
Proceedings Paper:
Peng, Xiaolan, Xie, Xurong, Huang, Jin et al. (6 more authors) (2023) ChallengeDetect : Investigating the Potential of Detecting In-Game Challenge Experience from Physiological Measures. In: CHI 2023 - Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, 23-28 Apr 2023 Conference on Human Factors in Computing Systems - Proceedings . Association for Computing Machinery, Inc , DEU , pp. 1-29.

https://doi.org/10.1145/3544548.3581232

ChallengeDetect: Investigating the Potential of Detecting In-Game Challenge Experience from Physiological Measures

Xiaolan Peng
Institute of Software, Chinese Academy of Sciences
Beijing, China
xiaolan@iscas.ac.cn

Xurong Xie
Institute of Software, Chinese Academy of Sciences
Beijing, China
xurong@iscas.ac.cn

Jin Huang
Institute of Software, Chinese Academy of Sciences
Beijing, China
huangjin@iscas.ac.cn

Chutian Jiang
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
cjiang893@connect.hkust-gz.edu.cn

Haonian Wang
Beijing Normal University, Department of Artificial Intelligence
Beijing, China
aiwhn@mail.bnu.edu.cn

Alena Denisova
Department of Computer Science, University of York
York, United Kingdom
alena.denisova@york.ac.uk

Hui Chen
Institute of Software, Chinese Academy of Sciences
Beijing, China
chenhui@iscas.ac.cn

Feng Tian*
Institute of Software, Chinese Academy of Sciences
Beijing, China
tianfeng@iscas.ac.cn

Hongan Wang
Institute of Software, Chinese Academy of Sciences
Beijing, China
hongan@iscas.ac.cn

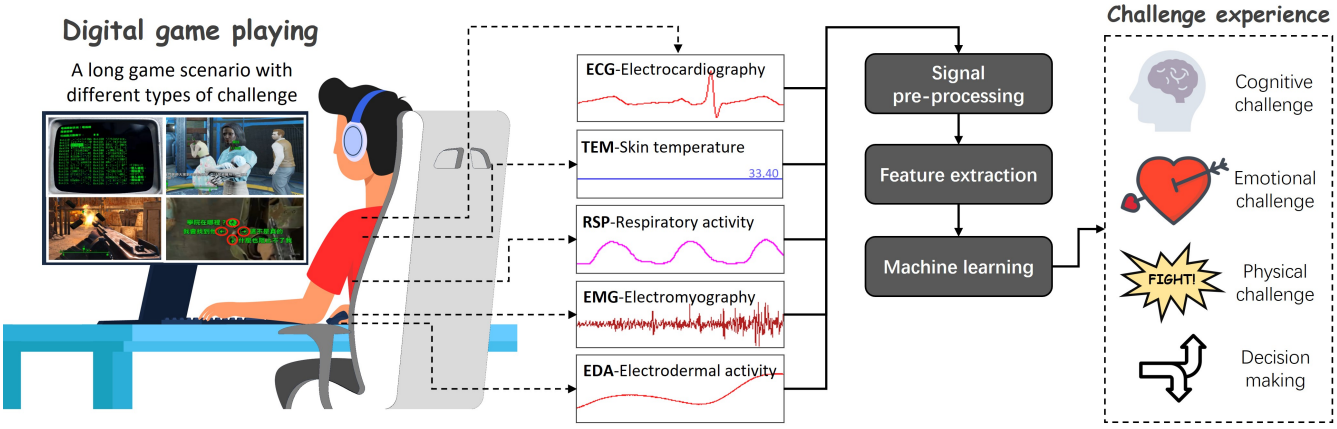

Figure 1: Detecting perceived challenge in a video game (Fallout 4) from physiological measures

ABSTRACT

Challenge is the core element of digital games. The wide spectrum of physical, cognitive, and emotional challenge experiences provided by modern digital games can be evaluated subjectively using a questionnaire, the CORGIS, which allows for a post hoc evaluation of the overall experience that occurred during game play. Measuring this experience dynamically and objectively, however, would allow for a more holistic view of the moment-to-moment experiences of players. This study, therefore, explored the potential of detecting perceived challenge from physiological signals. For this, we collected physiological responses from 32 players who engaged in three typical game scenarios. Using perceived challenge ratings from players and extracted physiological features, we applied multiple machine learning methods and metrics to detect challenge experiences. Results show that most methods achieved a detection accuracy of around 80%. We discuss in-game challenge perception, challenge-related physiological indicators and AI-supported challenge detection to inform future work on challenge evaluation.

*Corresponding author

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581232

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; *User studies*; • **Applied computing → Computer games**.

KEYWORDS

perceived challenge, video games, physiological signals, machine learning, player experience

ACM Reference Format:

Xiaolan Peng, Xurong Xie, Jin Huang, Chutian Jiang, Haonian Wang, Alena Denisova, Hui Chen, Feng Tian, and Hongan Wang. 2023. ChallengeDetect: Investigating the Potential of Detecting In-Game Challenge Experience from Physiological Measures. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3544548.3581232

1 INTRODUCTION

Digital games are a diverse medium that allows for a range of complex, nuanced and subtle experiences. Amongst a range of meaningful and eudaimonic player experiences [13, 14, 19], perceived challenge [5, 12, 21, 54] is a complex and multi-faceted player experience that arises from one's interaction with a game's intrinsic challenges at a particular skill level. It typically manifests itself in three types: physical (performative), cognitive, and emotional challenge, which is believed to exist independently from or jointly with one another [5, 12, 14, 54] and can change dynamically with time as the player progresses through a long game scenario with rich game contents and designs [56]. With this fact, it is vital to comprehensively and dynamically evaluate different types of challenge experiences, which would help to understand how different challenges are perceived by players throughout the gameplay, and further design games with specific challenge types and levels to adapt to players' skills, experiences, and motivation.

Players' perceived experience of challenge in video games can be assessed using the Challenge Originating from Recent Gameplay Interaction Scale (CORGIS). CORGIS is a validated tool that allows for the subjective quantification of several types of perceived challenge in video games, including the aforementioned cognitive, performative, and emotional challenge [21]. However, due to the discrete nature of the tool, the CORGIS does not allow for a continuous assessment of this experience – because questionnaires are administered after a game play session, the evaluation of challenge can only be done as an overall reflection of the entire session. Physiological metrics, on the other hand, provide a continuous, real-time, objective and quantitative assessment of player status. Such measures have long been leveraged to analyze or detect various experiences and in-game aspects, including game difficulty. Usually, specific affective states and experiences, like boredom and anxiety/frustration, can be mapped to the easy, normal and hard modes of a game, respectively. With this mapping, various physiological signals can be used to, first, detect specific emotions and then determine the game's difficulty level [8, 9, 18, 43, 49]. However, this method is not ideally suited for digital games with emotional challenge, as the diverse and complex emotional experiences triggered by emotional challenge do not perfectly align with the mapping of 'difficulty-emotion'. Considering that the process in which players make different efforts to overcome different challenges may also lead to different challenge-related physiological responses, we, therefore, argue in this paper that perceived challenge as a player experience should be explored in conjunction with players' physiological signals.

To investigate the potential of detecting a range of experienced challenge from physiological signals, we first constructed three game scenarios that cover the different challenge types and then conducted an experiment to collect 32 participants' physiological signals (electrocardiography (ECG), electrodermal activity (EDA), respiratory activity (RSP), electromyography (EMG) and skin temperature (TEM)) while interacting with the game scenarios. Participants' in-game challenge experiences were collected at discrete points throughout the game using the CORGIS questionnaire. With 80 physiological features extracted, several machine learning methods and metrics were applied to detect challenge experiences, among which the majority achieved an accuracy of around 80% (with the highest accuracy up to 85%) on challenge activation detection. Moreover, challenge-related features were selected and refined with feature importance techniques. Finally, we conclude by discussing players' perception of in-game challenge, challenge-related physiological indicators and also AI-supported challenge prediction to inform future work related to game challenge.

In this paper, we contribute a pipeline to detect in-game challenge experience with physiological signals, which consists of three parts:

(1) An experimental study that examines players' various types of in-game challenge experience and also offers a novel dataset[1] for challenge detection. The dataset consists of 32 players' physiological signals and their continuously reported challenge experience of playing three game scenarios.
(2) The first attempt to comprehensively and dynamically detect different types of perceived challenge based on physiological measures by using multiple machine learning methods.
(3) A set of challenge-rated physiological features selected and refined by feature importance techniques which could effectively indicate different types of perceived challenge.

2 RELATED WORK

2.1 Challenge in digital games

Challenge deals with the obstacles that players have to overcome and the tasks which they have to perform to make progress [1, 22, 59]. Overcoming different types of challenges requires different abilities and efforts from the player. Physical challenge depends on the skills such as speed, accuracy, endurance, dexterity and strength of the player [16, 37, 59]. Cognitive challenge requires players to use their cognitive abilities such as memory, observation, reasoning, planning and problem solving [16, 59]. And emotional challenge deals with tension within the narrative or difficult material presented in the game and can only be overcome with a cognitive and affective effort from the player [12]. Research into emotional challenge has considered how this experience differs from other, more traditional types of challenge [5], the diverse emotional responses it evokes [5, 54], the required game design characteristics [5, 14], as

[1]The data of this paper is released at GitHub: https://github.com/XIEXurong/ChallengeDetect.git.

well as challenge-related psychological theories [14]. By resolving the tension within the narrative, identifying with the characters, and exploring emotional ambiguities, players encountering emotional challenge would be put in a more reflective state of mind and experience more diverse, impactful and complex emotional experiences [5, 12, 21, 54–56].

In digital games, emotional challenge may exist independently or jointly with more traditional types of physical and cognitive (functional) challenge [5, 12, 14, 54]. In the first work to introduce emotional challenge, Cole et al. [12] suggest that emotional and functional challenge appear to be, at least partially, mutually exclusive. Bopp et al. [5] then explored the different game experiences that could feature emotional and functional kinds of challenge. To explore the tension between different types of challenge, Peng et al. [54] investigated how emotional challenge shapes player experience (PX) when presented separately or jointly with functional challenge. They show that adding emotional challenge to a game period dominated by functional challenges affects the emotional responses reported by the players [54]. In the latest work to investigate eudaimonic experiences, Cole and Gillies [14] suggests that different types of challenge can exist independently of one another, but practically they are inter-related and overlap at many points. Although the boundaries of different types of challenge are somehow blurred [22], it is important to note that digital games rarely provide players with exclusively one type of challenge. Instead, they require players to overcome them simultaneously or sequentially in a long-game scenario.

2.2 Challenge evaluation

Before emotional challenge research became a prominent direction in games HCI, game challenge had long been evaluated or explored as a similar concept to the "game difficulty". To offer a balanced level of challenge, the difficulty of the game needs to be assessed and adjusted to match the player's abilities and preferences. For games with static levels of difficulty (e.g. easy, normal and hard), challenge levels are largely identified based on researchers or designers' experience and the challenge balancing is done by directly altering game design characteristics, such as interaction speed [16, 34, 36], enemy agents [41], multi-tasks setting [34, 35], etc. However, such approaches have several downsides, for example, as players' skills and preferences are not discrete and evolve with time, it is not always possible to balance the difficulty of the game to keep the player 'in the zone' at all times. To help address this issue, dynamic difficulty adjustment (DDA) is used in digital games to continuously adapt the game difficulty which is often inferred or calculated based on the player's in-game performance [2, 15, 20, 51, 57, 62, 65] or their affective states (e.g. [48]). In both cases, these adaptations are most commonly afforded by relatively traditional physical and/or cognitive challenge [10, 18, 26, 27, 46].

Challenge as a player experience could be evaluated by gathering subjective data from players using validated questionnaires. Traditional physical or cognitive challenge has been considered as a component involving several challenge-related items in some questionnaires measuring broader PX [22]. With the efforts to better define game challenges, more comprehensive questionnaire tools have recently been developed to evaluate the wide range of challenge experience, including emotional challenge. One such tool for assessing perceived challenge is the Challenge Originating from Recent Gameplay Interaction Scale (CORGIS) [21] – a validated questionnaire that differentiates and measures four types of perceived challenge in digital games: cognitive, performative, emotional, and decision making challenge. Unlike the Video Game Demand Scale (VGDS) [6] that measures demand – a conceptually similar experience to challenge [24] – the CORGIS assesses the experience of the player exclusively inside the game, while the VGDS also covers the social aspects of gaming and other facets external to the on-screen experience.

While questionnaires like the CORGIS and VGDS are well suited for measuring retrospective recall of one's overall gaming experience, these tools do not allow for a continuous assessment of challenge that can be directly mapped to the different events inside the game. For this, physiological measures are commonly employed to evaluate various player experiences, among which many are targeted at affective/emotional states related to game difficulty level [3, 8, 9, 18, 43, 45]. Considering that a digital game which is too easy or too hard could ultimately lead to boredom or anxiety, respectively, flow and anxiety/frustration are often mapped directly onto the difficulty levels of 'easy', 'normal' and 'hard' [9, 43]. With such mapping, signals from various physiological measures, including ECG, EDA, RSP, EMG and TEM have been used to first detect a specific affective state and then adjust the difficulty of the game accordingly. This method has been widely adopted in biofeedback-controlled DDA in games dominated by traditional physical or cognitive challenge, such as Pong [18, 43], Tetris [8, 9] and first-person shooting games [49].

As many modern digital games offer richer and more complex experiences, the aforementioned methods might not be suitable for assessing the varied perceived challenge of players. Especially considering that emotional challenge often appears jointly with the more conventional types of physical and/or cognitive challenge, leading to more diverse and complex emotional responses from players. This makes the detection of perceived challenge difficult and impractical if based solely on specific kinds of emotional responses. For example, if physiological signals indicate that a player is anxious, it may be induced by a high level of physical or cognitive game difficulty, or it may also be caused by emotional challenge the player had encountered in the game, e.g. feeling worried about the safety of a non-player character (NPC). Besides, how well a player can overcome emotional challenge in a game cannot be as easily quantified unlike the conventional physical or cognitive challenge experiences, where the performance of the player is assessed based on their successful completion of a task. Hence, we argue that challenge should be directly explored with physiological signals to help understand how different challenge types are perceived by players throughout the entire gaming session and further promote real-time challenge adjustment for different players.

2.3 PX detection using machine learning methods

Various machine learning and pattern recognition methods have been used to detect PX from physiological signals. In general, machine learning models can learn a set of parameters to capture the

PX pattern characteristics from statistical information of features and labels of training data, and use these parameters to compute the pattern distribution from features of test data. For example, Chanel et al. [9] used an SVM (Support Vector Machine) to analyze physiological data (GSR and EEG) from 20 people playing Tetris on different difficulty levels. The SVM model obtained an accuracy of 53.33% on classifying three emotional states including boredom, anxiety and engagement. With a multi-modal database that contains peripheral physiological signals (ECG, EDA, RSP, EMG, temperature), accelerometer signals, facial and screening recordings as well as player's self-reported emotional experiences, Yang et al. [64] investigated physiological-based emotion detection (both valence-affect dimensional emotions and several categorical emotions) by using multiple machine learning classifiers (Linear SVM, radial basis function (RBF), SVM, Decision tree, Random Forest) among which Linear SVM achieved the highest average accuracy. In addition to using traditional machine learning methods, Maier et al. [45] applied an end-to-end deep learning CNN (convolutional neural network) to estimate the state of boredom, stress, and flow when playing Tetris on different difficulty levels based on 15.5 hours' physiological signals (EDA and blood volume pulse). The CNN model achieved an accuracy of 67.50% in recognizing high flow vs low flow states and 49.23% in distinguishing the state of boredom, stress, and flow. To adapt task difficulty based on automated affect recognition when playing Pong in a competitive mode, Darzi and Novak [18] used an SVM algorithm to classify perceived difficulty using 5 physiological responses (RSP, EDA, ECG, and 2 facial electromyograms). The SVM model achieved an accuracy of 84.3% on the perceived difficulty for two-class classification. As a proof of concept, Darzi and Novak [18] also conducted a small closed-loop study to demonstrate the technical feasibility of such physiological-based real-time adaptation.

