepare before data acquisition but also may bring much discomfort to users and thus affects their game experience. Instead, some researchers attempted to monitor or evaluate user's physiological responses by inexpensive non-medical or commercial devices, such as NeuroSky Mindeset EEG headset, Contec pulse oximeter, and Mindfield® eSense [11], [12], [21], and investigate the contributions of the physiological responses in flow experience detection for game users. These works provide new insight and improve the applicability of relevant scientific research in practice to some extent. However, they mostly still stay in the stage of preliminary exploration and hypothesis verification with very simple experiments.

Table I shows an overview of typical related works mentioned above. On the one hand, most existing works focus on validating the correlation between emotional experiences and different physiological responses by statistical analysis, while few studies aiming at further investigation. On the other hand, they are mostly confined to tiny samples and a single test game. The inference derived from the small-scale and simple experiment is generally difficult to be applied to the universal analysis of GUX. The game experience detection model resulted from a single game may not be possible to be applied to other games as well. Given these deficiencies, this paper aims to extend existing works and investigate a more general flow experience detection model by low-cost wearable devices-based physiological signals capture. Some basic comparative information of our work can be found in the last column of Table I

III. EXPERIMENT SETUP

Psychophysiological research suggests that the physiological responses are the substrate of psychological activities. In other words, many emotional states or cognitive processes of humans can be characterized or quantified by correlative physiological measurements [13], [22]. A psychophysiological study usually starts with an assessment of a psychological concept by physiological measures to an observation of the correlation between them via an experiment [22]. Accordingly, the core idea of this paper is to assess the flow experience during gameplay with the GSR (also known as the electrodermal response or skin conductance) and HR of the subjects and observe the intrinsic correlation between the flow state and various physiological indicators based on an experimental

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

manipulation. Concerning the principles of psychophysiological research [22], a data collection experiment is designed as follows.

A. Physiological Measures and Devices

In this paper, we mainly record the GSR and HR of the subjects by utilizing two inexpensive and commercially available products of Mindfield® eSense: the Skin Response Sensor[1] and Pulse Sensor[2]. Both physiological signals and devices are briefly introduced as follows.

1) GSR: GSR is a measurable physiological response resulting from the activity of the skin sweat glands. It is controlled by the autonomic nervous system (mainly the sympathetic nervous system) but beyond the range of our self-perception. GSR usually shows significant physiological changes when someone feels stress, anxiety, panic, etc. Therefore, it is extensively used as a good indicator of some mental states like concentration [11]. The main goal of the Skin Response Sensor is to produce an electrical current to the skin by applying a safe and unnoticeable electrical voltage using two electrodes, thereby measuring the activity of sweat glands based on the changes of the small electrical current. Generally, GSR consists of two components: the tonic skin conductance level (SCL) and the phasic named skin conductance responses (SCR) [23]. The SCL reflects the changes in skin conductance caused by activities in the absence of stimuli. It is often used as an indicator of physiological arousal, of which the value usually decreases as someone calming down gradually and finally reaches a stable level. In contrast, SCR reflects the changes in skin conductance when receiving an external stimulus, which generally occurs within 1.5-6.5s after the stimulus. It is also an important indicator of emotion measurement.

2) HR and HRV: HR refers to the number of heartbeats per minute. It is another frequently-used indicator of positive and negative emotions [17]. HRV reflects the ability to change the frequency of the heart rhythm. It is an important mechanism for the stability of the cardiovascular system, as well as a vital factor in balancing the activity of the cardiac sympathetic and vagus. It represents the small changes of the continuous cardiac cycle (i.e. R-R interval or N-N interval), and is often used as a quantitative indicator of the activities of the autonomic nervous system [18]. The Pulse Sensor measures the HR by recording a 1-channel ECG signal based on a chest strap with two electrodes. The sampling rate of the device is 500Hz, and only the time between two heartbeats is transmitted to the app.

B. Test Games

Considering the expectations and requirements of the experiment as well as the subject's experimental experience, the test games should include the following features:

a) The game content is simple or widely known to avoid an extra burden on subjects.
b) The game operation is relatively easy so that any subject can quickly master and play the game.
c) The game challenge is controllable and easy to adjust for subjects with different skill levels.
d) The time of complete gameplay is relatively short and adjustable as the experiment demands, which is also conducive to the fast arouse of certain game experiences.
e) Since one hand of the subject needs to be used for the placement of sensors, the test game is required to be playable by one hand to reduce the interference of hand movement on the physiological signals.

According to the principles above, we chose and designed the following two games as the test games of our experiment.

(a) Game1 (b) Game2

Fig. 2. The screenshots of two test games

1) Game1: It is an iPad-based adapted Whack-A-Mole game in which the traditional characters are substituted by Ultraman and monsters (as shown in Figure 2(a)). In short, the player's mission is to hit a certain number of monsters back into the hole by clicking on the screen. The game will be over once the player hits the Ultraman or continuously misses multiple monsters. In order to increase the player's interest in the game, a piano sound will be produced every time the player clicks on the screen, and thus continuous clicking will generate a complete song. The game challenge is adjusted by the monster number that the player is required to hit, the speed and frequency of monsters appearing, and the frequency of Ultraman appearing.

2) Game2: It is a PC-based obstacle avoidance game, as shown in Figure 2(b). The player's mission is to control the movement of the character to avoid the falling bars and get the scores. The game will be over once the character is hit by the bars. The goal of the game is that the player continues playing the game as long as possible until achieving the required score. The game challenge is controlled by adjusting the length, falling speed, and falling frequency of the bars.

C. Participants and Procedure

34 students (18 male and 16 female), ranging in age 18 to 30, at the University of Science and Technology Beijing, China, participated in this study. Each subject did two experiments by playing Game1 and Game2. The interval between the two experiments was more than two weeks. For both experiments, subjects were all informed beforehand of free and voluntary participation, and they fully understood what would happen in the experiments, including the information that would be collected. Additionally, all the collected data

[1]https://www.mindfield.de/en/Biofeedback/Products/Mindfield%C2%AE-eSense-Skin-Response.html

[2]https://www.mindfield.de/en/Biofeedback/Products/Mindfield%C2%AE-eSense-Pulse-%7C-optimal-HRV-Biofeedback.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

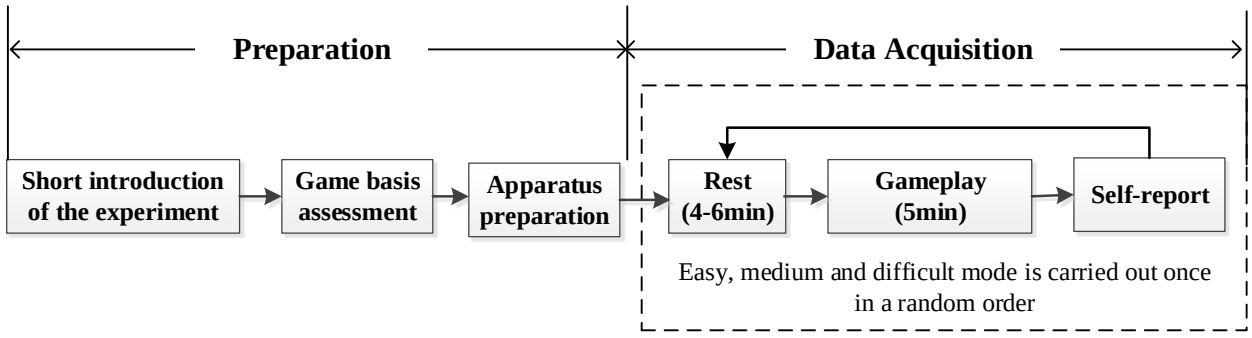


Fig. 3. Experiment process and data acquisition

was stored offline with a safe access principle and was only used for non-commercial research, which also was told to participants.

