Predicting Video Game Players' Fun from Physiological and Behavioural Data.

One Algorithm Does not Fit All.

Alexis Fortin-Côté,
Cindy Chamberland, Mark Parent,
Sébastien Tremblay
and Philip Jackson
School of Psychology,
Université Laval, Québec, Canada
Email: alexis.fortin-cote.1@ulaval.ca

Nicolas Beaudoin-Gagnon,
and Alexandre Campeau-Lecours
Department of Mechanical Engineering,
Université Laval, Québec, Canada

Jérémy Bergeron-Boucher
and Ludovic Lefebvre
Ubisoft Québec
Quebec City, Quebec, Canada

***Abstract*—Finding a physiological signature of a player's fun is a goal yet to be achieved in the field of adaptive gaming. The research presented in this paper tackles this issue by gathering physiological, behavioural and self-report data from over 200 participants who played off-the-shelf video games from the Assassin's Creed series within a minimally invasive laboratory environment. By leveraging machine learning techniques the prediction of the player's fun from its physiological and behavioural markers becomes a possibility. They provide clues as to which signals are the most relevant in establishing a physiological signature of the fun factor by providing an importance score based on the predictive power of each signal. Identifying those markers and their impact will prove crucial in the development of adaptive video games. Adataptive games that tailor their gameplay to the affective state of a player in order to deliver the optimal gaming experience. Indeed, an adaptive video game needs a continuous reading of the fun level to be able to respond to these changing fun levels in real time. While the predictive power of the presented classifier remains limited with a gain in the F1 score of 15% against random chance, it brings insight as to which physiological features might be the most informative for further analyses and discuss means by which low accuracy classification could still improve gaming experience.**

***Keywords*—Affective computing; Machine learning; Biomedical measurement, Video Game**

I. Introduction

In recent years, studies have increasingly linked video games to their potential social, cognitive and motivational benefits [1]. This led to an important growth in the interest of serious gaming and learning environments that use games for other reasons than pure entertainment [2]. Fun within games has proven to be a positive factor in learning and behavioural changes [3] and is also documented as an important factor in the satisfaction of players in entertainment-driven games, a 30-billion-dollar market [4]. It thus seems to be an important target to take into account in the development of any games, regardless of whether they were designed for educational or entertainment purposes. The optimization of the player's fun during gameplay would ensure that the experience is positive. To this end, a continuous assessment of the fun throughout the gameplay would allow a real-time adaptation that are tailored to the player preferences.

As with many subjective experiences, there are different definitions of the concept of fun, for it could be used as a label for different states across individuals. For instance, some people experience fun through the relief of fear by reaching a safe point in a horror game [5] or by overcoming a level after repeated failures [6]. While context and specific triggers of fun may vary, the necessary condition for something to be fun can be described as evoking a state of positive valence to a person [7]. Yet, fun remains a challenge to capture, contrarily to the assessment of difficulty and skill levels which are accurately measured through in-game behaviour [8]. Like any human affective or cognitive state, fun is continuous and unfolds over time and over multiple gaming sessions [9]. However, it is often reduced to a holistic rating in the study of user experience due to practical reasons [7], [10]. Yet this approach makes it near impossible to pinpoint the specific factors which contribute to the player's fun. Most importantly, such measures of the player's fun prevent any real-time adaptation of gameplay as only general appreciation can be assessed.

Traditionally, game designers have tried to identify player preference profiles [11], [12] and created content to respond to some or all player profiles as a way to ensure a sense of fun within the game. However, this led to either large population not being targeted by the game, or game content that did not match certain player profiles. Furthermore, this approach could only be used during the game design as players profiling requires lengthy psychometric assessment, which is difficult to implement at a larger scale.

To overcome these limits, game developers have started modelling player experiences in real time from behavioural cues [13], [14]. For instance, dynamic difficulty adjustment algorithms have been developed to assess subjective difficulty from gameplay and adapt the game to maintain an optimal flow level [8]. However, these approaches rely on the assumption that every player experience fun in the same context and that all players react to difficulty the same way.