Although physical and cognitive challenge, usually assessed through game difficulty, has widely been detected with physiological-based affective computing methods, little research has directly explored different challenge types and levels with physiological signals, especially with a focus on emotional challenge. Recently, to explore the possibility of detecting a wider range of challenge types based on physiological measures, Peng et al. [56] conducted a primary study using logistic regression models to detect different types of game challenges. A primary detection accuracy over 60% suggests the potential for further development of a real-time challenge measurement instrument. However, the study was limited to only one game scenario and the challenge types were determined by the researchers without a quantitative evaluation. Moreover, the study analyzed only one dominant challenge type with each in-game event, which largely limited the understanding of concurrent and inter-related challenge experiences. With increasing interest in defining and evaluating challenge experience, there is a need to investigate the experience of perceived challenge comprehensively and dynamically by combining quantitative, subjective challenge evaluation metrics with AI-supported computational methods, thus improving our understanding of how different types of perceived challenge manifest themselves in real-time and for different in-game scenarios.

3 EXPERIMENT

To explore the player experience of perceived challenge throughout game play and to also collect data for challenge detection, we designed and conducted an experiment. Specifically, we first constructed three game scenarios using a series of game quests from Fallout 4 [63]. The game scenarios were sought after to cover an entire spectrum of different game challenges. Then, the game events of each scenario were segmented so as to locate valid data for the following step. After a pilot study to refine experimental settings and instructions, a total of 41 participants were recruited to play the game scenarios during which their physiological signals (EDA, ECG, EMG, RSP and TEM) were recorded. Upon completion of the gaming session, they were asked to report on their overall challenge experience and emotional responses to playing the game and afterwards each participant was asked to recall their perceived challenge experience every 30 seconds. All the collected data were prepared for the analysis and detection of challenge in the next step.

3.1 Game scenarios

Fallout 4, an action roleplaying game (RPG), was chosen for this study as it provides narrative-rich content, supports modification, and has the potential for generating various types of perceived challenge [14, 54, 56]. Three **game scenarios** (A, B and C) were constructed with a series of Fallout 4 **game quests**[2] by three authors familiar with the game content and/or because of their expertise in the field of player experience and challenge, in particular. The constructed game scenarios were expected to meet the following criteria: 1) each game scenario should be possible to complete within one hour; 2) at least three types of challenge (physical, cognitive and emotional) are present; 3) a certain type of challenge might appear within different in-game scenes and different types of challenge might appear in a different order; and 4) the game events and main characters that players meet should also be different within different game scenarios.

3.1.1 Scenario A. Scenario A was constructed using the *War Never Changes*, *Out of time*, *Reunions* and *Institutionalized* quests. The *War Never Changes* quest begins with players walking around inside a house and interacting with their spouse, infant son (Shaun) and a robot housekeeper (Codsworth). The player then talks to a person and obtains the clearance to enter Vault 111. After a short period of time inside their house, the player learns that nuclear detonations have hit some nearby cities – this prompts the player with their spouse and son to rush toward Vault 111. Inside, each of them enters a separate cryo pod. Then the player witnesses three unknown figures shoot their spouse and take their son. When the player gets out of the pod, the game progressed to the *Out of Time* quest during which the player escapes from Vault 111, finds ruins all over the world and learns from Codsworth that 210 years have passed. After that, the player progresses to the *Reunions* quest to find Kellogg who killed his spouse. In *Reunions*, the player needs to fight against large amounts of guarding synths first and then talks to Kellogg to learn that Shaun is living at the Institute – an evil organization that creates synths. After killing Kellogg, the player progresses to the

[2]Detailed walkthrough of the quests can be found at https://fallout.fandom.com/wiki/Fallout_4_quests

Institutionalized quest to find their son. In a room at the Institute, the player encounters a scared and confused child who responds that he is Shaun but keeps asking 'Father' to help. Then the 'Father' enters the room and tells that the child is just a synthetic Shaun and he is the real Shaun. He is also the director of the Institute now. Finally, the player needs to decide whether to join their son in the Institute or leave their son.

3.1.2 Scenario B. Scenario B was comprised of the *Fire Support*, *Call to Arms* and *Blind Betrayal* quests. In the *Fire Support* quest, the player meets Paladin Danse, a commander of the Brotherhood of Steel in a battle against a swarm of feral ghouls. From Danse, the player learns that the Brotherhood is a just and united group and then responds to assist them to find a transmitter. In *Call to Arms*, the player follows Danse to an abandoned rocket silo to find the transmitter. Along the way, the player fights alongside Danse. In addition to encountering robbers and exploring the labyrinth-like silo, they go through many battles with the Brotherhood's biggest enemy – synths. Throughout the quest, Danse shows his resentment towards the synths. The *Blind Betrayal* quest takes place when the head of the Brotherhood, Elder Maxson, reveals to the player that Danse is a synth himself and orders the player to execute Danse. The player then confirms the news with Quinlan and locates Danse with Haylen. When the player finds Danse hiding in a bunker, the player then has a choice after talking to Danse about whether to save or execute him. If the player decides to save Danse, they need to convince Danse to escape with them. They then have to face Maxson when leaving the bunker. Again, the player must talk to Maxson to spare Danse, otherwise, Danse is executed by Maxson.

3.1.3 Scenario C. Scenario C was constructed from the *Human Error* quest. Before playing, the player learns that a caravan person has been killed in strange circumstances near a peaceful settlement called Covenant and their task is to investigate what has happened. In *Human Error*, the player begins by taking a SAFE test to enter the Covenant. In Covenant, the player talks to Honest Dan who is also an outsider of Covenant and agrees to help Dan to find Amelia Stockton, a girl missing from the caravan. The player then looks for information in Covenant by talking to the inhabitants, searching their houses and cracking office terminals to get clues, etc. Finally, the player learns that Amelia is hiding in a secret Compound, which the player then reaches together with Dan. Inside Compound, the player fights along with Dan against groups of scattered armed guards at the labyrinth-like Compound and also learns that the Compound are conducting cruel experiments to develop the SAFE questionnaire. After killing all guards, the player finally meets a doctor who is the person that the guards are trying to protect. The doctor's work is to reveal any hidden synths by perfecting the SAFE test, even by killing real human residents. She argues that the girl Amelia is most likely a synth infiltrator. The doctor then asks the player to support her to continue the experiment. Dan interferes with the conversation and disagrees with the doctor. The player needs to choose whether to kill the doctor and release Amelia or to support her and then kill Dan.

3.2 Game events

Although each game quest has a sequence of fixed tasks for players to complete, players in our experiment still have the opportunity to be involved in other trivial matters, such as encountering the blank screen of quest transitions, reading instruction tips, seeking help, exploring the open game world, etc. These matters may happen uncertainly in the game play and usually cause noise in the physiological signals. To remove them from the analysis, for each game scenario, a sequence of **game events** was segmented so as to locate valid data of each player. Particularly, each game event is a relatively independent in-game task that every player must complete in the game, such as being involved in an unavoidable fight, interacting with an important NPC, or searching for necessary information. These game events were segmented by the first author and Appendix A shows the list of the sequential game events within each game scenario.

3.3 Participants and procedures

3.3.1 Pilot study. A pilot study with 5 participants was conducted to refine the basic settings and instructions for each game scenario. Basic settings of arm equipment, character skills, game difficulty, etc. were kept at a normal level. Several instruction tips (see Appendix A) were given to the players in a written form before progressing to a specific game quest (each tip takes around 5 minutes to read).

3.3.2 Participants. For the main experiment, we recruited a total of 41 participants (19 male and 22 female, age M=24.3, SD=2.01) with no experience with Fallout series prior to the experiment. Each participant was randomly assigned to play one of three game scenarios (A, B or C). In this experiment, participants used a mouse and a keyboard to play the game on a PC with headphones (Figure 2-a). During the session, they were only allowed to use basic commands like moving, shooting, loading ammunition and restoring health, as well as interacting with NPCs and other key objects of the quests. Other advanced options like looting items, changing equipment and upgrading skills were not allowed.

3.3.3 Procedure. Each participant signed a consent form prior to playing the game. This was followed by an introduction to basic game operations. Participants then wore physiological sensors and went through a test period to familiarise themselves with the operations. Each participant had a rest period before starting the main part of the experiment. These procedures lasted approximately 25 minutes.

When playing the game, participants first read the written instructions to learn the basic background information. During playing, their physiological signals and gameplay screening were recorded synchronously. An experiment facilitator was present in the room outside of the participants' field of sight to provide help if necessary. After playing, participants removed the physiological sensors and filled out survey scales measuring their overall challenge experience and emotional responses. To complete each scenario (A, B and C), a player took on average 55, 50 and 40 minutes, respectively, with the survey reporting overall challenge experience and emotional responses adding another 15 minutes to the experiment time. Participants who completed this part were rewarded with 30 USD.

After playing the game, each participant was also trained (additional 5 minutes) to use the software developed by the authors for collecting their retrospective in-game challenge experience, which they had to fill out within two days after the completion of the study. They were rewarded another 40-80 USD according to the quality of their ratings.

3.4 Data collection

3.4.1 Physiological signals. Physiological signals, including ECG, EDA, RSP, EMG, and TEM, were recorded at a sampling rate of 1000 Hz using Biopac's MP150 system. For ECG, EDA and EMG, disposable AgCl electrodes were used. Skin temperature was collected by using the TSD202B of Biopac and RSP data was acquired with a BioNomadix respiration transducer belt, also from Biopac. Figure 2-b shows the placements of the various electrodes for physiological recordings and Figure 2-c is an example of the recorded physiological signals.

3.4.2 Overall player experience of each game scenario. Players' experienced challenge of each game scenario was measured using the CORGIS [21], which consists of 30 items (7-point Likert scale, 1 = strongly disagree and 7 = strongly agree) to evaluate cognitive, emotional, performative (physical), and decision making challenge types. Players' diverse emotional responses were measured using the Emotion Annotation and Representation Language (EARL) [60] to rate 48 kinds of emotions (9-point Likert scale, 0 = not at all and 8 = quite a lot). These data were collected immediately after each game scenario was completed.

3.4.3 Players' in-game challenge ratings. To capture players' perceived challenge experience, participants were given two days to reflect on their experience every 30 seconds by watching their gameplay screen recording post study. Participants used a rating software developed by the authors to play a collection of short videos one by one and answer a set of slightly modified CORGIS questions ("the game" in each item of the original CORGIS was changed to "the game period"). The short videos (30 seconds for each video), split from the gameplay screen recording, were assessed by the participants answering the questions about their experience exactly within the 30s game period. To facilitate data quality checking, for each 30s video, the software automatically and randomly selected one question from the CORGIS's 30 items and repeated it in the software's question sets for participants to answer. If a player's answers to the repeated questions are inconsistent, it may indicate that the rating is less reliable. In our experiment, answers that differ by more than 2 points on the repeated questions are deemed inconsistent, and a player's rating data is judged invalid if they have over 10% inconsistent answers across all the videos. The post-study rating of scenarios A, B and C took 4.38 (SD = 1.15), 3.66 (SD = 0.90) and 2.88 (SD = 0.82) hours respectively (calculated using valid data from 32 participants).

3.5 Results of the overall player experience

We excluded the data from 9 participants for the following reasons: among which 2 stopped playing due to 3D dizziness, 2 refused to do the post-study rating, 3 with invalid post-study retrospective data, one with irregular heart rate and one whose EMG electrodes became loose during the experiment. The final dataset was comprised of the data from 32 participants (age M=24.5, SD=2.46; 10, 11 and 11 participants for the game scenario A, B, and C, respectively), which was included in the analysis. None of them reported having fatigue during play, measured by Simulator Sickness Questionnaire [39].

Table 1 shows the results of participants' overall challenge experience and the high-scoring emotional responses for each game scenario. The rating for each type of perceived challenge was calculated as the average score of all items for that particular type using the CORGIS.

4 IN-GAME CHALLENGE EXPERIENCE

4.1 Data preparation

In our experiment, different participants played each game scenario at different paces. Scenarios A, B and C took 54.3 (SD = 4.87), 49.3 (SD = 5.68) and 38.7 (SD = 9.03) minutes to complete, respectively. For each game event (see Appendix A), both the time to start the event and the time taken to get through the event differed between individual players. To learn participants' perceived challenge experience across game play, we watched each participant's game play screen recording and marked their start and end time of each game event. Then we used the marked time stamp to locate each participant's challenge ratings for each game event. Particularly, as most of the segmented events lasted more than 30 seconds, the challenge experience with each event is indicated by the challenge rating at the time median of the event (see Figure 3). This allowed us to evaluate each participant's challenge experience with each game event. As the sample of players for each scenario was not large enough for statistical analysis, we instead report on the novel descriptive in-game challenge findings.

4.2 Descriptive analysis

Figures 4-a.1, 4-a.2 and 4-a.3 show the mean ratings of participants' challenge experience for each game event in scenarios A, B and C, respectively. Figures 4-b.1, 4-b.2 and 4-b.3 show example ratings of different participants (Appendix B offers more detail on Figure 4).

4.2.1 Dynamical challenge experience. Figures 4-a.1, 4-a.2 and 4-a.3 show that **different types of perceived challenge changes with in-game events during game play**. For example, in scenario A, physical challenge (phyC) is low before event 4 when the player stays at the house. It increases at event 5 and event 10 when the player rushes toward or escapes the Vault. The highest phyC is observed in events 14-16 when the player is involved in fighting large amounts of synths. With emotional challenge (emoC), it is low before event 6 and then increases to a medium level when the player witnesses three unknown figures shooting their spouse and taking their son in event 7. After that, emoC decreases and then increases until finding the scared and confused child (the synthetic son) in event 20 and talking to the 'Father' (the real son) in events 21-23. Decision making (DM) challenge, on the other hand, is low almost across the whole scenario until events 21-23, when the player is faced with important decisions. Finally, cognitive challenge (cogC) has small fluctuations and increases a little in event 2 when talking and event 10 when looking for a path. The highest cogC level is

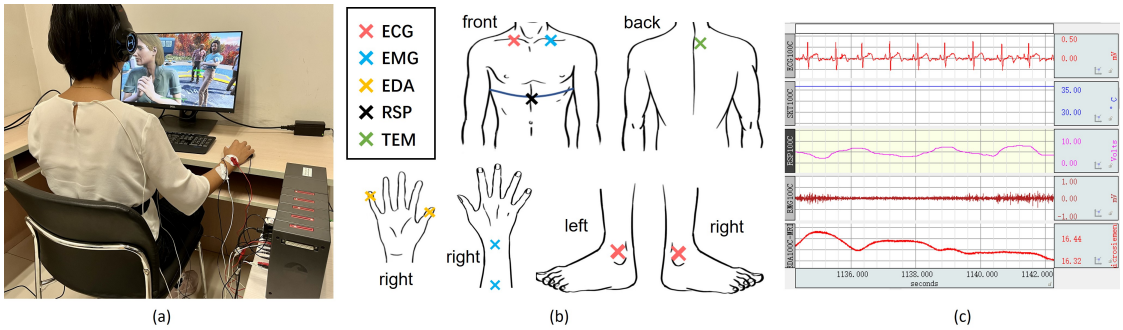

Figure 2: (a) Experimental setting, (b) electrode placement, and (c) example of physiological signals.

Table 1: The Overall Challenge Experience and Emotional Responses for each game scenario.