The experiment process can be summed up in the preparation stage and data acquisition stage, as shown in Figure 3.

In the preparation stage, after a short introduction of the experiment (including the experiment procedures and requirements, the test game, some notes, etc.), subjects played the test game several times to learn and master the test game. Then, subjects were required to finish a game basis assessment to learn about their game characteristics, including the time and frequency of playing games on weekdays, the self-assessment of the game skill, the extent of interest in games,etc. According to the preliminary investigation, most of the subjects usually spend less than one-hour playing games every day. Despite the differences in the gaming interest, the game skills of subjects are mostly similar with a small gap, and both test games are of interest and easy to master for all subjects.

Depending on the subject's performance in the game trial and their self-report, three challenge levels, namely easy, medium, and difficult, were set for each subject, attempting to elicit different experience states of subjects. Specifically, the challenge level of medium mode was roughly close to the game skill level of the subject. It was reported as a relatively smooth and pleasant game experience by the subject, which was mostly close to the flow state. Accordingly, an easy or difficult mode was determined by decreasing or increasing the challenge level of medium mode to a certain level until the subject felt a sense of boredom or anxiety more or less.

The data acquisition stage includes three consecutive turns, corresponding to the three challenge modes (i.e. easy, medium, and difficult). The three modes were ordered randomly for the sake of decreasing the side effects of time in questionnaires and physiological signals [14]. Each turn further consists of three sessions: rest, gameplay, and self-report.

The purpose of the rest session was to restore the subject to a calm state and hence acquire the baseline of every physiological signal. Given the different recovery abilities of an individual, the rest time usually lasted for 5 minutes with about 1-minute deviation. Some soothing music was played to assist the subject in returning to calm as soon as possible during the rest session.

In the gameplay session, the subject was asked to play the test game in a certain challenge mode repeatedly for 5 minutes, during which the subject was required to a new game with the same game setting once he/she failed. In order to motivate the subjects' enthusiasm to strive for an excellent game performance in the gameplay session, some incentives and hints were offered to subjects without influencing the experiment. Generally, the experience states may vary over time during gameplay because of the different challenge levels, design, and the pacing of the game in practice. Since the content and operation of the test game are relatively simple and repetitive, a short time of gameplay of the same game setting is hard to arouse multiple experience states with significant differences. Therefore, the game experience of the subject in the same gameplay session can be considered consistent.

After each gameplay session, the subject was required to finish a flow experience assessment scale (FEAS). The FEAS consists of 12 questions. Hereinto, the first three questions focus on overall subjective perception of the subject about the gameplay, e.g. how the subject feels about the boredom, flow (a short interpretation was presented by a similar concept of "enjoy" considering that the term "flow" is too professional for most of the subjects), and anxiety level during the gameplay. The other nine questions are about the measurement of the flow experience from the nine dimensions of the flow descriptions of Csikszentmihalyi. This part is adapted from the FFS used in [8] (as discussed in section II). Particularly, the FFS is simplified from 36 questions to 9 questions (i.e. the 4 questions in each group are reduced to 1 question) to avoid the negative impact of tedious and lengthy self-assessment on subjects. Since the relevance between each dimension and flow experience has been identified by Jackson et al. [5]–[8], this paper did not include more verification. The answer to each question of FEAS is designed as a five-point Likert scale, and all questions are presented in Chinese for the ease of understanding.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

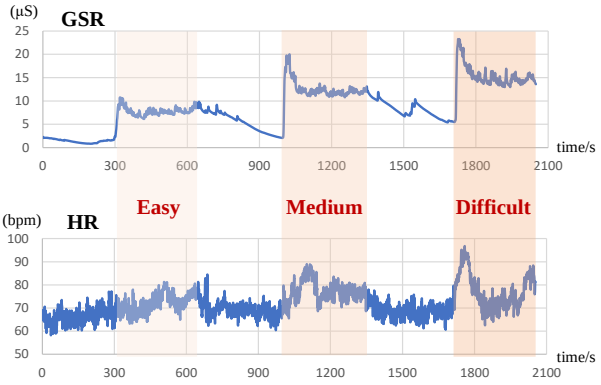


Fig. 4. An example of the visualization of one set of data from Dataset1

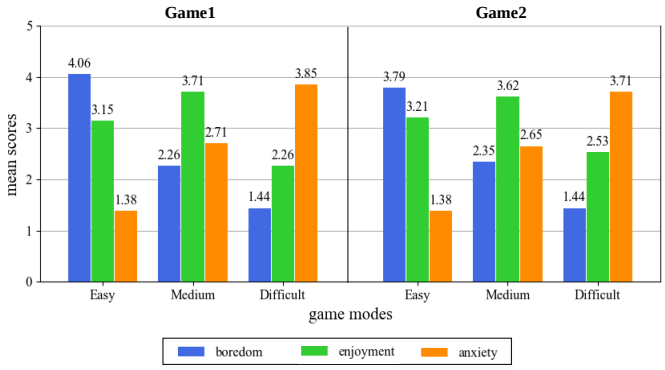


Fig. 5. The mean scores of the subjects' overall subjective perception of boredom, enjoy, and anxiety of FEAS in three challenge modes

IV. Data Processing

A. Datasets and Annotation

After two rounds of experiments, we collected two sets of data, named Dataset1 and Dataset2. In particular, Dataset2 consists of only 33 sets of valid data since one set of data was excluded for technical reasons.