Although human-computer interaction initially operated primarily on the basis of behavioural markers, research in affective computing now provides access to other dimensions of a player's state through the use of psychophysiological measures [15]. Changes in physiological responses have long been recognized as potential markers of affective and cognitive states but has gained popularity in the last decades through advances in recordings, interpretation and analysis [16]. As opposed to subjective ratings, psychophysiological measures are mostly independent from bias and can be measured continuously without breaking flow [17]. These markers have, therefore, the potential to add a number of important features to help infer the fun of a player. Nonetheless individual physiological markers are non-specific measures of the player's experience [18]. Thus, by collecting both behavioural and physiological markers, the strength of each can be leveraged in the continuous inference of the fun.

Over the last decade, an increasing number of studies have shown associations between physiological markers and a wide range of cognitive and affective states such as workload [19], attention [20], and various discrete emotions [21], [22]. In gaming research, the focus has mostly been on direct biofeedback as a way to improve the player's experience [23], [24]. However, few studies have attempted to predict either affective or cognitive states within video games and even less using a commercially available game. Although the approach presented here focuses on fun, it is general enough to adapt to other cognitive and affective states, such as stress and engagement, given a different measure.

This study thus aimed to find a specific physiological signature of the player's fun during video game session. To this end, a multitude of data sources were exploited: 1) The signals coming from several physiological measures; 2) The responses to a wide range of questionnaires aiming to capture individual differences; 3) The game events, which provide information on the player's actions in the game and the state of play; 4) A proxy of the subjective state of the player during the game obtained through a replay of the session during which the participant rated the fun experienced on a continuous scale. Finding a signature from those data sources would open the door to new possibilities in affective gaming by adapting content to the player's experience and reinforcing motivation for the game. This research is part of the FUN*ii* project introduced in previous papers [25], where participants played off-the-shelf popular games from Ubisoft's Assassin's Creed series. Preliminary results were reported in [26], in which only a subset of the modalities (limited number of physiological measures) from a small set of 63 participants were leveraged. The current paper proposes an approach that takes into account the full dataset using all of the modalities for a larger sample of 218 participants.

The paper is structured as follows: the methodology and the materials used in the creation of the dataset are presented in sec. II. Analysis, pre-processing and labelling of the dataset is presented in sec. III. The modelling and fun prediction from the dataset are then presented in sec. IV. Finally, a discussion and conclusion including future works are presented.

II. Methods and Materials

A. Participants

Two hundred twenty eight participants aged between 18 and 35 years old ($M = 25.49$, $SD = 4.54$; 212 male, 16 females), were recruited from Université Laval and Ubisoft Québec's volunteer database. This last database and the type of game proposed to the participants are the main reason for the heavy gender imbalance of the dataset. None of the participants reported any mental health diagnosis, cognitive impairment, uncorrected vision, or health issue that could impact the physiological and cognitive measures gathered during the experiment. Furthermore, participants were required not to have played the specific games used in the experiment. This project was approved by Université Laval's Ethics Committee (#2012-272).A monetary compensation of $20 was given to participants.

B. Materials

The sample was split based on two different games from the Assassin's Creed series: 103 participants played the missions "The Prophet" (S5M3) and "The Escape" (S9M3) from Assassin's Creed Unity (ACU) and 115 played the missions "A Spoonful of Syrup" (S4M1) and "Survival of the Fittest" (S5M3) on Assassin's Creed Syndicate (ACS). These missions were selected based on their relatively short completion time and differences in subjective difficulty and fun during pilot studies. Both games were the latest opus of the series at the time of experimentation to ensure that we could reach enough players that had never played the game. The games were played on a high-end PC using an Xbox 360 controller for Windows. Finally, game sessions were recorded using a dedicated video monitoring card (Blackmagic WDM).