Scenario	Cronbach's α	Cognitive Challenge	Emotional Challenge	Physical Challenge	Decision-Making	High-scoring emotional responses
A	0.935	4.05 (1.06)	5.43 (0.89)	4.24 (1.89)	5.14 (1.03)	anxiety, sadness, doubt, tension, shock
B	0.924	3.79 (1.03)	5.32 (0.66)	4.09 (1.34)	5.18 (0.86)	trust, love, affection, courage, interest
C	0.933	4.88 (0.91)	5.70 (0.60)	5.22 (1.20)	5.51 (0.65)	tension, anxiety, interest, courage, empathy

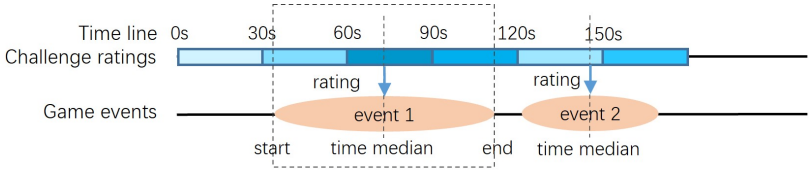

Figure 3: The challenge experience with each event is indicated by the challenge rating at the time median of the event.

achieved in event 18 when fighting against Kellogg – the player needs to find Kellogg inside the building.

4.2.2 Co-existing and independent types of experienced challenge. Figures 4-a.1, 4-a.2 and 4-a.3 also show that **with many game events, different types of perceived challenge can co-exist with each other**. For example, in scenario A, cogC and phyC co-exist in event 10 when escaping from the Vault and event 18 during demanding fighting against Kellogg; emoC and DM co-exist in event 23-a when joining the son and event 23-b when leaving the son. In scenario B, cogC, emoC and DM nearly co-exist in events 14-15 and events 18-19-a during which the player needs to talk to NPCs to get information and make choices; for event 19-b, two participants decided to kill Danse directly and they reported high levels of emoC and DM. In scenario C, for event 2 when answering a set of psychological questions, cogC, emoC and DM co-exist; and for events 10-14 during which the player is involved in searching and fighting, cogC and phyC co-exist; cogC, emoC and DM co-exist in events 16-18 when the player is involved in important conversations and decision making; for event 19-b1, two participants decided to kill Dan and they peaked at almost all four types of challenge. Noting that some types of experienced challenge can also exist in a relatively independent way, such as the emoC in event 7 of scenario A, the phyC in event 1 of scenario B, the cogC in event 7 of scenarios B and C.

4.2.3 Individual rating differences. Figures 4-b.1, 4-b.2 and 4-b.3 show that **although the basic settings of each game scenario were the same, different participants rated quite different levels of challenge experience with the same game event**. This may be due to the varied skill, experience with and motivation in the game play between participants. Figure 4-b.1, for example, shows that with event 7 of witnessing the death of the spouse and the robbery of the son, participants reported different levels of challenge experience. The maximum, minimum, median, first quartile and third quartile for the rating of emoC with event 7 are: 5.89, 2.89, 5.00, 4.67 and 3.86 respectively. Another example, as shown in Figure 4-b.2, with event 1 of fighting against a group of feral ghouls, the maximum, minimum, median, first quartile, third quartile and outlier for the rating of phyC are: 6.00, 3.20, 5.40, 6.00, 4.60 and 1.70 respectively. In this study, considering the significant individual differences which may exist in understanding the challenge, data of outliers are also included for the following analysis and detection.

5 CHALLENGE DETECTION APPROACH

5.1 Data preparation

5.1.1 Physiological signal pre-processing. Signal pre-processing was conducted with AcqKnowledge software of

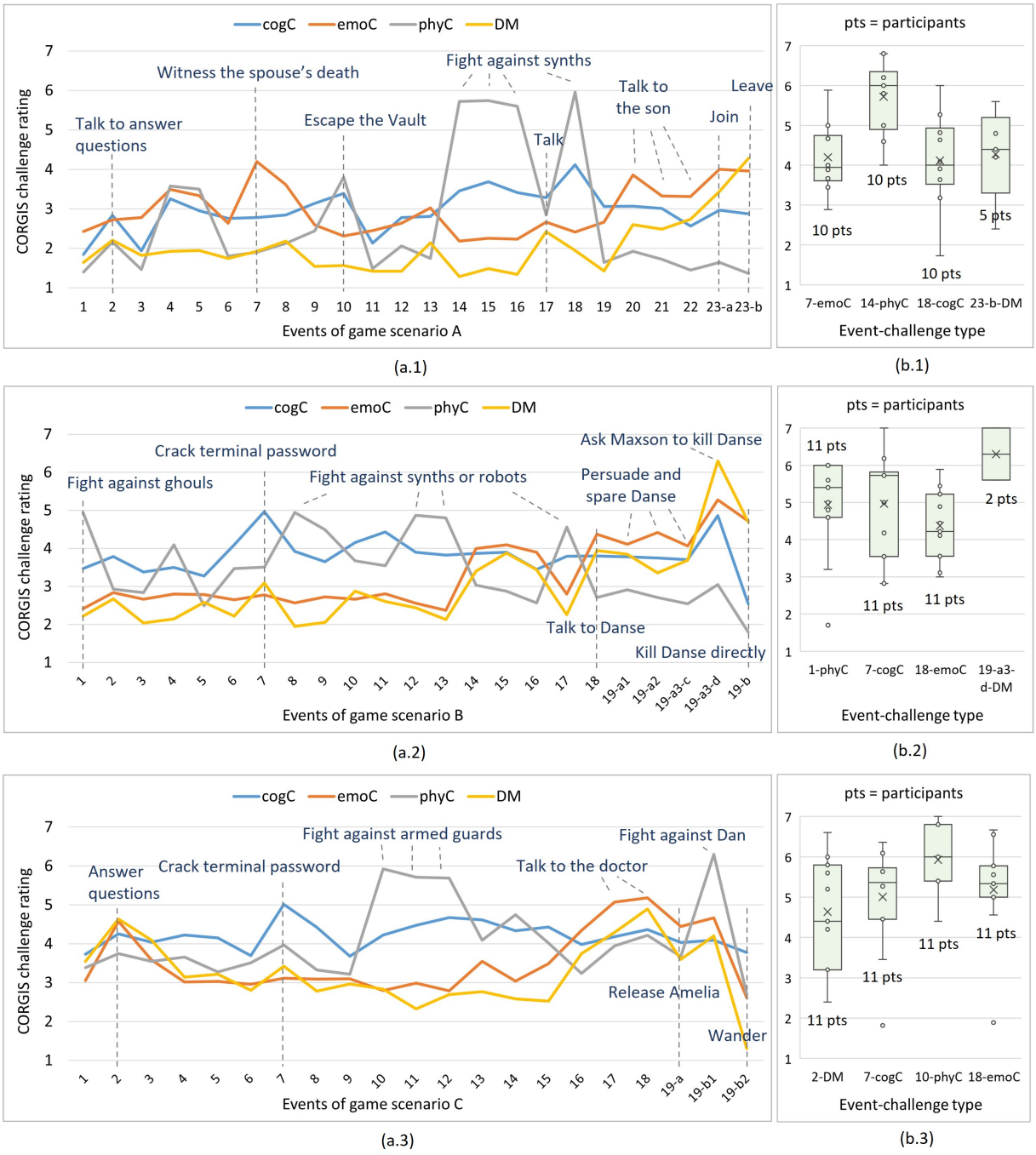

Figure 4: (a.1, a.2, a.3) The mean ratings with each event and (b.1, b.2, b.3) the example ratings of different participants.

Biopack[3]. Specifically, for ECG data, z-score normalization was first applied to reduce individual differences by making each participant's data have unit variance and zero mean. Median filters of 200-ms and 600-ms width were used to remove baseline wander caused by the participant's movement or breathing [33]. Then, R-wave peaks were detected by Pan-Tompkins QRS detector [53], and the RR intervals (heart rate variability, HRV) were computed as the time between two successive R-wave peaks. For EDA data, baseline wander was removed by a high pass filter [61] with 0.02

[3]Detailed information can be found at https://www.biopac.com/product/acqknowledge-software/

Hz. For RSP data, a band-pass filter with 0.1 Hz to 5 Hz was used to remove unusual breathing cycles. Penh Analysis of Acqknowledge was then used to detect respiratory circulation by setting the 65% of exhalation volume as a threshold.

5.1.2 Feature extraction. A total of 80 physiological signal features were extracted using Matlab (these can be found in Appendix C). For each kind of signal of each participant, z-score normalization was applied first to reduce individual differences. Then, each frame of physiological features was computed using signals inside a sliding time window with 10-second width and 5-second shift (see Figure 5). Although there is still no consensus on the optimal size of the time window with peripheral physiological analysis [64], the most commonly used time windows are 60, 30 and 10s [42]. In this study, to investigate the fine-grained dynamic physiological response in greater depth, the time window was set to a 10s width. Further, to avoid any possible failure to capture information between two consecutive time windows, a 5s sliding shift was used.

With feature extraction, specifically, ECG time-domain features were computed using the R-wave peak values, RR intervals, as well as their first-order differences (Diffs) between two successive tokens. To compute ECG frequency-domain features, the RR intervals were re-sampled to 8 Hz using a cubic spline interpolation [52]. Then, Fast Fourier Transform-based method on 256 samples for each frame (reset to 32-second window width) was applied and the frequency bands were grouped into three ranges: very-low-frequency (VLF) (0–0.04 Hz), low-frequency (LF) (0.04–0.15 Hz), and high-frequency (HF) (0.15–0.4 Hz) [33]. The power and power spectral density (PSD) in these ranges were utilized as features. For RSP data, the same time-domain features as ECG data were extracted. For TEM data, the amplitudes and their Diffs were used. For EMG data, the signal was first down-sampled to 32 Hz, followed by a 6-level wavelet decomposition with Daubechies5 [11]. For EDA data, time- and frequency-domain features [61] of low-frequency (LF) (0.02–0.5 Hz), high-frequency (HF) (0.5–1 Hz), and very-high-frequency (VHF) (>1 Hz) ranges were computed.

Finally, to remove the potentially noisy physiological periods during game play, only the feature frames located in the game event periods (marked by the start and end time of each game event) were deployed for detection (this can be found in Figure 5).

5.1.3 Training labels. Participants' perceived challenge ratings were used as labels for model training. As Figure 5 shows, the physiological feature samples were determined by the feature frames and challenge ratings located in the game event periods. In this study, 12,436 labeled data samples were determined over the 32 participants in total.

In the following detection work, we could directly use the challenge ratings of each data sample as a training label, and the model would directly predict the challenge ratings for evaluation data. However, with perceived challenge experience, the demonstrated individual rating differences make the distributions of challenge ratings for different participants rather different from each other. This may make the averaged distribution represented by the trained models significantly mismatch to that of the target participants. To address this issue, we compute the mean μ_y and standard deviation σ_y for all challenge ratings $\mathbf{y}$ within each participant, and normalize the training label of that player using the following formula

$$\mathbf{y}_{\text{train}}^{(\text{norm})} = \frac{\mathbf{y}_{\text{train}} - \mu_y}{\sigma_y}. \tag{1}$$

For evaluation, we invert the normalization for the predicted labels to produce the predicted challenge ratings with original distribution, which is computed as

$$\hat{\mathbf{y}}_{\text{test}} = \hat{\mathbf{y}}_{\text{test}}^{(\text{norm})} * \sigma_y + \mu_y \tag{2}$$

for each participant.

5.2 Machine learning methods

As the first study to explore perceived challenge detection with physiological signals, multiple widely used machine learning methods were applied and tested. We used regression models to predict the challenge ratings using the physiological features introduced in section 5.1.2. These methods include Linear Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting Decision Trees (GBDT), Random Forest (RF), and Deep Neural Networks (DNN) with different architectures. The detailed model descriptions and configurations are provided in Tables 14, 15 and 16 of Appendix D. Training and evaluation of all the models are implemented using Matlab.

5.3 Challenge evaluation metrics

5.3.1 RMSE and MAE. To evaluate the machine learning models for challenge detection, the widely used metrics Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are computed between the true and predicted challenge ratings of evaluation data. They reveal how far the predicted ratings are off from the true ratings. For the ith challenge they are computed as

$$\text{RMSE}_i = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_{n,i} - \hat{y}_{n,i})^2} \tag{3}$$

$$\text{MAE}_i = \frac{1}{N}\sum_{n=1}^{N}|y_{n,i} - \hat{y}_{n,i}| \tag{4}$$

where N denotes the number of data samples for evaluation.

To determine whether a certain type of challenge has been activated, an additional 2-class classification metric is exploited for the evaluation. This metric may facilitate a simpler and more discrete understanding of challenge, which could be easily utilized in some game challenge adjustment designs and applications.

5.3.2 Challenge activation detection: five 2-class classifications. The challenge activation detection is a 2-class classification for each challenge to detect whether the challenge experience is activated to a medium level. Based on the challenge ratings we used, each challenge with a rating equal to or higher than 4 is regarded as activated. This produces four 2-class classifications for all four challenge types. We also added the fifth 2-class "challenge" label denoting "No Challenge", which is activated when all four challenge types have ratings less than 4. The proportions of the activated (positive) class of cogC, emoC, phyC, DM and No Challenge in all data are (41.2%, 30.4%, 34.9%, 23.0%, 39.7%).

For the 2-class classification, the evaluation data is separated into four parts: True Positive (TT), True Negative (TN), False Positive

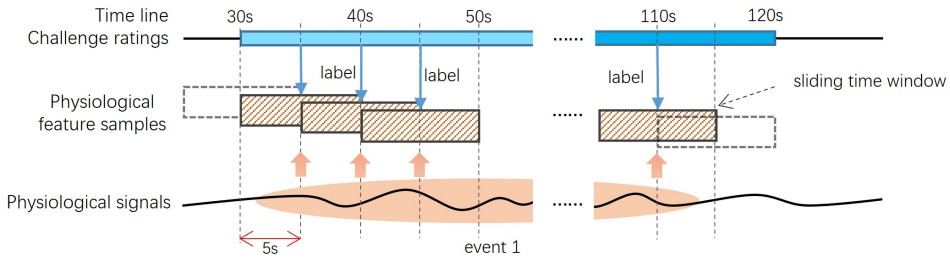

Figure 5: Physiological feature samples were determined by features frames and challenge ratings located in the game event periods.

(FP), and False Negative (FN). The accuracy is hence computed as

$$\text{Acc}_{2-c} = \frac{TT + TN}{TP + TN + FP + FN}. \tag{5}$$

and the 2-class F1 score of the activated (positive) class is computed as

$$F1_{2-c} = 2\frac{\text{Pre} * \text{Rec}}{\text{Pre} + \text{Rec}} \tag{6}$$

where $\text{Pre} = \frac{TP}{TP+FP}$ and $\text{Rec} = \frac{TP}{TP+FN}$ denote the precision and recall of the activated (positive) class respectively.

6 RESULTS OF DETECTION

6.1 Performance of evaluation

6.1.1 10-folds cross validation. We employed 10-folds and leave-one-participant-out cross validation methods for evaluation in this section. In the 10-folds cross validation, the whole dataset with 12,436 labeled data samples were equally and randomly split into 10 subsets. We trained the models 10 times. For each time we used 1 subset as the evaluation data and the other 9 subsets as the training data. Finally, each subset was used for evaluation once. The evaluation results of all 10 subsets were averaged for each challenge.