Figure 4 is a representative of the visualization of one set of raw data of a subject during a complete experiment. The shaded parts represent three gameplay sessions and the labels of "Easy", "Medium", and "Difficult" represent the challenge mode of the test game. Note that the three modes of each subject are randomly ordered in the experiment. It can be seen that the GSR signal shows relatively significant differences between the challenge modes from a macro perspective, and the values generally increase as the game challenge (of course, some individual may show different changes). Besides, there is always a sharp increase in GSR at the beginning of each gameplay session, and then it returns to a stable state. Considering the specific time that the subject needs to spend to enter a stable state in each gameplay session, we considered the last four minutes as valid gameplay data. Another critical point is that the GSR level at the end of each rest session (i.e. the baseline level of each gameplay session) can be different due to the individual differences or hysteresis effect. As a result, the overall GSR level of the corresponding game session may show some deviation, which will have an impact on the analysis of GUX. One proposed solution is to eliminate the corresponding baseline from each gameplay session, which will be described in the section of "Feature Extraction". In contrast, the HR signal shows the less noticeable macroscopic differences between the challenge modes. Compared with the rest session, the HR during each gameplay session seems more variability.

The rating scores of each item of FEAS were averaged across the three challenge modes, and an analysis of variance (ANOVA) with all answers was applied to test the statistical significance of the results. Figure 5 represents the mean scores of the first part regarding the subjects' overall subjective perception of each gameplay session. The blue, green, and orange bars indicate the intensity of the three feelings, respectively. It can be seen that subjects mostly tend to have a relatively strong sense of boredom in easy mode ($F = 88, p < 0.001$), enjoyment (i.e. flow) in medium mode ($F = 17, p < 0.001$), and anxiety in difficult mode ($F = 79, p < 0.001$), which is consistent with the flow theory presented in Figure 1. Moreover, the sense of boredom diminishes as the challenge increases while the sense of anxiety increases($p_{(B)}$[3] < 0.001 for all pairs). The sense of enjoyment in the easy mode is weaker than that in medium mode ($p_{(B)} = 0.076$) but stronger than difficult mode ($p_{(B)} = 0.02$). For the second part regarding the flow state assessment, each dimension except the ninth one was statistical significance ($p < 0.05$). The total scores of the nine dimensions also showed significantly different distributions ($F = 7, p < 0.01$). Figure 6 presents the mean of the total scores of the nine dimensions in different challenge modes. Both test games obtained the highest reported score in medium mode. In other words, subjects are more likely to be in a flow state when the game challenge is roughly balanced with their skill level. Whereas the score gap between easy modes and medium modes are relatively small, especially for Game2 ($p_{(B)} > 0.05$). It can be speculated that the difficulty of the game itself and the game content and types also have an impact on the flow experience to some extent.

The datasets corresponded to the three challenge modes, which are annotated as boredom, flow, and anxiety state, based on the subjects' subjective assessment and the flow theory discussed in Section I. At the same time, both boredom and anxiety states are also marked as a non-flow state. The last-minute of each rest session (i.e. one minute before each gameplay) is selected to compute the baseline of the corresponding game session, marking it as baseline data. It should be noted that the annotation of various experience states aims to distinguish the differences of GUX without presenting the meticulous feeling of each individual. On the one hand, as mentioned before, the flow state varies over time and may present a bit different features in each stage. According to the results of the subjects' self-assessment, the either easy

[3] $p_{(B)}$ represents the p-value of multiple comparisons by using Bonferroni method.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

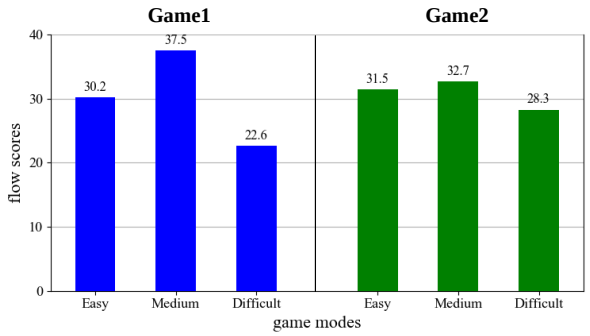


Fig. 6. The flow scores of FEAS in three challenge modes

or difficult mode can still arouse a certain degree of flow experience despite its lower intensity. The flow score of easy mode is even close to that of medium mode. On the other hand, the intensity of the flow state is also influenced by the subjects' personal characteristics such as the intension, motivation, mental state. It is usually uncontrollable in the experiment and cannot be reflected in the annotation as well.

B. Feature Extraction

Biological and physiological signals are generally non-stationary data of which the statistical features are the function of time. Therefore, physiological signals relevant analysis usually requires to extract various features to characterize the physiological response. Note that here the features are computed over the data of the entire gameplay session, separately. The features of GSR and HR signal are computed independently but used together in the subsequent classification.

For GSR, we extracted 39 time-domain features and 6 frequency-domain features. The time-domain features included some common statistical features such as the mean value ($\overline{x}$), standard deviation (std_x), median value (x_{median}), minimum value (x_{min}), maximum value (x_{max}), value range (x_{range}), the ratio of the minimum and maximum (x_{min_ratio} and x_{max_ratio}) as well as the ith-order difference (e.g. first-order difference ($1d_x$) and second-order difference ($2d_x$) in this paper). Some of these statistical features are illustrated in Table II. In addition, some features related to SCR were also computed as important indicators, including the number of SCR (n_scr) and the mean (scr_mean), minimum (scr_min), maximum value (scr_max) of SCR amplitude over a period of time, and the mean (scr_time_mean), minimum (scr_time_min), maximum value (scr_time_max) of the duration of each SCR. The frequency-domain features were acquired by extracting the power spectrum between 0.08Hz and 0.2Hz. We first applied Fourier transform to the signal and calculated the power spectrum based on Formula 1, and then calculate the mean, median, minimum, and maximum value of the power spectrum.

$$P_x = \frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2 \tag{1}$$

TABLE II
THE DESCRIPTION OF SOME COMMON STATISTICAL FEATURES

Feature name	Connotation
$\overline{x}$	the mean value of x signal. $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
std_x	the standard deviation of x signal. $std_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$
x_{range}	the difference between the maximum and the minimum of x signal $x_{range} = x_{max} - x_{min}$
x_{min_ratio}	the minimum ratio of x signal. $x_{min\_ratio} = \frac{x_{min}}{N}$
x_{max_ratio}	the maximum ratio of x signal. $x_{max\_ratio} = \frac{x_{max}}{N}$
$1d_x$	the first difference of x signal. $1d_x = x_{i+1} - x_i, (i = 1, 2, ..., N-1)$
$\overline{\lvert 1d_x \rvert}$	the mean value of the absolute value of the first difference of x signal. $\overline{\lvert 1d_x \rvert} = \frac{1}{N-1}\sum_{i=1}^{N-1}\lvert x_{i+1} - x_i \rvert$
$\widetilde{\lvert 1d_x \rvert}$	the mean value of the absolute value of the first difference of the normalized x signal. $\widetilde{\lvert 1d_x \rvert} = \frac{1}{N-1}\sum_{i=1}^{N-1}\lvert \tilde{x}_{i+1} - \tilde{x}_i \rvert$ where $\tilde{x}_i$ is the normalized x signal.
$2d_x$s	the second difference of x signal. $2d_x = x_{i+2} - x_i, (i = 1, 2, ..., N-2)$

Similarly, we also calculated the common statistical features of HR and HRV (parts of formulas are shown in Table II). Given the importance of HRV in emotion detection, we further extracted other time-domain features of HRV, including NNVGR, SDNN, SDANN, RMSSD, SDSD, NN50, pNN50. The comments and formulas of these features are shown in Table III. They generally reflect the overall change of HR and the activities of sympathetic or parasympathetic nerves, which plays a vital role in emotion analysis. Besides, seven frequency-domain features in relation to HRV were computed based on the signal spectrum using the Fourier transform and Formula 1. The HRV spectrum of humans in the normal state ranges from 0Hz to 0.04Hz, and different frequency bands show differences under various stimulations. Thus, the method of frequency division is often used for the frequency domain analysis of HRV. Table IV presents a detailed description of the frequency-domain features of HRV.