The performance of different machine learning models described in section 5.2 and Appendix D in 10-fold cross validation for all challenges is shown in Table 2. For each machine learning model, e.g., the LR1, all results in different metrics are generated by the same trained model. Four main trends are presented in the Table:

(1) The machine learning models using the label normalization method introduced in section 5.1.3 by equation (1) and (2) achieve the average performance in RMSE and MAE as low as 0.84 and 0.61 respectively. These results imply that, in the best case, the predicted challenge rating may have the error of ±0.84 on average and ±0.61 as the median. Evaluating by challenge activation detection, the average detection accuracy over all challenges is up to 85.1%. All the models significantly outperform both the random selection with accuracies of 50% and the constant model with accuracies of 66.2%, where the constant model simply predicts the class with the highest proportion for each challenge. In addition, we also train models without using label normalization, which produce much worse results than those shown in Table 2 with label normalization.
(2) For the performance on each challenge as well as for all challenge types, the feed-forward DNN models using residual connection outperform most other models. Moreover, the multi-task learning (MTL) for DNN training using the additional task of challenge activation detection consistently improves the performance and yields the best model denoted by "DNN4" (detailed configuration can be found in Table 16 in Appendix D). However, the deeper DNN (9-hidden-layer) has no advantage over the 5-hidden-layer models while having a much larger number of parameters. Meanwhile, with higher interpretability than the black-box DNN model, the Random Forest model produces a sufficiently good performance similar to that of DNN4.
(3) In general, the performance difference among different challenge types is not significant. Specifically, comparing the RMSE and MAE among different challenges, detecting the rating level of cogC and emoC is relatively easier (with lower RMSE and MAE) than phyC and DM. With the challenge activation metric, detection for cogC and DM are relatively easier (with higher Acc_{2-c}). Using this metric, the detection of 'No Challenge' has the worst performance.
(4) Most of the models produce standard deviations (Std) smaller than 0.02 in RMSE and MAE over the 10 subsets.

Figure 6 shows examples of predicting challenge ratings using the DNN4 for each challenge type. We randomly select 50 samples from the fifth subsetand predict the ratings using the models trained with the other subsets. These samples are not sequential and are from different participants. Here, the ratings predicted by the DNN4 (the blue points) yield a good fit to the true ratings (red points) for all four challenges.

6.1.2 Leave-one-participant-out cross validation. The 10-fold cross validation shows the in-domain challenge detecting performance when all target participants are seen in the training data. However, when in practice, the target participants could be new users unseen in the training stage. The leave-one-participant-out cross validation is similar to this, in which the data of each participant was treated as an evaluation set once, while the data of the other 31 participants was used for training. Therefore, the evaluation data at each time was totally out-of-domain and unseen in the training data. Finally, the evaluation results of all 32 participants were averaged for each challenge type.

Table 2: Average performance of 10-fold cross validation over all challenges for different machine learning models with label normalization method introduced in section 5.1.3 by equation (1) and (2). The Acc_{2-c} and $F1_{2-c}$ are metrics of challenge activation detection. Std denotes standard deviation; Chal denotes Challenge; noC denotes No Challenge. The configurations of different models are provided in Tables 14, 15 and 16 of Appendix D.

Metric	Chal		LR1	LR2	KNN	SVM	DT	GBDT	RF	NN	DNN1	DNN2	DNN3	DNN4	DNN5
RMSE	cogC	Mean	0.92	0.87	0.86	0.83	0.91	0.77	0.76	0.89	0.81	0.77	0.77	**0.74**	0.76
		Std	0.02	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.02	0.02	0.01	0.03
	emoC	Mean	0.91	0.88	0.88	0.84	0.89	0.75	0.76	0.89	0.81	0.77	0.78	**0.75**	0.76
		Std	0.02	0.02	0.02	0.01	0.02	0.02	0.03	0.02	0.02	0.02	0.01	0.02	0.01
	phyC	Mean	1.23	1.19	1.18	1.13	1.27	1.09	1.08	1.17	1.09	1.03	1.04	**1.00**	1.03
		Std	0.02	0.03	0.02	0.03	0.03	0.02	0.04	0.03	0.03	0.04	0.03	0.01	0.03
	DM	Mean	1.06	1.05	1.03	1.02	1.06	0.91	0.91	1.04	0.94	0.90	0.90	**0.87**	0.89
		Std	0.02	0.02	0.02	0.03	0.02	0.03	0.04	0.02	0.02	0.02	0.03	0.04	0.03
	Avg	Mean	1.03	1.00	0.99	0.96	1.03	0.88	0.88	1.00	0.91	0.87	0.87	**0.84**	0.86
		Std	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.03
MAE	cogC	Mean	0.72	0.67	0.64	0.64	0.68	0.59	0.59	0.70	0.63	0.58	0.59	**0.56**	0.57
		Std	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.02
	emoC	Mean	0.72	0.67	0.66	0.64	0.67	0.57	0.59	0.70	0.62	0.58	0.59	**0.56**	0.57
		Std	0.02	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.01
	phyC	Mean	0.95	0.89	0.84	0.83	0.93	0.81	0.82	0.90	0.80	0.75	0.74	**0.71**	0.72
		Std	0.01	0.02	0.02	0.02	0.03	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.02
	DM	Mean	0.81	0.79	0.73	0.72	0.77	0.68	0.68	0.79	0.68	0.65	0.64	**0.62**	0.63
		Std	0.02	0.01	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02
	Avg	Mean	0.80	0.75	0.71	0.71	0.76	0.66	0.67	0.77	0.68	0.64	0.64	**0.61**	0.62
		Std	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Acc_{2-c}(%)	cogC	Mean	79.8	82.8	82.4	82.9	82.1	85.8	**87.2**	81.5	84.2	85.3	85.6	86.8	86.1
		Std	0.9	0.7	0.6	1.0	1.2	1.3	0.7	1.4	1.3	0.6	1.0	0.9	1.3
	emoC	Mean	81.1	81.9	82.7	82.9	82.0	84.7	84.3	81.8	83.7	84.4	85.0	85.1	**85.6**
		Std	0.9	0.8	0.7	0.7	1.1	0.9	1.0	0.6	1.0	0.7	0.6	1.1	0.9
	phyC	Mean	79.1	81.5	80.9	81.1	79.1	83.1	82.7	80.8	82.5	84.1	84.0	**85.1**	84.6
		Std	1.1	1.1	1.0	1.1	1.3	0.7	1.7	1.2	1.2	1.1	0.8	0.5	0.7
	DM	Mean	83.5	84.4	84.2	83.6	83.8	86.2	85.7	83.6	85.2	86.1	86.5	**87.3**	**87.3**
		Std	1.2	0.8	0.9	1.1	1.1	0.7	1.0	1.1	0.9	1.2	0.8	1.1	0.5
	noC	Mean	71.6	78.4	77.4	75.1	77.8	79.8	78.0	75.3	78.1	80.1	80.6	**81.4**	81.0
		Std	1.3	0.9	0.8	0.8	1.6	1.4	1.3	1.8	1.7	1.4	1.0	1.1	0.9
	Avg	Mean	79.0	81.8	81.5	81.1	81.0	83.9	83.6	80.6	82.7	84.0	84.3	**85.1**	84.9
		Std	1.1	0.9	0.8	1.0	1.3	1.0	1.1	1.2	1.2	1.0	0.8	1.0	0.9
$F1_{2-c}$(%)	Avg	Mean	63.8	70.1	69.1	66.7	68.3	73.3	72.0	66.7	70.9	73.2	74.1	**75.6**	75.5
		Std	2.2	0.9	0.6	1.3	1.1	1.3	1.5	2.0	1.0	1.2	1.1	0.9	0.7

The performance of different machine learning models for all challenge types in leave-one-participant-out cross validation is shown in Table 3. For each machine learning model, e.g., the LR1, all results in different metrics are generated by the same trained model. We observed different trends from the 10-folds cross validation:

(1) Consistent performance degradation is obtained compared to the models in 10-folds cross validation. This may be caused by the high mismatch between training and test data. We expect this degradation can be alleviated using additional data collected from a larger sample.
(2) Different models have different performances evaluated by different metrics. Especially, the simple methods, for example, Linear Regression1 produces relatively good results in all metrics and achieves the best average accuracy of challenge activation detection. This may be due to the overfitting of the more complex model, such as the DNN models. The Random Forest model achieved the best results across all metrics.
(3) With the various performance among different challenge types, we observe a similar trend to the 10-folds cross validation case, except that the challenge activation detection for cogC is harder (with lower Acc_{2-c}) than emoC, phyC and DM.
(4) The results have much larger Stds over different participants than those in 10-folds cross validation. This implies the high variability among participants. This variability among different participants is shown in Table 4 using the results obtained from Random Forest model in leave-one-participant-out cross validation.

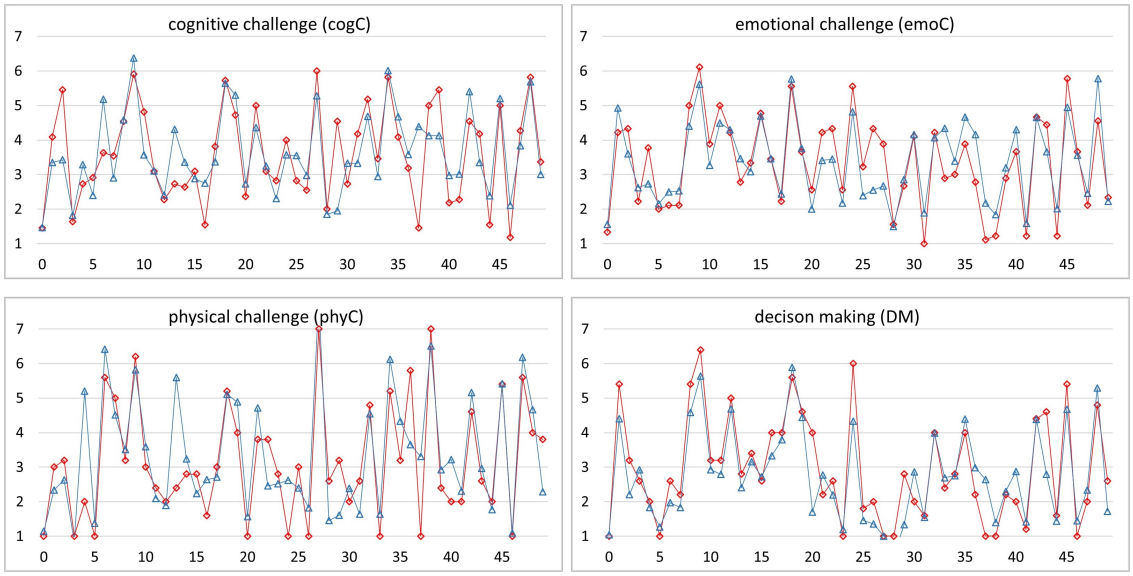

Figure 6: Examples of prediction to the challenge ratings by DNN4 with lab normalization for each challenge in 10-fold cross validation. 50 samples are randomly selected from the fifth subsets and predicted by the model trained with the other subsets. The red points denote the true ratings and blue points denote the predicted ratings.

Figure 7 shows examples of predicting challenge ratings using Random Forest for each challenge type. For each challenge type, 200 sequential samples are selected from different participants and predicted by the model trained using other participants' data. It shows that the predicted ratings for emoC and phyC fit the true ratings well. The prediction for DM has a trend with a significant positive correlation to the true ratings. Nevertheless, only a weak correlation was obtained by the predicted trend for cogC compared to the true ratings. This is consistent with the fourth trend discussed above that the cogC has lower Acc_{2-c} than emoC, phyC and DM.

6.2 Feature importance

To investigate the ability and importance of different physiological features for representing challenge information, we adopt Linear-Regression-based feature selection and Random-Forest-based feature selection. In Linear-Regression-based feature selection, we train a Linear Regression model using all data with 80 features from 32 participants. The physiological features with interception are used as input features, namely $\mathbf{x} = [\mathbf{x}, 1]^T$. We select 44 most important features by keeping only the features with significant coefficients ($p < 0.05$) for at least one challenge. They are shown in the upper part of Figure 8. The trained model interceptions for the cogC, emoC, phyC, and DM are $(0.40, 0.14, 0.02, -0.35)$, which represent the normalized ratings on zero point (also the average point due to the z-score) of physiological features. The other coefficient values represent the increments of the corresponding normalized challenge ratings based on the interception when the related features increase by 1.

In Random-Forest-based feature selection, we train a Random Forest model using all data and normalized labels. The importance of a feature is computed as the increase on MSE when permuting the values of the feature for the data without being selected in training a tree [17]. Then, we sort the features by their importance values and keep the first 20 features for each challenge type. Finally, we select 34 features in total. They are shown in the lower part of Figure 8.

Table 5 shows the top 10 important features for each challenge selected by the two methods. Further, 24 most important features are obtained by the intersection between selections of the two models, which are shown by red boxes in Figure 8. Figure 8 also shows the importance of different physiological channels. Considering both selection methods, the ECG and EMG are the two most important channels. Finally, we train the machine learning models to view the performance of the selected features, which is shown in Table 6. Compared to the corresponding models in Tables 2 and 3 using all 80 features, the selected features obtain similar results and even shows slight improvement on some metrics.

7 DISCUSSION

Perceived challenge is a key player experience that arises from one's interaction with a game's intrinsic challenges at a particular skill level. As the player progresses through the game, this experience can change dynamically with time, based on the in-game content and with skill acquisition/mastery on the player's behalf. Therefore, it is important to be able to assess this experience in real-time to balance game difficulty or to adjust the in-game content to match the player's skills and to cater to the individual preferences and differences between players.

In this paper, we propose a novel pipeline to assess in-game challenge experience using physiological measures. To do so, we first conducted an experimental study and found that players' perceived challenge changes could fluctuate over time, co-existing with one

Table 3: Average performance of leave-one-participant-out cross validation over all challenges for different machine learning models. Label normalization is applied to all the models. The Acc_{2-c} and $F1_{2-c}$ are metrics of challenge activation detection. LR denotes Linear Regression; DT denotes Decision Tree; RF denotes Random Forest; noC denotes No Challenge. The configurations of different models are provided in Tables 14, 15 and 16 of Appendix D.

Metric	Chal		LR1	LR2	KNN	SVM	DT	GBDT	RF	DNN4
RMSE	cogC	Mean	0.91	1.11	0.95	0.94	1.09	1.00	**0.90**	0.92
		Std	0.27	0.28	0.26	0.27	0.32	0.28	0.27	0.27
	emoC	Mean	**0.91**	1.13	0.97	0.96	1.08	1.01	**0.91**	0.93
		Std	0.25	0.28	0.26	0.26	0.27	0.28	0.27	0.26
	phyC	Mean	1.20	1.44	1.27	1.22	1.39	1.30	**1.17**	1.23
		Std	0.33	0.36	0.35	0.33	0.35	0.35	0.32	0.35
	DM	Mean	**1.08**	1.32	1.15	1.13	1.26	1.17	1.09	1.11
		Std	0.33	0.35	0.34	0.35	0.34	0.33	0.33	0.35
	Avg	Mean	1.03	1.25	1.09	1.06	1.20	1.12	**1.02**	1.05
		Std	0.29	0.32	0.31	0.3	0.32	0.31	0.3	0.31
MAE	cogC	Mean	**0.74**	0.89	0.77	0.76	0.89	0.82	0.75	0.75
		Std	0.23	0.23	0.22	0.23	0.27	0.24	0.23	0.23
	emoC	Mean	**0.75**	0.91	0.80	0.79	0.86	0.82	**0.75**	0.77
		Std	0.22	0.23	0.23	0.22	0.22	0.23	0.22	0.22
	phyC	Mean	0.97	1.13	0.99	0.96	1.10	1.03	**0.94**	0.97
		Std	0.28	0.3	0.29	0.28	0.29	0.29	0.27	0.29
	DM	Mean	**0.86**	1.05	0.91	0.88	0.98	0.93	**0.86**	0.87
		Std	0.28	0.29	0.29	0.3	0.29	0.28	0.29	0.3
	Avg	Mean	0.83	1.00	0.87	0.85	0.96	0.90	**0.82**	0.84
		Std	0.25	0.26	0.26	0.26	0.27	0.26	0.25	0.26
$Acc_{2-c}(\%)$	cogC	Mean	**77.8**	74.2	76.2	77.0	74.2	76.0	77.3	77.1
		Std	22.3	20.0	22.3	22.0	20.6	21.1	22.7	22.4
	emoC	Mean	**80.6**	75.9	78.9	79.7	77.1	79.0	80.1	80.0
		Std	16.3	17.3	16.6	16.3	16.1	16.1	17.1	16.2
	phyC	Mean	77.2	74.0	75.6	76.7	74.1	76.6	**78.8**	77.0
		Std	13.5	14.3	15.5	15.2	15.1	13.6	13.3	14.6
	DM	Mean	**81.3**	77.5	79.9	80.2	78.4	79.9	81.0	80.5
		Std	15.4	16.5	16.4	16.6	17.3	16.1	15.9	16.0
	noC	Mean	71.3	**72.9**	69.5	71.3	72.7	72.4	70.7	71.5
		Std	20.4	16.7	19.3	18.4	17.8	18.1	20.0	18.8
	Avg	Mean	**77.6**	74.9	76.0	77.0	75.3	76.8	**77.6**	77.2
		Std	17.6	16.9	18.0	17.7	17.4	17.0	17.8	17.6
$F1_{2-c}(\%)$	Avg	Mean	37.3	38.7	36.3	36.8	38.5	**39.6**	38.0	34.8
		Std	21.6	18.5	20.8	20.4	17.9	19.9	21.2	17.4

another and displaying significant differences between individual players. We then detected different types of challenge experience based on 32 players' physiological measures and their continuously reported challenge experience by using multiple machine learning methods. Our results show that the majority of these methods can achieve an accuracy of around 80% on challenge activation. Finally, by adopting feature importance techniques, a set of 24 challenge-rated physiological features were selected and refined to indicate different challenge types.

Generally, we hope that the pipeline would inspire game researchers, designers and developers in the field of human-computer interaction to view challenge as an important player experience. The pipeline also suggests that challenge experience could be practically assessed through physiological sensing methods. With this pipeline, it is possible to explore perceived challenge in digital games more thoroughly in future studies to understand how players perceive different game challenges and to develop novel biofeedback game interactions based on challenge prediction. First, we propose to establish datasets towards the wide spectrum of challenge types. A specific dataset may comprise multimodal user data, such as physiological signals like EEG, EDA, EMG etc., as well as more continuous or discrete user data including other player experiences. These data may be induced by various challenge structures as well as different difficulty modes, involving users with varied skill levels, motivations and backgrounds. Second, akin to our pipeline, more sophisticated and advanced AI-based models and methods should be explored to compute and predict perceived challenge, for instance, by employing AI and mathematical techniques to model the temporal aspects of how perceived challenge develops over time. Moreover, considering that emotional concepts

Table 4: Performance for different participants using Random Forest model in leave-one-participant-out cross validation. Each column shows the result of one participant.

Scenario	Metric	Participants										
A	RMSE	1.04	0.95	0.87	0.74	1.01	**0.55**	0.70	0.71	1.02	1.03	
	MAE	1.35	1.14	1.15	0.87	1.21	**0.76**	0.89	0.83	1.22	1.22	
	Acc_{2-c}	85.1	60.8	87.8	68.2	62.7	**95.2**	94.8	91.6	83.7	83.1	
	$F1_{2-c}$	23.6	22.0	27.8	29.4	23.6	**38.7**	19.7	22.5	27.4	25.4	
B	RMSE	1.12	0.95	0.98	0.92	0.92	0.63	0.80	0.76	0.80	0.40	**0.37**
	MAE	1.32	1.17	1.21	1.07	1.19	0.75	0.95	0.93	0.99	0.54	**0.**
	Acc_{2-c}	58.3	66.0	59.5	63.7	81.8	65.8	62.6	73.4	81.1	90.6	**98.1**
	$F1_{2-c}$	32.2	38.1	29.0	40.3	27.5	31.8	26.2	19.9	18.2	**94.5**	59.5
C	RMSE	0.70	0.78	0.62	0.78	0.68	1.09	**0.55**	0.94	0.81	1.21	0.95
	MAE	0.86	0.93	0.85	1.00	0.87	1.38	**0.65**	1.23	0.97	1.48	1.13
	Acc_{2-c}	**99.6**	81.4	93.7	77.1	70.1	78.1	87.5	83.2	71.0	55.6	70.9
	$F1_{2-c}$	**99.8**	61.0	37.5	68.6	44.0	17.3	72.6	26.8	45.5	20.9	43.3

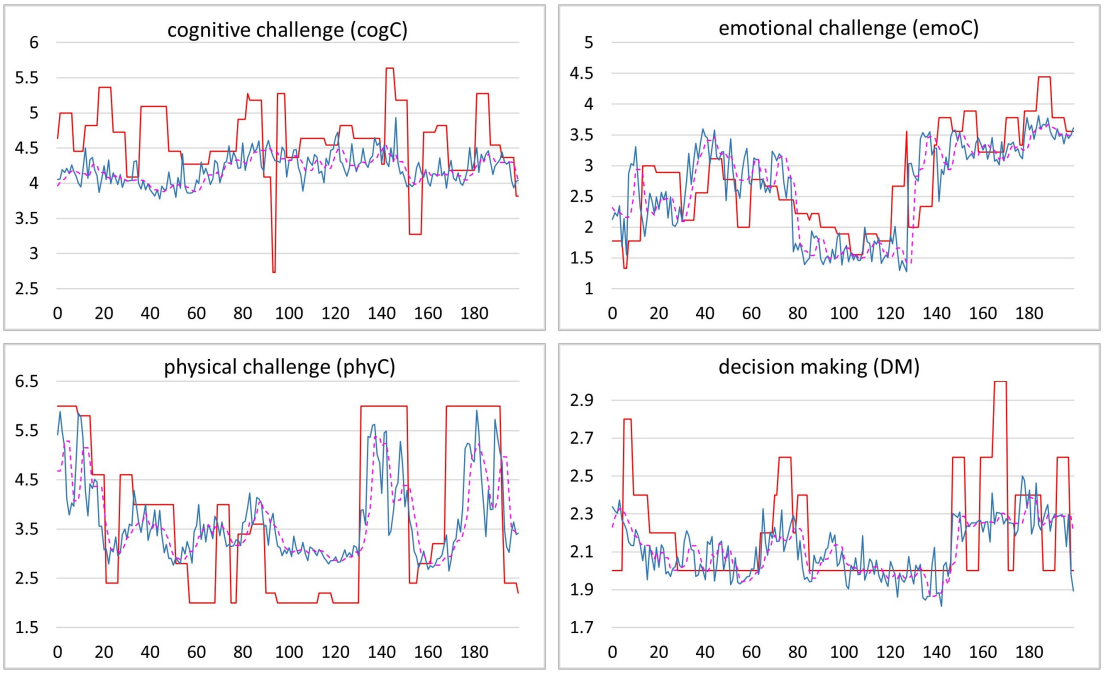

Figure 7: Examples of prediction to the challenge ratings by Random Forest for each challenge in leave-one-participant-out cross validation. For each challenge type, 200 sequential samples are selected from different participants and predicted by the models trained with the other participants. The red curves denote the true ratings and blue curves denote the predicted ratings. The magenta curves are the predicted ratings smoothed by a median filter such that the predicted trends are shown more clearly.

that have been extensively used in affective gaming tend to offer some ambiguity, the proposed pipeline may also assist in bridging the gap between challenge-related game design elements and complex emotionally-charged player experiences. For instance, it is believed that when playing modern digital games with emotional challenge, players would first encounter the emotional challenge, which is then thought to lead to complex and diverse emotional experiences.

7.1 In-game challenge perception

The descriptive results of the experimental study demonstrate that it is possible to discern players' perceived challenge experience dynamically as it changes with time and that challenge can be perceived with significant individual differences. We also observed that different types of challenge experience can co-exist or manifest themselves relatively independently. These descriptive findings

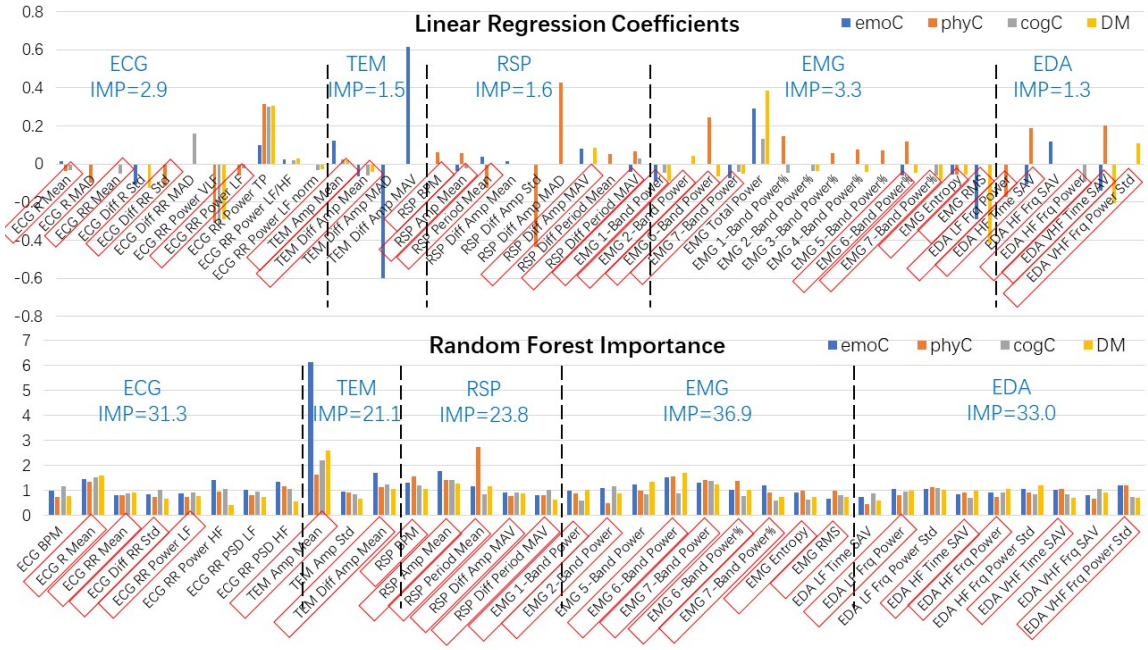

Figure 8: Upper: significant coefficients of selected features for Linear Regression. Lower: importance of selected features by Random Forest. The physiological channel IMP denotes the importance of the physiological channel computed by summing the absolute importance values (or significant coefficients) of the selected features in that channel. The features in red boxes are the intersection between the two selections.

Table 5: The top-10 important features for each challenge type selected by sorting the absolute significant coefficient (COEF) and importance (IMP) values of Linear Regression (LR) and Random Forest (RF), respectively.

	cogC		emoC		phyC		DM	
	Top-10 features	COEF	Top-10 features	COEF	Top-10 features	COEF	Top-10 features	COEF
LR	ECG RR Power VLF	-0.31	TEM Diff Amp MAV	0.62	RSP Diff Amp Std	-0.44	EMG RMS	-0.42
	ECG RR Power TP	0.30	TEM Diff Amp MAD	-0.60	RSP Diff Amp MAD	0.43	EMG Total Power	0.38
	ECG Diff RR MAD	0.16	EMG RMS	-0.29	ECG RR Power TP	0.31	ECG RR Power TP	0.30
	EMG Total Power	0.13	EMG Total Power	0.29	ECG RR Power VLF	-0.31	ECG RR Power VLF	-0.29
	EDA HF Frq Power	-0.09	EDA VHF Time SAV	-0.14	EMG 6-Band Power	0.25	EDA VHF Time SAV	-0.21
	EMG 7-Band Power%	-0.06	TEM Amp Mean	0.12	EDA VHF Time SAV	0.20	ECG Diff R Std	-0.13
	TEM Diff Amp Mean	-0.06	EDA HF Frq SAV	0.12	EDA HF Time SAV	0.19	EDA VHF Frq Power Std	0.11
	ECG RR Mean	-0.05	ECG Diff R Std	-0.11	EDA LF Frq Power	-0.17	EMG 1-Band Power	-0.09
	EMG 1-Band Power%	-0.05	EDA HF Time SAV	-0.11	EMG 1-Band Power%	0.15	EMG 7-Band Power%	-0.08
	EMG 1-Band Power	-0.05	ECG RR Power TP	0.10	RSP Period Mean	-0.13	RSP Diff Amp MAV	0.08
	Top-10 features	IMP	Top-10 features	IMP	Top-10 features	IMP	Top-10 features	IMP
RF	TEM Amp Mean	2.21	TEM Amp Mean	6.12	RSP Period Mean	2.73	TEM Amp Mean	2.58
	ECG R Mean	1.52	RSP Amp Mean	1.79	TEM Amp Mean	1.65	EMG 6-Band Power	1.70
	RSP Amp Mean	1.44	TEM Diff Amp Mean	1.70	EMG 6-Band Power	1.57	ECG R Mean	1.59
	EMG 7-Band Power	1.38	EMG 6-Band Power	1.52	RSP BPM	1.57	EMG 5-Band Power	1.34
	TEM Diff Amp Mean	1.25	ECG R Mean	1.47	EMG 7-Band Power	1.43	RSP Amp Mean	1.27
	RSP BPM	1.20	ECG RR Power HF	1.41	RSP Amp Mean	1.42	EMG 7-Band Power	1.23
	EMG 2-Band Power	1.19	ECG RR PSD HF	1.35	EMG 6-Band Power%	1.38	EDA HF Frq Power Std	1.23
	ECG BPM	1.18	EMG 7-Band Power	1.32	ECG R Mean	1.36	RSP Period Mean	1.16
	EDA LF Frq Power Std	1.09	RSP BPM	1.31	EDA VHF Frq Power Std	1.20	EDA HF Frq Power	1.07
	ECG RR Power HF	1.08	EMG 5-Band Power	1.25	ECG RR PSD HF	1.16	RSP BPM	1.06

Table 6: Performance of the Linear Regression1, Random Forest, and DNN4 models with label normalization as in Tables 2 and 3 using the selected features.

Model	Cross Validation	Selection Method	RMSE	MAE	Challenge-Act-Detect	
					Acc_{2-c} (%)	$F1_{2-c}$ (%)
Linear Regression1	10-folds	Linear Regression	1.03	0.80	78.9	63.7
		Random Forest	1.03	0.81	79.0	63.6
		Intersection	1.03	0.81	78.9	63.6
Random Forest		Linear Regression	0.88	0.67	83.4	71.5
		Random Forest	0.85	0.65	84.3	73.3
		Intersection	0.86	0.66	83.8	72.3
DNN4		Linear Regression	0.84	0.61	85.1	75.5
		Random Forest	**0.81**	**0.59**	**85.6**	**76.7**
		Intersection	0.82	0.60	85.4	76.1
Linear Regression1	Leave-one--participant--out	Linear Regression	1.02	0.82	77.7	37.2
		Random Forest	1.02	0.83	**77.8**	37.7
		Intersection	1.02	**0.82**	**77.8**	37.6
Random Forest		Linear Regression	**1.01**	**0.82**	77.6	37.9
		Random Forest	**1.01**	**0.82**	**77.8**	**38.3**
		Intersection	**1.01**	**0.82**	77.7	38.1
DNN4		Linear Regression	1.06	0.84	77.5	36.2
		Random Forest	1.06	0.85	77.1	35.2
		Intersection	1.06	0.85	77.2	35.2

highlight the importance of measuring players' challenge experience dynamically, comprehensively and on an individual basis.

The results of our experiment highlight significant correlations (measured by Pearson's correlation coefficient) between some types of reported in-game challenge experience. Specifically, a strong and positive correlation between emotional challenge (emoC) and decision making (DM) ($r = 0.783$, $p < 0.01$) challenge is prominent in all three game scenarios. One reason may be that there is a lack of discriminant validity with the two types of challenge due to their conceptually overlapping characteristics in digital games [21]. There is also a possibility that the three game scenarios in our experiment have some comparable emoC and DM characteristics. It is noteworthy that the two types of challenge have their unique contribution in building up game challenge. As shown in our experiment, emotional challenge often happens in conjunction with some kind of moral decision, however, this is not always the case. For example, in event 7 of scenario A when participants witnessed the death of their spouse without being able to act upon it in the game, they reported a high level of emoC but a low level of DM.