To reduce the impact of individual differences, two baseline processing methods were used to calculate the baseline with respect to the features of GSR and HR signals. The first one is used to subtract the mean of the corresponding baseline data from each physiological value and then extract all physiological features from the difference values. The other one is used to calculate the features of baseline data and sample data, and then subtract the corresponding baseline feature from the sample features. Furthermore, the Z-score standardized method was carried out over all samples for the normalization of each feature. The purpose of feature normalization is to avoid numerical problems caused by the disequilibrium of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

TABLE III
PARTS OF IMPORTANT TIME-DOMAIN FEATURES OF HRV

Feature name	Connotation
NNVGR	mean of all N-N intervals. $NNVGR = \frac{1}{N}\sum_{i=1}^{N} NN_i$ where NN_i is the ith N-N intervals.
SDNN	standard deviations of all N-N intervals. $SDNN = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(NN_i - \overline{NN})^2}$ where $\overline{NN}$ is the mean of all N-N intervals.
SDANN	standard deviation of the means of several N-N intervals segmented by a specific time session. $SDANN = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{nn}_i - \overline{\overline{nn}})^2}$ where $\overline{nn}_i$ is the mean of the ith segmentation and $\overline{\overline{nn}}$ is the mean of all $\overline{nn}_i$.
RMSSD	root-mean-square of the difference between adjacent N-N intervals. $RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(NN_{i+1} - NN_i)^2}$
SDSD	standard deviation of the difference between adjacent N-N intervals. $SDSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(NN'_i - \overline{NN'})^2}$ where NN'_i is the difference between the ith adjacent N-N interval.
NN50	the number of interval pairs whose difference between adjacent N-N intervals is greater than 50ms.
pNN50	the ratio of NN50 to all sequential N-N interval pairs. $pNN50 = \frac{NN50}{N-1}$

TABLE IV
FREQUENCY-DOMAIN FEATURES OF HRV

Feature name	Connotation
TP	total frequency power of HRV in 0∼0.4Hz
VLF	ultra-low frequency power of HRV in 0∼0.04Hz
LF	low frequency power of HRV in 0.04∼0.15Hz
HF	high frequency power of HRV in 0.15∼0.4Hz
LF/HF	the ratio of LF to HF
LF_D	HRV energy density of low frequency. $LF\_D = \frac{LF}{LF+HF} \times 100\%$
HF_D	HRV energy density of high frequency. $HF\_D = \frac{HF}{LF+HF} \times 100\%$

feature ranges of all individuals and thereby improving the performance of classifiers. The Z-score algorithm is shown in Formula 2 where X_i is the ith dimension feature of original data; μ_i and σ_i present the mean and standard deviation of X_i, respectively; and X_i^* is the ith dimension feature after standardization.

$$X_i^* = \frac{X_i - \mu_i}{\sigma_i}, i = 1, 2, \ldots, p \quad (2)$$

TABLE V
THE CONFUSION MATRIX OF TWO-WAY CLASSIFICATION

predicted state \ true state	flow state	non-flow state
flow state	TP (True positive)	FP (False positive)
non-flow state	FN (False negative)	TN (True negative)

V. RESULTS AND ANALYSIS

A. Methods and Indicator

In this section, we aim to construct multiple classification models to detect and analyze the flow experience of game users. Five classifiers were applied, including Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF). To reduce the features' redundancy and to improve the utilization of effective features, eight feature-selection algorithms were also applied, including ANOVA test, recursive feature elimination with logistic regression (RFE-LR), RFE with random forest (RFE-RF), L1-based feature selection with logistic regression (L1-LR), L1-based feature selection with linear SVC (L1-SVC), random forest-based feature selection (RFFS), extremely randomized trees-based feature selection (ETFS), and principal component analysis (PCA). The five classifiers were used in paired with each feature selection algorithm, and the combination with the highest accuracy was selected for performance analysis (in the following, we use the classifier name to represent the corresponding classification model for convenience). For each dataset, 30% of sample data were used as test data while others were used for model training.

The classification was carried out in two ways: 1) detecting flow experience from non-flow experience (i.e. two-way classification) and 2) distinguishing boredom, flow, and anxiety state (i.e. three-way classification). The classification results were reported and assessed by some common indicators such as Accuracy (Acc), Precision (P), Recall (R, also called Sensitivity), F1-score, and AUC score. Take the two-way classification as an example where flow state is defined as the positive condition. The confusion matrix is shown in Table V. Accordingly, the Acc, P, R, and F1-score are calculated by Formula 3, 4, 5, and 6. AUC score refers to the area under the ROC curve that plots true positive rate (TPR, shown in formula 5) vs. false positive rate (FPR, shown in formula 7) at different classification thresholds. In addition, a ten-fold cross-validation method was employed to evaluate the stability and generalization of classification models, in which the C-score denotes the mean of all test accuracy of the classifier.

Dataset1 and Dataset2 were analyzed separately for the sake of investigating the differences resulted from the game itself. Then a mixed dataset constructed by both datasets was used to explore the feasibility of a more general classification model. Lastly, we conducted a comparison test with the existing reported results in ref [14] for further verifying the proposed strategies.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

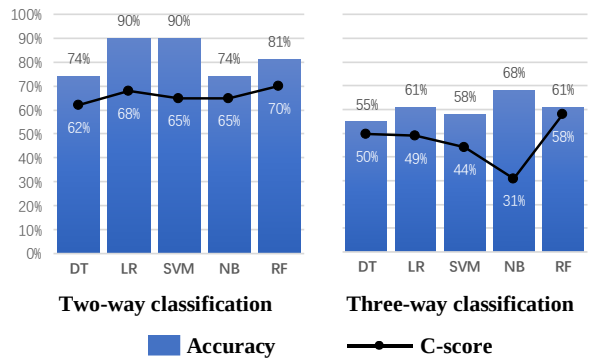


Fig. 7. The performances of five classification models of Dataset1

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = TPR = \frac{TP}{TP + FN} \tag{5}$$

$$F1\text{-}score = \frac{2PR}{P + R} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

B. Dataset1

Figure 7 shows the performances of five classification models of Dataset1 in two-way classification and three-way classification, in which the bars represent the accuracy and the polylines represent the C-score. The result indicates that the flow experience is distinguishable from others based on GSR and HR signals. However, the accuracies of most classification models are relatively low, especially for three-way classification. In particular, LR and SVM models show superior performance in two-way classification and get the highest accuracy of 90%, but lapse into mediocrity in three-way classification. NB model presents the highest accuracy in three-way classification but shows poor generalization (i.e. significantly low C-score).