We also observed a moderate and positive correlation between cognitive (cogC) and physical challenge (phyC) ($r = 0.536$, $p < 0.01$). The correlation may be attributed to the co-existing challenge properties of the game itself. For example, in the game scenarios A and C, when players fight against synths/armed guards, they need to engage their cognitive efforts to progress or to adopt some fighting strategies. In our experiment, emoC and phyC also show a moderate but negative correlation with each other ($r = -0.420$, $p < 0.01$). This could be because, in the selected game scenarios, the physical challenge of fighting and emotionally challenging narrative aspects of the game are often presented separately from one another. Even though it is rare that these two kinds of activities can be found simultaneously in games, in general, as game design evolves, we believe that it is possible for physical challenge to co-exist with emotional challenge in ways other than employing emotionally complex narratives.

Another interesting observation we made looking at the data of players' overall challenge experience was that the levels of emoC and DM were significantly higher than those of phyC (scenario A and B) and cogC (scenario A, B and C) (see Table 1). However, this phenomenon could hardly be observed from the data of players' perceived in-game challenge (see Figure 4). With some in-game events, the phyC ratings were higher than those of emoC. This could be because physical challenge is able to induce hedonic player experience that focuses on momentary pleasures. Emotional challenge, on the other hand, is a type of eudaimonic player experience [13] which focuses on players' intrinsic need of fulfilment, meaning and long-term importance [47]. In our experiment, most events related to phyC took place midway through the game scenario, so it is possible that these experiences did not have a lasting effect on our participants throughout their whole game play.

7.2 Challenge-related physiological indicators

By using feature importance techniques, a set of challenge-rated physiological features effectively indicated different types of challenge experience. The Random Forest based selection mainly highlighted the important features without providing information on the effects. Meanwhile, the Linear Regression based selection showed the features' positive/negative effects for each challenge type. Therefore, the intersection subset of refined 24 important features are considered to be meaningful in answering the question of how the physiological signal reflects the perceived challenge types. This may not be the exact subset of the most important features, but it

suggests that the selected features are likely to be important, and vice versa. For instance, although the "TEM Diff Amp MAV" and "TEM Diff Amp MAD" features have unusually significant effects on emotional challenge when using Linear Regression, they have relatively low importance when using the Random Forest approach. Hence, these unusual effects are likely to be not reliable.

We also verified if the selected features were able to achieve similar results to the application of all 80 features. This finding contributed to the interpretability and analyzability of the models. As a simple example: the features "RSP BPM" and "RSP Period Mean" have significant positive and negative effects on physical challenge, respectively, as shown in the top part of Figure 8. This is consistent with the common experience that extensive physical activity may lead to shortness of breath. Furthermore, the negative effects of the "EMG Entropy" feature on emotional and physical challenge types suggest a relatively heterogeneous distribution (compared to the average state) of frequencies on EMG channel. Meanwhile, the coefficients of "EMG 6-Band Power%" show the opposite effect on these two challenge types on the high-frequency EMG band. The positive/negative effects between emotional challenge and decision making is consistent with the relatively high correlation between ratings of these two challenges. Moreover, the "ECG RR Mean" and "EDA HF Frq Power" features have a unique negative effect on cognitive challenge.

With these primary challenge-related physiological indicators, we hope that future work focusing on a deeper analysis and an investigation of the important features will allow for an exploration of challenge experience using neuro-scientific methods. For example, further exploration of the activation mechanism of users' peripheral nervous system when encountering different types of challenge. The results also provide evidence that more compact and fast models for challenge detection could be used in the future and might signify the possibility of using only parts of the physiological channels for challenge detection.

7.3 AI-supported challenge detection

To explore challenge detection using physiological signals, we have compared the effectiveness of multiple widely used machine learning methods (LR, KNN, SVM, DT, GBDT, RF, DNNs) to detect different types of perceived challenge. Both 10-folds and leave-one-participant-out methods were employed for giving implications in multi-aspects. Results of the challenge detection show that different machine learning models with label normalization are able to achieve the accuracy (ACC_{2-c}) of around 80% for the prediction performance, which is superior to results of the random selection and constant model. This suggests the feasibility of our tested pipeline for detecting challenge with physiological measures, including data collection, feature and label processing, as well as the applications of machine learning methods. These results are an important first step in demonstrating the potential of using these detection methods in future applications of real-time challenge detection and dynamic challenge/difficulty adaptation techniques.

In our study, although the results of 10-folds cross validation show that DNN with 5-hidden-residual-MTL performed best, methods with leave-one-participant-out cross validation cater more to the application of predicting different types of challenge of a new player. In this case, if a future study or application has a training dataset similar to our study, the simple model of linear regression and random forests should be appropriate methods as they offer the best combination of performance, interpretability, and model simplicity, both with full features and when using only the most significant features. More complicated models may encounter the problem of overfitting with limited user data. On the other hand, the results of 10-folds cross validation in this study may present us with an ideal prediction vision with large amounts of user data in the future. This means that if a larger data sample could be collected as training data, the DNN model should have the potential to achieve good performance (e.g., with $ACC_{2-c} \geq 85\%$) in both the 10-folds and leave-one-participant-out cases.

In the setting of 10-fold cross validation method, detection performance for all challenge types is sufficiently good yet the performance diversity among different challenge types is not significant, as shown in both Table 2 and Figure 6. With performance varying between different challenge types in leave-one-participant-out method, a similar trend to the 10-folds method can be observed in Table 3, except that the challenge activation detection for cogC becomes harder than the other three challenges. A similar pattern is shown in Figure 7 where only a weak correlation is obtained by the predicted trend for cogC compared to the true ratings. Considering the proportion of positive class (with rating ≥ 4) of cogC is the largest (41.2%) among all challenges, this performance degradation for cogC is not likely to be caused by overfitting. A possible reason is the connection between cognitive challenge and the cognitive abilities of individual players, which could result in a mismatch between training and test participants larger than in other challenge types in the setting of leave-one-participant-out method.

Last but nonetheless an important point, the different results produced by the 10-folds and leave-one-participant-out cross validation highlight the non-negligible differences between individual players. We expect this mismatch to be alleviated if the participant number of training data is increased. Nevertheless, we consider certain methods for dealing with individual differences to be important and useful, for example, the semi-supervised or unsupervised adaptation methods widely used to reduce a mismatch between target and source domains in the fields of machine learning and pattern recognition.

7.4 Limitations and future work

One limitation of our work is that the current methodologies focus on the dynamical detection of a wide range of challenge experiences – some temporal and contextual information about how a certain type of challenge develops over time has not been included in the modeling. Additionally, in our study, the segmentation of game events left out some minor details that, in our opinion, are not important when creating any type of challenge. These details, for example, seeking help or wandering around the open game world, may also influence the construction of and the development of challenge experience. Therefore, for modern digital games with multiple and complex challenge types, it is critical to shape or model challenge experience alongside the game progression. This would be helpful for understanding how challenge changes as the game progresses as well as understanding eudaimonic and other

player experiences related to game challenge, such as the palette of high-scoring emotional responses reported in this study.

The second limitation is the inherent bias of physiological signals. For example, heart rate (HR) and EDA measurements allow for an evaluation of player arousal, which increases their feasibility in commercial game production situations [23]. In our study, when evaluating players' experience of challenge, their reactions or responses largely occurred in high arousal states, which may facilitate the detection of challenge activation. This could also be one of the reasons that in many physiological-based affective gaming works, several types of high arousal emotions are often targeted as the most performative models. Despite this bias, physiological sensing has been a prominent fixture in games user research (GUR) since the late 1990s [58]. One of the reasons might be that: *"The role affective technologies [...] play in the gaming industry is certain to create some new and exciting user experiences"* [28]. In this regard, physiological-based challenge detection should share the same motivation with affective gaming in using biofeedback techniques to enhance game interaction [50, 58].

Another limitation was related to the three game scenarios selected from a single game, which may have limited the generalisability of the results as different challenge types may manifest themselves in multiple ways. For example, cognitive challenge requires the player to use their memory, attention, reasoning, planning and problem solving. One or more of these abilities may present themselves in different ways in different digital games to provide cognitive challenge. Therefore, we propose that future research should explore more game scenarios in other types of games with different challenge structures. Moreover, the potential relationships between perceived challenge and other eudaimonic and meaningful player experiences could be explored to provide further insights into how to balance challenge types and levels to provide a preferable player experience. In addition, other continuous gameplay data demonstrating the players' in-game performance could be used alongside the physiological metrics to strengthen the evaluation of different types of perceived challenge.

8 CONCLUSION

This study investigated the potential of detecting in-game perceived challenge from physiological signals. In this first study of its kind that evaluates different types of perceived challenge dynamically, objectively and individually, we conducted an experiment to collect physiological signals (EDA, ECG, EMG, RSP and TEM) from 32 players who played through three different game scenarios. Players' perceived challenge was continuously assessed using the CORGIS questionnaire at discrete points in the game. With 80 physiological features extracted, several machine learning methods and metrics were applied to detect four types of perceived challenge of players. It is inspiring that most methods achieved detection accuracy of around 80%, which highlights the potential for further exploration of real-time challenge measurement with physiological signals. Moreover, a set of challenge-related features were also selected and refined with feature importance techniques, which aim to help understand perceived challenge using objective player data and also inform future physiological-based studies of challenge as player experience.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant no. 62132010, 62172397), CAS Project for Young Scientists in Basic Research (Grant No.YSBR-040) and Youth Innovation Promotion Association CAS (Grant no. 2023119,2020113). We thank the reviewers for their input in improving this work.

REFERENCES

[1] Ernest Adams. 2014. *Fundamentals of game design.* Pearson Education.
[2] Gustavo Andrade, Geber Ramalho, Hugo Santana, and Vincent Corruble. 2005. Challenge-sensitive action selection: an application to game balancing. In *IEEE/WIC/ACM International Conference on Intelligent Agent Technology*. IEEE, 194–200. https://doi.org/10.1109/IAT.2005.52
[3] Riccardo Berta, Francesco Bellotti, Alessandro De Gloria, Danu Pranantha, and Carlotta Schatten. 2013. Electroencephalogram and physiological signal analysis for assessing flow in games. *IEEE Transactions on Computational Intelligence and AI in Games* 5, 2 (2013), 164–175. https://doi.org/10.1109/TCIAIG.2013.2260340
[4] Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*. Vol. 4. Springer.
[5] Julia Ayumi Bopp, Klaus Opwis, and Elisa D Mekler. 2018. "An Odd Kind of Pleasure" Differentiating Emotional Challenge in Digital Games. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12. https://doi.org/10.1145/3173574.3173615
[6] Nicholas David Bowman, Joseph Wasserman, and Jaime Banks. 2018. Development of the video game demand scale. In *Video Games*. Routledge, 208–233. https://doi.org/10.4324/9781351235266
[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/a:1010933404324
[8] Guillaume Chanel and Phil Lopes. 2020. User evaluation of affective dynamic difficulty adjustment based on physiological deep learning. In *International Conference on Human-Computer Interaction*. Springer, 3–23. https://doi.org/10.1007/978-3-030-50353-6_1
[9] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. 2008. 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*. 13–17. https://doi.org/10.1145/1457199.1457203
[10] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. 2011. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 41, 6 (2011), 1052–1063. https://doi.org/10.1109/TSMCA.2011.2116000
[11] Bo Cheng and Guangyuan Liu. 2008. Emotion recognition from surface EMG signal using wavelet transform and neural network. In *Proceedings of the 2nd international conference on bioinformatics and biomedical engineering (ICBBE)*. 1363–1366. https://doi.org/10.1109/ICBBE.2008.670
[12] Tom Cole, Paul Cairns, and Marco Gillies. 2015. Emotional and functional challenge in core and avant-garde games. In *Proceedings of the 2015 annual symposium on computer-human interaction in play*. 121–126. https://doi.org/10.1145/2793107.2793147
[13] Tom Cole and Marco Gillies. 2021. Thinking and doing: Challenge, agency, and the eudaimonic experience in video games. *Games and Culture* 16, 2 (2021), 187–207. https://doi.org/10.1177/1555412019881536
[14] Tom Cole and Marco Gillies. 2022. Emotional Exploration and the Eudaimonic Gameplay Experience: A Grounded Theory. In *CHI Conference on Human Factors in Computing Systems*. 1–16. https://doi.org/10.1145/3491102.3502002
[15] Thomas Constant and Guillaume Levieux. 2019. Dynamic difficulty adjustment impact on players' confidence. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12. https://doi.org/10.1145/3290605.3300693
[16] Anna Cox, Paul Cairns, Pari Shah, and Michael Carroll. 2012. Not doing but thinking: the role of challenge in the gaming experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 79–88. https://doi.org/10.1145/2207676.2207689
[17] Adele Cutler, D Richard Cutler, and John R Stevens. 2012. Random forests. In *Ensemble machine learning*. Springer, 157–175. https://doi.org/10.1007/978-1-4419-9326-7_5
[18] Ali Darzi and Domen Novak. 2021. Automated affect classification and task difficulty adaptation in a competitive scenario based on physiological linkage: An exploratory study. *International Journal of Human-Computer Studies* 153 (2021), 102673. https://doi.org/10.1016/j.ijhcs.2021.102673
[19] Alena Denisova, Julia Ayumi Bopp, Thuy Duong Nguyen, and Elisa D Mekler. 2021. "Whatever the Emotional Experience, It's Up to Them": Insights from Designers of Emotionally Impactful Games. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–9. https://doi.org/10.1145/3411764.3445286

[20] Alena Denisova and Paul Cairns. 2015. Adaptation in digital games: the effect of challenge adjustment on player performance and experience. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. 97–101. https://doi.org/10.1145/2793107.2793141

[21] Alena Denisova, Paul Cairns, Christian Guckelsberger, and David Zendle. 2020. Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS). *International Journal of Human-Computer Studies* 137 (2020), 102383. https://doi.org/10.1016/j.ijhcs.2019.102383

[22] Alena Denisova, Christian Guckelsberger, and David Zendle. 2017. Challenge in digital games: towards developing a measurement tool. In *Proceedings of the 2017 chi conference extended abstracts on human factors in computing systems*. 2511–2519. https://doi.org/10.1145/3027063.3053209

[23] Anders Drachen, Lennart E Nacke, Georgios Yannakakis, and Anja Lee Pedersen. 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*. 49–54. https://doi.org/10.1145/1836135.1836143

[24] Alex Flint, Alena Denisova, and Nick Bowman. 2023. Comparing Measures of Perceived Challenge and Demand in Video Games: Exploring the Conceptual Dimensions of CORGIS and VGDS. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3544548.3581409

[25] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232. http://www.jstor.org/stable/2699986

[26] Julian Frommel, Fabian Fischbach, Katja Rogers, and Michael Weber. 2018. Emotion-based dynamic difficulty adjustment using parameterized difficulty and self-reports of emotion. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. 163–171. https://doi.org/10.1145/3242671.3242682

[27] Kiel Gilleade, Alan Dix, and Jen Allanson. 2005. Affective videogames and modes of affective gaming: assist me, challenge me, emote me. *DiGRA 2005: Changing Views–Worlds in Play* (2005). http://www.digra.org/wp-content/uploads/digital-library/06278.55257.pdf

[28] Kiel M Gilleade and Alan Dix. 2004. Using frustration in the design of adaptive videogames. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*. 228–232. https://doi.org/10.1145/1067343.1067372

[29] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 315–323.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778. https://doi.org/10.48550/arXiv.1512.03385

[31] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation* 18, 7 (2006), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

[32] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012). https://doi.org/10.48550/arXiv.1207.0580

[33] Yu-Liang Hsu, Jeen-Shing Wang, Wei-Chun Chiang, and Chien-Han Hung. 2017. Automatic ECG-based emotion recognition in music listening. *IEEE Transactions on Affective Computing* 11, 1 (2017), 85–99. https://doi.org/10.1109/TAFFC.2017.2781732