Some typical causes may result in low classification accuracy. Firstly, the baseline processing in feature extraction just eliminates the effect of the individual initial value (i.e. the physiological value of an individual in a calm state). The differences in the values range and variability of the same physiological signal of each individual will still interfere with the classification effectiveness. For example, the values range of GSR of one subject in a boredom state can be close to that of another subject in an anxiety state. Secondly, although a relatively short gameplay time, the users' game experience may still change slightly as their increasing familiarity with the game (including game content, operation, etc.). As a result, the features calculated from the data spanning the entire gameplay session tend to drown out this subtle change. Thirdly, the number of experimental samples is relatively small to confine the generalization of a trained classification model.

Considering these challenges and in order to improve the classification accuracy, two data processing strategies were designed as follows:

Strategy1: Normalize the features in three challenge modes for each subject and then normalize the features of all subjects through Z-score standardized method. This strategy attempts to eliminate the negative effect of the values range and variability differences of a kind of physiological signals between individuals.

Strategy2: Fragment the data of each subject into multiple samples by adding a small time-window. For example, we fragmented each set of data in the 20-second units. On the one hand, it is conducive to convey the subtle changes of GUX during continuous gameplay. On the other hand, it increases the number of classification samples, which benefits the classification veracity to some extent.

For convenience, we name the original data processing method without using the proposed strategies as **Strategy0**. The comparisons of five classification models of Dataset1 using different strategies are shown in Figure 8. Hereinto, accuracy and C-score are denoted by bars and polylines, respectively. Different strategies are distinguished by the color of bars. It can be found that Strategy1 shows a significant contribution to promoting the classification performance. Both accuracies and C-score of all classification models were improved more or less. The best accuracies of two-way classification and three-way classification are 97% and 81%, respectively. In contrast, Strategy2 showed less potential in performance improvement. Some parameters, such as the accuracies of LR and NB, are even reduced compared with Strategy0.

Table VI and Table VII show the detailed results of two-way classification and three-way classification of Dataset1 using Strategy1. It is worth mentioning that the number of samples in the two-way classification is unbalanced. The sample size of the non-flow state (including boredom and anxiety states) is twice that of the flow state. By comparison, LR and SVM models show relatively superior performance over other classification models, and they show the best accuracies of 96% and 97% for two-way classification and 81% and 77% for three-way classification, respectively. The other parameters of LR and SVM models also achieved relatively high values. In other words, both classification models show preferable sensitivity and stability. According to Table VII, the precisions of boredom, flow, and anxiety state are mostly more than 70%, which indicates that the three states can be well classified by constructed classification models. By inspection, the feature selection algorithms also play an important role in improving classification performance except for the ANOVA test, in which RFE-LR, L1-LR, and RFFS methods are most frequently selected to construct the best classification model.

C. Dataset2

The comparisons of five classification models of Dataset2 using different strategies are shown in Figure 9. Compared with the results of Dataset1, the overall accuracy level of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

TABLE VI
THE RESULTS OF TWO-WAY CLASSIFICATION OF DATASET1 USING STRATEGY1

Classifier	Feature Selection	Accuracy	Precision	Recall	F1-score	AUC score	C-score
DT	ETFS	0.84	0.78	0.70	0.74	0.80	0.63
LR	**RFE-LR**	**0.96**	**0.91**	**1.00**	**0.95**	**0.98**	**0.95**
SVM	RFE-LR	0.97	0.91	1.00	0.95	0.98	0.92
NB	RFE-LR	0.90	0.77	1.00	0.87	0.93	0.86
RF	L1-LR	0.90	1.00	0.70	0.82	0.85	0.75

TABLE VII
THE RESULTS OF THREE-WAY CLASSIFICATION OF DATASET 1 USING STRATEGY1

Classifier	Feature Selection	Accuracy	Precision			C-score
			Easy/Boredom	Medium/Flow	Difficult/Anxiety	
DT	RFFS	0.65	0.67	0.64	0.62	0.58
LR	**RFE-LR**	**0.81**	**0.78**	**0.77**	**0.88**	**0.58**
SVM	RFE-RF	0.77	0.80	0.75	0.78	0.74
NB	RFE-RF	0.74	0.78	0.73	0.73	0.50
RF	L1-LR	0.67	0.75	0.64	0.67	0.72

TABLE VIII
THE RESULTS OF TWO-WAY CLASSIFICATION OF DATASET2 USING STRATEGY2

Classifier	Feature Selection	Accuracy	Precision	Recall	F1-score	AUC score	C-score
DT	RFFS	0.82	0.74	0.65	0.69	0.78	0.80
LR	L1-SVM	0.80	0.68	0.60	0.64	0.74	0.77
SVM	RFE-RF	0.87	0.78	0.77	0.78	0.84	0.85
NB	RFE-RF	0.74	0.63	0.32	0.42	0.62	0.66
RF	**RFE-RF**	**0.91**	**0.93**	**0.75**	**0.83**	**0.86**	**0.89**

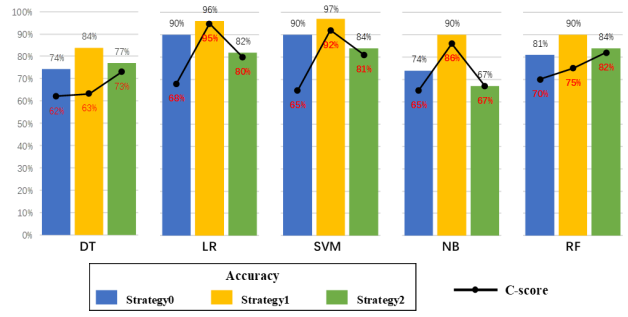
TABLE IX
THE RESULTS OF THREE-WAY CLASSIFICATION OF DATASET2 USING STRATEGY1

Classifier	Feature Selection	Accuracy	Precision			C-score
			Easy/Boredom	Medium/Flow	Difficult/Anxiety	
DT	**ETFS**	**0.80**	**1.00**	**0.58**	**0.91**	**0.68**
LR	**L1-LR**	**0.80**	**0.71**	**0.71**	**0.92**	**0.82**
SVM	L1-LR	0.77	0.80	0.67	0.79	0.80
NB	RFE-LR	0.77	0.80	0.62	0.83	0.77
RF	**L1-LR**	**0.80**	**0.80**	**0.67**	**0.91**	**0.80**

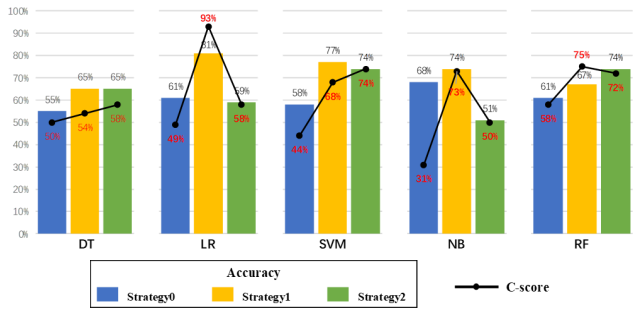
Dataset2 is lower. It can be concluded that the flow state, when playing Game2, is less distinguishable than that when playing Game1, which seems consistent with the results of the flow state assessment scale. Strategy1 was still effective in improving the performance of all models. Strategy2 showed a more significant effect in performance improvement in Dataset2 than Dataset1, and it obtained the best accuracy of 91% for two-way classification by RF and the best accuracy of 81% for three-way classification by SVM. Nevertheless, the improvement effect of Strategy1 is more prominent and stable, particularly for three-way classification.

Table VIII shows the detailed results of the two-way classification of Dataset2 using Strategy2, and Table IX shows that of three-way classification using Strategy1. Note that here we just present the most remarkable results as a reference in order to facilitate the analysis. It can be seen that most of the indicators of Dataset2 are decreased compared to Dataset1, especially the precision of flow state detection in both two-way classification and three-way classification. According to Table IX, the anxiety state is the easiest one to be classified in general, followed by boredom state, whereas many errors occur in recognition of flow state. By inspection, many marked flow state samples in the test set were mistakenly identified as boredom state. That happened because the content and rules of Game2 are relatively simpler so that users can easily master and get familiar with the game in a very short period. As a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

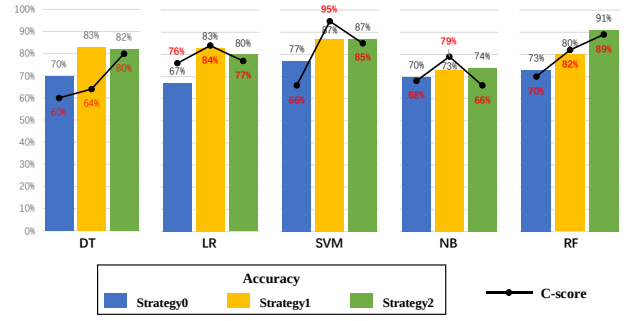


(a) Two-way classification

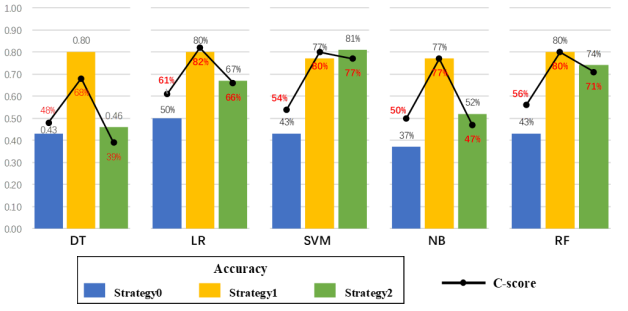


(b) Three-way classification

Fig. 8. The comparisons of five classification models of Dataset1



(a) Two-way classification



(b) Three-way classification

Fig. 9. The comparisons of five classification models of Dataset2

result, many users tend to enter a boredom state in the later stage of medium mode gameplay.

D. Mixed Dataset

To explore the possibility of a general classification model of flow experience, we constructed a mixed dataset by combing Dataset1 with Dataset2 and applied it for the classification analysis. The mixed dataset also makes up for the small sample size to some extent. Since the previous analysis has shown the positive role of the proposed strategies, here we mainly focus on the classification results by using both strategies.

Figure 10 shows the comparisons of five classification models of the mixed dataset using Strategy1 and Strategy2. The results indicate that the flow state can still be relatively well classified in spite of ignoring the game itself. In other words, there will be some commonness when users enter a flow state even while playing different games, and the commonness can be reflected by the physiological features. However, the overall accuracy level of the mixed dataset is slightly reduced compared with that of Dataset1 and Dataset2. For two-way classification, all classification models have comparable performances, in which SVM with Strategy2 got the best accuracy of 81%, followed by NB with Strategy1 and RF with Stragety2 of 80%. As for three-way classification, the performances of five classification models using Strategy1 were almost similar, and the best accuracy of 77% was obtained by LR and RF. On the contrary, Strategy2 showed apparently different effects on these classification models, in which RF had the best accuracy of 75% while NB gets a minimum of 49%.

Table X and Table XI show the detailed results of two-way classification and three-way classification of the mixed dataset using Strategy1. Compared with the results of both independent datasets, the precision, recall, and f1-score of two-way classification are reduced compared to that of Dataset1 but are superior to that of Dataset2 on the whole. One interesting point is that the classification performance seems to be neutralized with the intersection of two datasets. According to the results of three-way classification (as shown in Table XI), boredom state is the best classified in general, followed by flow state and then anxiety state.

E. Comparison Test

To further verify the proposed strategies, a comparison test with the three-way classification results based on the peripheral signals reported in ref [14] is presented. Considering the comparability of results, we used the same individual cross-validation method as in ref [14] to compute the test accuracy. That is, for each subject, we trained the classifier by the data of the other subjects, and then calculated the classification accuracy of the test subjects based on the trained model. The mean of the classification accuracies of all subjects was computed as the test accuracy of the constructed classification model. Each constructed classification model was iteratively applied in each dataset for several times, and the highest test accuracy was selected for the comparison.