[34] Gang Hu, Nabil Bin Hannan, Khalid Tearo, Arthur Bastos, and Derek Reilly. 2016. Doing while thinking: Physical and cognitive engagement and immersion in mixed reality games. In *Proceedings of the 2016 ACM conference on designing interactive systems*. 947–958. https://doi.org/10.1145/2901790.2901864

[35] Jin Huang, Xiaolan Peng, Rui Chen, Shengcai Duan, Feng Tian, and Hongan Wang. 2020. Negative Emotion, Positive Performance? A Glimpse into Emotional Influences on Moving Target Selection. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8. https://doi.org/10.1145/3334480.3382874

[36] Jin Huang, Xiaolan Peng, Feng Tian, Hongan Wang, and Guozhong Dai. 2018. Modeling a target-selection motion by leveraging an optimal feedback control mechanism. *Sci. China Inf. Sci.* 61, 4 (2018), 044101–1. https://doi.org/10.1007/s11432-017-9326-8

[37] Jin Huang, Feng Tian, Xiangmin Fan, Xiaolong Zhang, and Shumin Zhai. 2018. Understanding the uncertainty in 1D unidirectional moving target selection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3173574.3173811

[38] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456. https://proceedings.mlr.press/v15/glorot11a.html

[39] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220. https://doi.org/10.1207/s15327108ijap0303_3

[40] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*. https://doi.org/10.48550/arXiv.1412.6980

[41] Madison Klarkowski, Daniel Johnson, Peta Wyeth, Mitchell McEwan, Cody Phillips, and Simon Smith. 2016. Operationalising and evaluating sub-optimal and optimal play experiences through challenge-skill manipulation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5583–5594. https://doi.org/10.1145/2858036.2858563

[42] Sylvia D Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological psychology* 84, 3 (2010), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

[43] Changchun Liu, Pramila Agrawal, Nilanjan Sarkar, and Shuo Chen. 2009. Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *International Journal of Human-Computer Interaction* 25, 6 (2009), 506–529. https://doi.org/10.1080/10447310902963944

[44] Wei-Yin Loh. 2011. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1, 1 (2011), 14–23. https://doi.org/10.1002/widm.8

[45] Marco Maier, Daniel Elsner, Chadly Marouane, Meike Zehnle, and Christoph Fuchs. 2019. DeepFlow: Detecting Optimal User Experience From Physiological Data Using Deep Neural Networks. In *AAMAS*. 2108–2110. https://doi.org/doi:10.24963/ijcai.2019/196

[46] Regan L Mandryk and M Stella Atkins. 2007. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International journal of human-computer studies* 65, 4 (2007), 329–347. https://doi.org/10.1016/j.ijhcs.2006.11.011

[47] Elisa D Mekler and Kasper Hornbæk. 2016. Momentary pleasure or lasting meaning? Distinguishing eudaimonic and hedonic user experiences. In *Proceedings of the 2016 chi conference on human factors in computing systems*. 4509–4520. https://doi.org/10.1145/2858036.2858225

[48] Paraschos Moschovitis and Alena Denisova. 2022. Keep Calm and Aim for the Head: Biofeedback-Controlled Dynamic Difficulty Adjustment in a Horror Game. *IEEE Transactions on Games* (2022). https://doi.org/10.1109/TG.2022.3179842

[49] Lennart Erik Nacke, Michael Kalyn, Calvin Lough, and Regan Lee Mandryk. 2011. Biofeedback game design: using direct and indirect physiological control to enhance game interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 103–112. https://doi.org/10.1145/1978942.1978958

[50] Lennart Erik Nacke, Michael Kalyn, Calvin Lough, and Regan Lee Mandryk. 2011. Biofeedback game design: using direct and indirect physiological control to enhance game interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 103–112. https://doi.org/10.1145/1978942.1978958

[51] Thorbjørn S Nielsen, Gabriella AB Barros, Julian Togelius, and Mark J Nelson. 2015. General video game evaluation using relative algorithm performance profiles. In *European Conference on the Applications of Evolutionary Computation*. Springer, 369–380. https://doi.org/10.1007/978-3-319-16549-3_30

[52] Juha-Pekka Niskanen, Mika P Tarvainen, Perttu O Ranta-Aho, and Pasi A Karjalainen. 2004. Software for advanced HRV analysis. *Computer methods and programs in biomedicine* 76, 1 (2004), 73–81. https://doi.org/10.1016/j.cmpb.2004.03.004

[53] Jiapu Pan and Willis J Tompkins. 1985. A real-time QRS detection algorithm. *IEEE transactions on biomedical engineering* 3 (1985), 230–236. https://doi.org/10.1109/TBME.1985.325532

[54] Xiaolan Peng, Jin Huang, Alena Denisova, Hui Chen, Feng Tian, and Hongan Wang. 2020. A palette of deepened emotions: exploring emotional challenge in virtual reality games. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13. https://doi.org/10.1145/3313831.3376221

[55] Xiaolan Peng, Jin Huang, Linghan Li, Chen Gao, Hui Chen, Feng Tian, and Hongan Wang. 2019. Beyond horror and fear: Exploring player experience invoked by emotional challenge in VR games. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6. https://doi.org/10.1145/3290607.3312832

[56] Xiaolan Peng, Chenyu Meng, Xurong Xie, Jin Huang, Hui Chen, and Hongan Wang. 2022. Detecting challenge from physiological signals: A primary study with a typical game scenario. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7. https://doi.org/10.1145/3491101.3519806

[57] Johannes Pfau, Jan David Smeddinck, and Rainer Malaka. 2020. Enemy within: Long-term motivation effects of deep player behavior models for dynamic difficulty adjustment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10. https://doi.org/10.1145/3313831.3376423

[58] Raquel Robinson, Katelyn Wiley, Amir Rezaeivahdati, Madison Klarkowski, and Regan L Mandryk. 2020. "Let's Get Physiological, Physiological!" A Systematic Review of Affective Gaming. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 132–147. https://doi.org/10.1145/3410404.3414227

[59] Jesse Schell. 2008. *The Art of Game Design: A book of lenses*. CRC press. https://doi.org/10.1201/9780080919171

[60] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality*

and social psychology 52, 6 (1987), 1061. https://doi.org/10.1037/0022-3514.52.6.1061

[61] Jainendra Shukla, Miguel Barreda-Angeles, Joan Oliver, GC Nandi, and Domenec Puig. 2019. Feature extraction and selection for emotion recognition from electrodermal activity. *IEEE Transactions on Affective Computing* (2019). https://doi.org/10.1109/TAFFC.2019.2901673

[62] Pieter Spronck, Marc Ponsen, Ida Sprinkhuizen-Kuyper, and Eric Postma. 2006. Adaptive game AI with dynamic scripting. *Machine Learning* 63, 3 (2006), 217–248. https://doi.org/10.1007/s10994-006-6205-6

[63] Bethesda Game Studios. 2015. Fallout 4. Game [PlayStation 4], [Windows], [Xbox One]. Bethesda Softworks, Rockville, Maryland.

[64] Wenlu Yang, Maria Rifqi, Christophe Marsala, and Andrea Pinna. 2018. Physiological-based emotion detection and recognition in a video game context. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8. https://doi.org/10.1109/IJCNN.2018.8489125

[65] Georgios N Yannakakis and John Hallam. 2009. Real-time game adaptation for optimizing player satisfaction. *IEEE Transactions on Computational Intelligence and AI in Games* 1, 2 (2009), 121–133. https://doi.org/10.1109/TCIAIG.2009.2024533

A GAME EVENTS

A.1 Game Events of Scenario A

The description of events and instruction tips of scenario A are given in Table 7.

A.2 Game Events of Scenario B

The description of events and instruction tips of scenario B are given in Table 8.

A.3 Game Events of Scenario C

The description of events and instruction tips of scenario C are given in Table 9.

B CHALLENGE RATINGS

B.1 Challenge Ratings of Scenario A

The challenge ratings of each event in scenario A are given in Table 10.

B.2 Challenge Ratings of Scenario B

The challenge ratings of each event in scenario B are given in Table 11.

B.3 Challenge Ratings of Scenario C

The challenge ratings of each event in scenario c are given in Table 12.

C PHYSIOLOGICAL FEATURES

The 80 extracted physiological features are given in Table 13.

D MACHINE LEARNING METHODS

The model descriptions and configurations are given in Table 14-16.

Table 7: Game Events of Scenario A

Quest	Event	Description of Events and Instruction Tips
	Tips	Read a brief instruction to learn the background of the main story of Fallout 4.
War Never Changes	1	Walk around inside the house and interact with the house's objects and goods, such as bottles, boxes, their spouse and infant son Shaun, and the robot housekeeper Codsworth.
	2	Talk to Vault-Tec rep, a person who states that they are qualified to enter the nearby Vault 111.
	3	Interact with Shaun and chat with the spouse for a while.
	4	Learn from the television that nuclear detonations have abruptly hit some nearby cities.
	5	Rush toward Vault 111 with their spouse and son, and enter the Vault just moments before a nuclear detonation.
	6	Follow the vault's scientist and enter a cryo pod, with the spouse and Shaun entering another.
	7	Witness three unknown figures open the spouse's pod and a mysterious man shoot the spouse with a pistol and take their son.
	8	Fall out of the cryo pod, struggle to feet and then open the spouse's cryo pod and vow to find Shaun.
Out of Time	9	Progress through the hallways of the hault and open several doors.
	10	Find the way to escape the vault and kill several radroaches.
	11	Take the elevator to get out of the vault and reach the ground level.
	12	Retrace the path back to home and find ruins all over the world.
	13	Meet the robot Codsworth and talk to learn that 210 years have passed since nuclear detonation.
	Tips	Read an instruction to learn that there is a big evil-force called Institution and they aim to create synths to control the destroyed world. Learn that it is Kellogg who kills the spouse and the player is about to find Kellogg.
Reunions	14	Fight against the synths guarding for Kellogg's building.
	15	Fight against the synths and automated turrets guarding for Kellogg's building.
	16	Fight against the synths and energy weapons guarding for Kellogg's building.
	17	Talk to Kellogg and be told that Shaun is being held by the Institute but out of reach.
	18	Fight against Kellogg and his synth bodyguards.
Institution -alized	19	Follow a greeting voice of a hidden speaker named "Father" and take an elevator to enter a special room of Institute.
	20	Find a scared and confused child in the room and then talk to the child. The child says that he is Shaun but keeps asking "Father" to help.
	21	See an old man named "Father" enters the room and know that the child is just a synthetic Shaun. "Father" says he is the real Shaun and also the director of the Institute now.
	22	Talk to "Father" and listen to him to talk about the details of his kidnapping by the Institute.
	23-a	Be persuaded to join the Institute and finally agrees to join. (5 participants chose this)
	23-b	Be persuaded to join the Institute but finally disagrees to join. (5 participants chose this)

Table 8: Game Events of Scenario B

Quest	Event	Description of Events and Instruction Tips
	Tips	Read a brief instruction to learn the player's life experience of Vault 111. Learn that there is a big evil-force called Institution and they aim to create synths to control the destroyed world. Learn that there is another force of Brotherhood of Steel whose aim is to eliminate synths.
Fire Support	1	Assist several soldiers to fight against a group of feral ghouls.
	2	Talk to the soldiers' commander, Paladin Danse, know that Danse work for Brotherhood of Steel and agrees toassist them to get a transmitter.
Call to Arms	3	Follow Danse to walk.
	4	Follow Danse to walk, listen Danse to speak and may also fight against some wildlife on the way.
	5	Talk to Danse to learn the crime of synths in an abandoned rocket silo.
	6	Find the lab control terminal to unlock a door in the way.
	7	Crack the password on the terminal to unlock the door.
	8	Fight alongside with Danse against synths from the unlocked room.
	9	Fight alongside with Danse against the fire power of synths' weapon systems.
	10	Find the equipment to restore the power supply to the silo.
	11	Crack the password on the terminal to restore the power to the silo.
	12	Fight alongside with Danse against synths from the bottom of the silo.
	13	Fight alongside with Danse against synths to get the transmitter.
Blind Betrayal	14	Talk to Elder Maxson, be told Danse is a synth and be ordered to execute Danse.
	15	Talk to Proctor Quinlan and be confirmed that Danse is a synth.
	16	Talk to Scribe Haylen. Haylen says that Danse is a good friend. Be told where Danse is hiding.
	17	Fight against three robots before meeting Danse.
	18	Talk to Danse until making the important decisions to spare or execute Danse.
	19-a1	Decide to spare Danse and convince Danse to escape. (9 participants chose this)
	19-a2	Talk to Maxson and listen Danse to describe his loyalty to the Brotherhood.
	19-a3-c	Decide to spare Danse and persuade Maxson to spare Danse. (7 participants chose this)
	19-a3-d	Decide to request Maxson to execute Danse. (2 participants chose this)
	19-b	Decide to and execute Danse. (2 participants chose this)

Table 9: Game Events of Scenario C

Quest	Event	Description of Events and Instruction Tips
	Tips	Read a brief instruction about the player's life experience of Vault 111. Learn that there is a big evil-force called Institution and they aim to create synths to control the destroyed world. Learn that human residents are very scared of synths as it is very hard to distinguish synths from real human beings. Learn that a caravan person have been strangely killed near a peaceful settlement called Covenant. Get the task to enter the Covenant to investigate what has happened to the caravan.
Human Error	1	Talk to the guard person of Covenant and know that a SAFE test must be passed to enter Covenant.
	2	Take the SAFE test by answering a serious of psychological questions.
	3	Talk to Honest Dan who is also an outsider of Covenant. Agree to help Dan to investigate Covenant to find Amelia Stockton, the missing girl of the caravan.
	4	Search information in Covenant by talking to the residents and searching their houses, etc.
	5	Read pieces of information on a public terminal or/and talk to the robot.
	6	Search information in Covenant by talking to the residents and searching their houses, etc.
	7	Crack the password on the terminal to unlock a private office terminal.
	8	Read important information on the office terminal. Know that the girl, Amelia Stockton, is hided in a secret place called Compound.
	9	Talk to Dan and share the searched information with him.
	10	Fight along with Dan against the armed guards at the Compound.
	11	Search information at the Compound and may also fight against some armed guards on the way.
	12	Fight along with Dan against the armed guards at the Compound.
	13	Find way to enter more inside and search more information.
	14	Fight against some armed guards on the way and know that they are fighting to protect a doctor.
	15	Read some important information. Know that the Compound are doing some cruel experiments to develop the SAFE questionnaire.
	16	Talk to the doctor who explains her and others' hatred of synths.
	17	Learn that the doctor tries to reveal hidden synths who infiltrated the populace by perfecting the SAFE test, even by killing real human beings. Learn that the caravan people were killed because the Covenant thinks Amelia in the caravan is most likely a synth infiltrator.
	18	Be asked whether to support the doctor to continue the experiment. Dan joins the talk and says he does not agree to continue the experiment.
	19-a	Disagree with the doctor and release Amelia. (9 participants chose this)
	19-b1	Agree with the doctor and kill Dan (2 participants chose this).
	19-b2	Look around and operate the doctor's terminal.

Table 10: Challenge Ratings of Scenario A. Abbreviations: cogC = cognitive challenge, emoC = emotional challenge, phyC = physical challenge, DM = decision making

	Time Consuming	Ratings' Mean				Ratings' Standard deviation			
Event	(seconds)	cogC	emoC	phyC	DM	cogC	emoC	phyC	DM
1	177.30 (69.57)	1.85	2.42	1.40	1.64	0.95	0.75	0.63	0.91
2	135.30 (19.04)	2.84	2.72	2.16	2.20	1.03	1.03	1.21	1.24
3	88.70 (10.66)	1.94	2.78	1.46	1.82	0.90	0.78	0.67	0.76
4	36.60 (4.52)	3.25	3.49	3.58	1.92	1.36	1.24	2.00	1.03
5	140.00 (18.63)	2.95	3.33	3.50	1.94	1.23	1.10	1.94	1.03
6	212.80 (48.02)	2.75	2.63	1.80	1.74	1.30	1.05	0.83	1.06
7	71.00 (1.26)	2.78	4.20	1.90	1.92	1.07	0.82	0.92	1.27
8	64.20 (21.55)	2.85	3.61	2.12	2.18	0.79	1.13	1.44	1.57
9	105.40 (74.81)	3.14	2.59	2.44	1.54	1.11	0.94	1.81	0.59
10	351.00 (68.50)	3.39	2.31	3.80	1.56	0.88	0.92	1.65	0.64
11	93.80 (12.07)	2.13	2.44	1.48	1.42	0.96	1.19	0.48	0.63
12	116.60 (49.68)	2.78	2.63	2.06	1.42	1.13	1.00	1.11	0.52
13	213.30 (73.97)	2.81	3.02	1.74	2.14	1.27	0.85	1.01	0.34
14	41.70 (14.70)	3.45	2.18	5.72	1.28	1.16	1.11	0.87	0.43
15	40.60 (14.24)	3.68	2.26	5.74	1.48	1.61	1.10	1.30	0.63
16	39.80 (22.80)	3.41	2.23	5.60	1.34	1.31	1.05	1.13	0.41
17	85.10 (27.12)	3.28	2.67	2.84	2.42	1.26	1.09	1.76	1.35
18	81.30 (19.91)	4.12	2.41	5.96	1.94	1.12	1.28	1.28	1.18
19	91.67 (9.42)	3.06	2.67	1.64	1.42	0.91	1.14	0.71	0.61
20	60.60 (5.68)	3.06	3.86	1.92	2.60	0.83	0.77	0.96	1.36
21	138.00 (29.97)	3.01	3.32	1.72	2.48	0.77	1.01	0.93	1.54
22	279.67 (96.81)	2.57	3.31	1.44	2.73	1.06	0.87	0.66	1.76
23-a	105.00 (40.37)	2.96	4.00	1.64	3.44	0.81	0.50	0.79	1.59
23-b	102.00 (25.81)	2.87	3.96	1.36	4.28	1.19	1.39	0.54	1.06

Table 11: Challenge Ratings of Scenario B. Abbreviations: cogC = cognitive challenge, emoC = emotional challenge, phyC = physical challenge, DM = decision making

	Time Consuming	Ratings' Mean				Ratings' Standard deviation			
Event	(seconds)	cogC	emoC	phyC	DM	cogC	emoC	phyC	DM
1	56.91 (11.84)	3.47	2.41	4.94	2.22	1.59	0.82	1.47	0.83
2	121.18 (22.01)	3.79	2.84	2.93	2.67	1.04	0.83	1.32	1.14
3	64.91 (5.96)	3.38	2.67	2.84	2.04	0.95	0.81	0.86	0.93
4	162.36 (51.64)	3.50	2.80	4.09	2.15	0.96	0.72	1.09	0.72
5	46.00 (11.58)	3.27	2.79	2.49	2.58	0.99	0.82	1.01	0.69
6	62.00 (29.77)	4.09	2.65	3.47	2.22	1.28	0.84	1.16	0.81
7	116.36 (53.00)	4.97	2.78	3.51	3.09	1.35	0.87	1.34	1.19
8	50.64 (14.70)	3.92	2.57	4.95	1.95	1.07	0.70	0.59	0.80
9	29.64 (6.06)	3.65	2.73	4.49	2.05	1.06	0.78	0.82	0.82
10	52.36 (14.32)	4.16	2.67	3.67	2.87	1.15	0.77	1.07	1.15
11	56.00 (40.50)	4.43	2.81	3.55	2.60	0.84	0.74	1.26	1.11
12	48.09 (11.41)	3.90	2.56	4.87	2.44	1.14	0.67	0.96	0.82
13	24.09 (3.82)	3.83	2.37	4.80	2.13	1.01	0.70	0.66	0.80
14	128.55 (12.69)	3.87	4.00	3.04	3.40	0.85	0.86	1.04	1.18
15	159.91 (27.61)	3.90	4.09	2.87	3.87	0.97	1.02	1.03	0.99
16	105.09 (23.63)	3.45	3.90	2.56	3.44	0.99	1.00	0.80	0.99
17	26.45 (6.93)	3.79	2.80	4.56	2.25	1.02	0.78	1.13	0.74
18	137.82 (42.76)	3.80	4.37	2.71	3.95	0.98	0.89	0.73	0.79
19-a1	73.33 (10.80)	3.78	4.11	2.91	3.84	0.84	0.95	1.22	1.27
19-a2	61.00 (45.91)	3.75	4.42	2.71	3.36	1.05	0.52	0.59	1.12
19-a3-c	155.44 (9.95)	3.70	4.06	2.54	3.69	1.01	1.14	0.52	1.30
19-a3-d	110.29 (7.50)	4.86	5.28	3.05	6.30	0.23	0.72	0.21	0.70
19-b	45.50 (9.00)	2.55	4.72	1.80	4.70	1.00	1.28	0.20	0.70

Table 12: Challenge Ratings of Scenario C. Abbreviations: cogC = cognitive challenge, emoC = emotional challenge, phyC = physical challenge, DM = decision making

	Time Consuming	Ratings' Mean				Ratings' Standard deviation			
Event	(seconds)	cogC	emoC	phyC	DM	cogC	emoC	phyC	DM
1	48.27 (17.24)	3.73	3.05	3.38	3.55	1.45	1.15	1.66	1.48
2	158.27 (12.13)	4.26	4.58	3.75	4.64	1.58	1.16	1.73	1.27
3	77.36 (16.53)	4.04	3.59	3.55	4.09	1.69	1.40	1.57	1.80
4	171.45 (141.46)	4.22	3.02	3.65	3.15	1.32	1.19	1.72	1.47
5	31.91 (11.56)	4.15	3.03	3.27	3.22	0.92	0.90	1.55	1.54
6	220.45 (120.80)	3.69	2.96	3.51	2.80	1.64	1.21	1.46	1.31
7	76.18 (59.81)	5.01	3.11	3.96	3.42	1.26	1.04	1.69	1.40
8	99.18 (66.21)	4.42	3.09	3.33	2.78	1.57	0.96	1.71	1.60
9	34.00 (19.70)	3.68	3.10	3.22	2.96	1.75	1.48	1.64	1.46
10	30.36 (20.48)	4.22	2.80	5.93	2.84	1.17	1.10	0.73	1.61
11	75.27 (49.79)	4.47	2.99	5.71	2.33	0.98	1.30	1.01	1.25
12	43.64 (29.39)	4.67	2.79	5.69	2.69	1.23	1.02	0.65	1.45
13	86.91 (67.12)	4.61	3.55	4.09	2.76	1.29	1.54	1.81	1.70
14	86.82 (65.08)	4.33	3.04	4.75	2.58	1.17	1.16	1.98	1.65
15	62.91 (50.03)	4.43	3.48	4.02	2.53	1.79	1.40	1.97	1.40
16	92.64 (36.01)	3.98	4.34	3.24	3.75	1.66	1.53	1.59	1.63
17	54.55 (23.78)	4.18	5.07	3.95	4.29	1.57	1.30	1.80	1.95
18	53.36 (18.35)	4.36	5.18	4.22	4.89	1.71	1.20	1.92	1.56
19-a	47.25 (7.81)	4.03	4.44	3.68	3.60	1.42	0.52	1.56	1.61
19-b1	22.00 (1.00)	4.09	4.67	6.30	4.20	1.36	1.78	0.30	2.80
19-b2	96.00 (15.00)	3.77	2.61	2.70	1.30	0.05	0.06	1.70	0.30

Table 13: The 80 extracted physiological features. Abbreviations: Std = standard deviation, MAD = mean absolute deviation, MAV = mean absolute value, TP = total power or PSD of {VLF,LF,HF}, LF or HF norm = $\frac{\text{LF or HF}}{\text{total power or PSD of \{LF,HF\}}}$, Band power% = $\frac{\text{Band power}}{\text{Total power}}$, Entropy = - $\sum$ Band power% $\log$ Band power%, RMS = root mean square, SAV = sum absolute value.

ECG	Time	HR	BPM (Bit Per Minute)
		R	Mean, Std, MAD
		Diff R	Mean, Std, MAD, MAV
		RR (HRV-)	Mean, Std, MAD
		Diff RR	Mean, Std, MAD, MAV
	Frequency (HRV-)	RR power	VLF, LF, HF, TP, LF/HF, LF norm, HF norm
		RR PSD	VLF, LF, HF, TP, LF/HF, LF norm, HF norm
TEM	Amplitude		Mean, Std, MAD
	Diff Amplitude		Mean, Std, MAD, MAV
RSP	Time	Amplitude	Mean, Std, MAD
		Diff Amplitude	Mean, Std, MAD, MAV
		Period	Mean, Std, MAD, BPM (Breath Per Minute)
		Diff Period	Mean, Std, MAD, MAV
EMG	Wavelet decomposition		1-7 Band power, Total power, 1-7 Band power%, Entropy, RMS
EDA	Time		LF SAV, HF SAV, VHF SAV
	Frequency		LF SAV, LF power, LF power Std, HF SAV, HF power, HF power Std, VHF SAV, VHF power, VHF power Std

Table 14: Machine learning models and configurations

Model	Description	Configuration
Linear Regression1 (LR1) (linear+squared)	Linear Regression model uses a linear function of input features $\mathbf{x}$ to predict the label $\mathbf{y} \in \mathbb{R}^4$, which is the challenge ratings in this paper. This is written as $\hat{\mathbf{y}} = \mathbf{w}^T\mathbf{x}$. To estimate the parameters $\mathbf{w}$, mean square error (MSE) between the predicted and true labels of the training data is minimized.	Both of the linear and squared terms of the physiological features are used and can be written as $\mathbf{x} = [\mathbf{x}^2, \mathbf{x}, 1]^T$. We train one Linear Regression model for each challenge.
Linear Regression2 (LR2) (Gaussian kernel)	This model is similar to the Linear Regression1 but Gaussian kernel [4] is employed for the input features. In this case, the prediction function of Linear Regression becomes $\hat{\mathbf{y}}_{\text{test}} = \mathbf{y}_{\text{train}}(\mathbf{K}(\mathbf{x}_{\text{train}}, \mathbf{x}_{\text{train}}) + \lambda\mathbf{I})^{-1}\mathbf{K}(\mathbf{x}_{\text{train}}, \mathbf{x}_{\text{test}})$, where $\mathbf{K}(\mathbf{x1}, \mathbf{x2})$ is the Gaussian kernel function and the (i, j)th element of it is computed as $\exp(-\gamma\|\mathbf{x1}_i - \mathbf{x2}_j\|^2)$.	We simply set $\lambda = 0.01$ and $\gamma = \frac{1}{D*\sigma_x}$ where D is the number of input features and σ_x denotes the global standard deviation of training data over all features.
K-Nearest Neighbors (KNN)	KNN model simply predicts the challenge ratings for each data sample in test set by using the average challenge ratings over its k-nearest data samples in training set. We normalize all data by z-score using mean and variance of training data and compute the euclidean distance to find the k-nearest neighbors.	We set $k = 3$ for 10-folds cross validation. For leave-one-participant-out cross validation we set $k = 10$, as a larger k may reduce overfitting.
Support Vector Machine (SVM)	SVM regression model predict the challenge ratings by $\hat{\mathbf{y}} = \mathbf{w}^T\mathbf{x}$ and finds the parameters $\mathbf{w}$ by $\arg\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2}$ such that $\lvert\mathbf{y}_{\text{train}} - \mathbf{w}^T\mathbf{x}_{\text{train}}\rvert \leq \epsilon$.	We use the Gaussian kernel function similar to that used by Linear Regression2 for input features. We train one SVM model for each challenge.

Table 15: Machine learning models and configurations (continued from Table 14)

Model	Description	Configuration
Decision Tree (DT)	The Decision Tree model used in this paper is based on the Classification and Regression Trees [44]. It is a binary tree where each node selects the pair of feature and splitting threshold that generate the minimum MSE.	The splitting stops when the MSE drops below 10^{-6} or it results in leaf nodes having less data samples than 20. We train one Decision Tree model for each challenge.
Gradient Boosting Decision Trees (GBDT)	GBDT model [25] is an ensemble of Decision Trees, of which each tree is trained to minimize the MSE by predicting the negative shrunken gradient derived from the ensemble of the previously trained trees.	For each tree, the maximum splitting number is 10 and the minimum data samples of a leaf node is 10. The learning rate for gradient shrinkage is 0.5 and 100 trees are trained. We train one GBDT model for each challenge.
Random Forest (RF)	Random Forest [7] is an ensemble of Decision Trees where each tree is trained with random selected features and random selected training data.	For each tree, we set the feature selection rate to $\frac{1}{3}$, the training data selection rate to 0.75, and the minimum data samples of a leaf node to 3. We train one Random Forest model for each challenge.
Neural Network (NN) (1-hidden)	NN model commonly consists of a stack of hidden layers followed by an output layer. Each hidden layer consists of nonlinear operation following an affine transform. The output layer consists of an affine transform to predict the 4 challenge ratings.	This is a conventional shallow DNN setting with one 200-dimensional hidden layer using a Sigmoid nonlinear operation. The network are trained to minimize the MSE using Adam [40] with initial learning rate 0.01, which will be decayed by the factor of 0.5 for annealing. In order to alleviate over-fitting, we set the coefficient of L2-penalty to 0.0001 and use 10% of the training data as development set for tuning the learning rate. We set the minibatch size to 80 for 10-folds cross validation. For leave-one-participant-out cross validation a larger minibatch size seems to reduce overfitting, thus we set this to 80 sequences of successive samples with maximum length of 5.
Deep Neural Network1 (DNN1) (5-hidden)	The DNN refers to Neural Network with multiple hidden layers [31].	In this model, we use five hidden layers where each layer consists of a ReLU function [29] followed by batch-normalization [38] and dropout [32] operations with 90% retention as the nonlinear operation. The output layer consists of an affine transform to predict the 4 challenge ratings, following a stack of 64-dimensional Affine, ReLU and batch-normalization operations. Other settings are the same as the NN model.

Table 16: Machine learning models and configurations (continued from Table 15)

Model	Description	Configuration
DNN2 (5-hidden+residual)	Residual connection [30] in DNN is an additional connection between two layers with a relative short path.	Residual connection with one hidden layer connecting the output of the first and fourth hidden layers is employed in this model. Other settings are the same as the DNN1 model.
DNN3 (9-hidden+residual)	-	This model is configured similar to the DNN2 model except for using nine hidden layers and an additional residual connection connecting the output of the fifth and eighth layers.
DNN4 (5-hidden+residual+MTL)	Apart from the MSE, we use multi-task learning (MTL) to minimize two additional objects.	We add four output layers with the same structure as the first one, but have 7-dimensional final-affine-transform followed by Softmax function. Furthermore, we add one similar output layer with 5-dimensional final-affine-transform and Sigmoid function, to "predict" the four rating level classifications and five challenge activation classifications respectively during training. The training labels of these additional task are four one-hot vectors for the 7-class tasks and five 0/1 values for the 2-class tasks for each data sample. The 2-class tasks are described in sections 5.3.2. The labels of 7-class tasks are generated by dividing each challenge rating into seven discrete levels by rounding the rating to the nearest integer $l \in \{1, ..., 7\}$. For simplification, we minimize the MSE between true and predicted labels for these additional tasks. In the evaluation stage, only the first (original) output layer producing challenge ratings are used to compute the evaluation metrics introduced in section 5.3.
DNN5 (9-hidden+residual+MTL)	-	This model is configured similar to the DNN4 model except for using nine hidden layers and an additional residual connection connecting the output of the fifth and eighth layers.