The comparison results are shown in Figure 11. The first group of bars are the test accuracies of the three classifiers

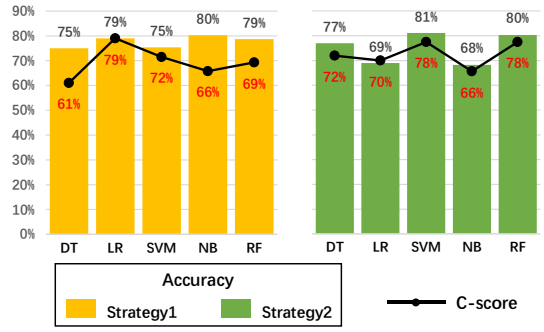
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

TABLE X
THE RESULTS OF TWO-WAY CLASSIFICATION OF MIXED DATASET USING STRATEGY1

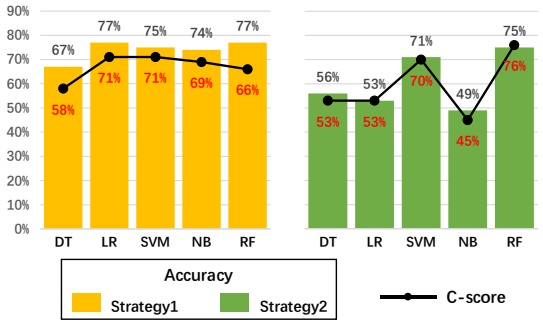
Classifier	Feature Selection	Accuracy	Precision	Recall	F1-score	AUC score	C-score
DT	L1-SVM	0.75	0.73	0.64	0.68	0.74	0.61
LR	L1-SVM	0.79	0.77	0.68	0.72	0.77	0.79
SVM	ETFS	0.75	0.92	0.44	0.59	0.71	0.72
NB	**ETFS**	**0.80**	**0.74**	**0.80**	**0.77**	**0.80**	**0.66**
RF	ETFS	0.79	0.93	0.52	0.67	0.75	0.69

TABLE XI
THE RESULTS OF THREE-WAY CLASSIFICATION OF MIXED DATASET USING STRATEGY1

Classifier	Feature Selection	Accuracy	Precision			C-score
			Easy/Boredom	Medium/Flow	Difficult/Anxiety	
DT	RFE-RF	0.67	0.93	0.67	0.47	0.58
LR	**L1-LR**	**0.77**	**0.73**	**0.81**	**0.78**	**0.71**
SVM	L1-LR	0.75	0.89	0.76	0.64	0.71
NB	L1-LR	0.74	0.88	0.69	0.67	0.69
RF	**RFFS**	**0.77**	**0.81**	**0.85**	**0.65**	**0.66**



(a) Two-way classification



(b) Three-way classification

Fig. 10. The comparisons of five classification models of mixed dataset

(i.e. LDA, QDA, and RBF SVM) reported in ref [14], while the last three groups are the results of Dataset1, Dataset2, and the Mixed Dataset. Hereinto, the blue, orange, and green bars indicate the results with Strategy0, Strategy1, and Strategy2, respectively. It can be seen that the accuracies of Strategy0 of each dataset are very similar to that of ref [14]. The best accuracies of the four groups are 59%, 60%, 66%, and 54%. By comparison, the accuracies of independent datasets are relatively higher than that of the mixed dataset. Strategy2 shows lower accuracy than Strategy0. It seems that the subtle emotional fluctuations experienced by the subjects during the same challenge mode are more prominent when using the individual cross-validation method. In contrast, the accuracy of each classification model with Strategy1 is greatly improved, and it still show superiority compared with the reported results in ref [14]. Three datasets have the highest accuracy of 76%, 82%, and 71%, respectively. It can be inferred that the individual differences in physiological signals are an important factor that has an impact on classification. The Strategy1 proposed in this paper shows great potential to effectively reduce its negative interference and improve the classification performance.

VI. CONCLUSION AND FUTURE WORKS

Inspired by the affective computing in HCI, this paper aims to understand game users' cognitive/emotional experiences during gameplay through their physiological responses. Combined with the flow model of Csikszentmihalyi, an experiment was designed to acquire the user's GSR and HR signals in different game experience states using wearable Mindfield® eSense sensors. Two types of games were used to motivate users' boredom, flow, and anxiety experiences by adjusting the game challenge relative to users' skill level. Hereinto, an adapted flow experience assessment scale was designed to obtain the user's subjective reports, and the results were used

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3010853, IEEE Internet of Things Journal

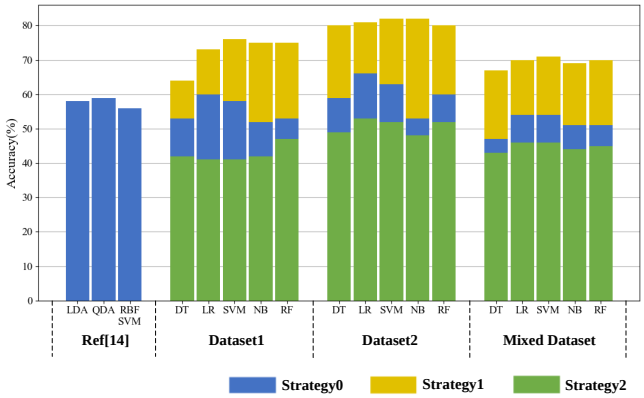


Fig. 11. The comparison with the reported results. The first three bars are the test accuracies of ref [14] using three classifiers: LDA, QDA, and RBF SVM. The last three groups of bars are the test accuracies of different constructed classification models in the three datasets.

as the reference of the ground truth. It is worth mentioning that the three states are relatively speaking since it is actually hard to induce a strong sense of boredom or anxiety using brief and simple gameplay according to our observation. Based on the collected datasets, different classification models were constructed and applied to detect the flow experience in gameplay. The performance of each model was evaluated from perspectives of two-way classification (i.e. flow state vs. non-flow state) and three-way classification (i.e. flow state vs. boredom and anxiety states). Despite different performances of various classification models, the best accuracies of two-way classification and three-way classification of both datasets were over 90% and 80%, respectively (specifically 97% and 81% for Dataset1, and 91% and 81% for Dataset2). According to the results, we conclude that different game challenge settings will arouse user's different cognitive or emotional experience, and the flow experience of users can be captured by some discernible features of GSR and HR signals of users even though some differences may appear in different games. Therefore, it is feasible to understand the GUX and the game itself through the users' physiological responses captured by low-cost wearable devices.

In spite of the positive results in this paper, some barriers are still encountered in practical GUX research. Some problems and gaps remain to be discussed and solved in future studies.

Firstly, the cognitive or emotional processing of game users is usually complicated and multidimensional. As a result, there is still no consistent theory and metric on GUX. Most of the descriptions and studies just capture a limited dimension of GUX as well. As mentioned before, the three experience states (i.e. boredom, flow, and anxiety) set in this paper are just rough descriptions to differentiate between different GUX. They can be relatively classified but still deviate from the real feelings of game users to some extent. Besides, the gameplay duration usually ranges from short fragment time to a long time, while the GUX states will change with the game progress. In this case, it is difficult to precisely identify the experience state of game users and its corresponding duration, which also increases the difficulty of the time setting of the gameplay session in the data acquisition experiment.

Secondly, the sample size in the paper is still limited in both experiment subjects and test games although it is larger than most of the existing studies (as shown in Table I). As Matias et al. [2] mentioned in the review work, although statisticians advise a minimum sample size for a reliable exploration of a large-scale effect, the results from limited samples are still implausible to demonstrate anything beyond the samples. It just provides a potential direction for further study. Accordingly, the constructed classification models based on the limited samples have relatively weak generality and extendibility in practice. Although a few studies have figured out some clues and proved something from the limited samples, more verification and efforts are still necessary to achieve a more general evaluation mechanism for GUX.

Lastly, individual differences are inevitably a key issue in the physiological responses-based GUX research. In addition to the differences in individual intrinsic physical qualities (e.g. the baseline of various physiological signals and their variability), some extrinsic factors associated with playing games, such as the game skill level, gaming motivation, tendency, interest, and so on, vary with each individual as well. According to the analysis in section III, these factors will have a certain impact on GUX, which is also challenging to establish a generalized evaluation metric. Moreover, game users from different communities, areas, and countries may show different characteristics, such as gaming habits and preferences. That is, the GUX may have group differences. Whilst the subjects in our experiment are mostly Chinese students with many similar characteristics. It also increases the limitation of the classification models in this paper. More extensive works are needed in the future study.

REFERENCES

[1] R. Bernhaupt, *Evaluating user experience in games: Concepts and methods*. Springer, 2010.
[2] J. M. Kivikangas, G. Chanel, B. Cowley, I. Ekman, M. Salminen, S. Järvelä, and N. Ravaja, "A review of the use of psychophysiological methods in game research," *journal of gaming & virtual worlds*, vol. 3, no. 3, pp. 181–199, 2011.
[3] M. Csikszentmihalyi, "Flow: The psychology of optimal experience," 1991.
[4] M. Csikszentmihalyi and I. S. Csikszentmihalyi, *Optimal experience: Psychological studies of flow in consciousness*. Cambridge university press, 1992.
[5] S. A. Jackson, "Factors influencing the occurrence of flow state in elite athletes," *Journal of applied sport psychology*, vol. 7, no. 2, pp. 138–166, 1995.
[6] S. A. Jackson, "Athletes in flow: A qualitative investigation of flow states in elite figure skaters," *Journal of applied sport psychology*, vol. 4, no. 2, pp. 161–180, 1992.
[7] S. Jackson, "Athletes in flow: Towards a conceptual understanding of flow state in elite athletes," *Research Quarterly for Exercise and Sport*, vol. 65, no. 3, pp. 122–136, 1994.
[8] S. A. Jackson and H. W. Marsh, "Development and validation of a scale to measure optimal experience: The flow state scale," *Journal of sport and exercise psychology*, vol. 18, no. 1, pp. 17–35, 1996.
[9] J. Chen, "Flow in games (and everything else)," *Communications of the ACM*, vol. 50, no. 4, pp. 31–34, 2007.
[10] E. Brown and P. Cairns, "A grounded investigation of game immersion," in *CHI'04 extended abstracts on Human factors in computing systems*, 2004, pp. 1297–1300.

[11] A. Sinha, R. Gavas, D. Chatterjee, R. Das, and A. Sinharay, "Dynamic assessment of learners' mental state for an improved learning experience," in *2015 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2015, pp. 1–9.

[12] C.-C. Wang and M.-C. Hsu, "An exploratory study using inexpensive electroencephalography (eeg) to understand flow experience in computer-based instruction," *Information & Management*, vol. 51, no. 7, pp. 912–923, 2014.

[13] L. Nacke and C. A. Lindley, "Flow and immersion in first-person shooters: measuring the player's gameplay experience," in *Proceedings of the 2008 conference on future play: Research, play, share*, 2008, pp. 81–88.

[14] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 41, no. 6, pp. 1052–1063, 2011.

[15] K. Katahira, Y. Yamazaki, C. Yamaoka, H. Ozaki, S. Nakagawa, and N. Nagata, "Eeg correlates of the flow state: A combination of increased frontal theta and moderate frontocentral alpha rhythm in the mental arithmetic task," *Frontiers in psychology*, vol. 9, p. 300, 2018.

[16] L. Harmat, Ö. de Manzano, T. Theorell, L. Högman, H. Fischer, and F. Ullén, "Physiological correlates of the flow experience during computer game playing," *International Journal of Psychophysiology*, vol. 97, no. 1, pp. 1–7, 2015.

[17] R. L. Mandryk and M. S. Atkins, "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies," *International journal of human-computer studies*, vol. 65, no. 4, pp. 329–347, 2007.

[18] D. Giakoumis, D. Tzovaras, K. Moustakas, and G. Hassapis, "Automatic recognition of boredom in video games using novel biosignal moment-based features," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 119–133, 2011.

[19] Y. Ohmoto, S. Takeda, and T. Nishida, "Distinction of intrinsic and extrinsic stress in an exercise game by combining multiple physiological indices," in *2015 7th international conference on games and virtual worlds for serious applications (vs-games)*. IEEE, 2015, pp. 1–4.

[20] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games," in *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*, 2008, pp. 13–17.

[21] A. Plotnikov, N. Stakheika, A. De Gloria, C. Schatten, F. Bellotti, R. Berta, C. Fiorini, and F. Ansovini, "Exploiting real-time eeg analysis for assessing flow in games," in *2012 IEEE 12th International Conference on Advanced Learning Technologies*. IEEE, 2012, pp. 688–689.

[22] J. T. Cacioppo, L. G. Tassinary, and G. Berntson, *Handbook of psychophysiology (3rd ed.)*. Cambridge university press, 2007.

[23] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments," *Psychophysiology*, vol. 49, no. 1, pp. 1017–1034, 2013.

Xiaozhen Ye received her B.S. degree from the School of Computer and Communication Engineering, University of Science and Technology Beijing, China, where she is currently pursuing the Ph.D. degree. Her current research interests include serious games and game user experience, and video-based cyber-physical interactions.

Huansheng Ning received his B.S. degree from Anhui University in 1996 and his Ph.D. degree from Beihang University in 2001. Now, he is a professor and vice dean of the School of Computer and Communication Engineering, University of Science and Technology Beijing, China. His current research focuses on the Internet of Things and general cyberspace. He is the founder of the Cyberspace and Cybermatics International Science and Technology Cooperation Base.

Per Backlund holds a PhD from Stockholm University and is currently a Professor of Information Technology at University of Skövde. He has been an active researcher in the field of serious games since 2005 with a specialization in game based and simulation based training. Professor Backlund has had the role of project manager and principal investigator in several research projects in serious games applications for different application areas such as traffic education, rescue services training and prehospital medicine.

Jianguo Ding holds the degree of a Doctorate in Engineering (Dr.-Ing.) from the faculty of mathematics and computer Science in University of Hagen, Germany. He is currently a senior lecturer(Docent) at school of informatics in University of Skövde, Sweden. His current research interests include distributed systems management and control, intelligent technology, probabilistic reasoning and critical infrastructure protection.