Four distinct sources of data were leveraged in this study: 1) **Physiology**: A set of physiological measures were recorded during game sessions using a Biopac MP150 system. Cardiac activity was monitored with an electrocardiogram (ECG) using a lead II configuration. Respiratory activity was monitored using a respiration (RSP) belt transducer placed around the player's chest. Electro dermal activity (EDA) was monitored on the left thenar and hypothenar eminences. Muscle activity of the right abductor pollicis longus (APL) was monitored using electromyography (EMG) for a small subset of participants. Furthermore, eye movements and pupil size were recorded using the Smart Eye Pro eye-tracking system. Other measures included blinks, fixations, saccades, and head and gaze orientation. Head and gaze orientation might capture larger scale movement like head shake or shrug. Finally, twenty facial action units, for which intensities are rated on a scale from 0 to 5, were extracted from a video recording of the participant during gameplay using Noldus FaceReader 5.0. Data from all different sources were synchronized using Noldus Observer XT 11 and in-house routines in Matlab 2015b.

2) **Questionnaires**: Self-reports were included in the study to capture individual differences, such as an immersion questionnaire [27]. Participants also reported the subjective difficulty and fun on a 1 to 5 scale, 1 being the lowest intensity and 5 the highest, for each mission played, and completed the short version of the NASA Task Load Index [28].

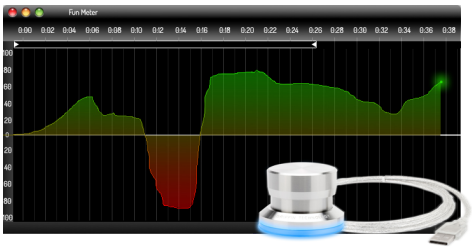

Fig. 1. Interface of the "Funmeter" with its associated control knob.

3) **Game events**: A set of game events were also recorded during the gameplay. The player's activities were recorded at regular intervals, each second for ACU and each two seconds for ACS. Activities range from simple activities like walking, sprinting and leaping to more complex activities like "In conflict," meaning that the player is involved in a fight or "Investigated," meaning that enemies are searching for the participant's character. Game events during gameplay were tracked online. Controller inputs were recorded using the Xinput v1.2.1 API.

4) **Fun**: A custom software called "Funmeter", of which the interface can be seen in Fig. 1, was developed as part of the project to measure the player's level of fun continuously. This software allowed players to rate the level of fun after the mission by watching a replay of their previous game session and scoring the level of fun on a linear scale from -100 to 100. Ratings were controlled by the player through a knob (PowerMate USB, Griffin technology) with visual feedback and sampled at 30 Hz. The interpretation of the meaning of the fun level was left to the participant.

C. Procedure

Upon arrival for a two-hour-long session, participants were given a brief overview of the project. Electrodes and physiological sensors were then installed on the participants. Once signal quality was confirmed, baseline activity for ECG, RSP, EDA and EMG signals were recorded for 3 minutes during which participants were asked to clear their mind while fixing a cross on a white screen and while a white noise was played in headphones as to create a baseline for those signals. The eye-tracking system was then calibrated for the participant. Participants were asked to read through a tutorial about the game to learn the game controls and mechanics. They then had a 5-minute trial session during which they had to achieve specific objectives to ensure that they knew everything necessary to complete the experiment. Participants then played the selected missions in counterbalanced order. They had an undisclosed maximum of approximately 15 minutes to complete each mission. Following each mission, players watched a replay of their last game session and rated their fun level continuously using the Funmeter software. They then reported their subjective experience regarding difficulty, fun and completed the NASA-TLX and the immersion questionnaires. Electrodes and sensors were removed and participants were debriefed.

D. Data Processing

From all the 218 participants, 25 were discarded because of technical issues. For all 193 remaining participants, of which 9 were female, the two missions played were kept for a total of 384 game sessions. General statistics of these game sessions broken-down by missions are shown in the Table I.

From the dataset available, a total of 16 different modalities were extracted, which are presented in the Table II. Sub-signals were extracted from the main signals like the heart rate and the respiration rate, which are derivatives of the electrocardiogram and the respiration intensity. ECG, RSP, EDA and EMG signals were normalized using their baseline value acquired during the first 3 minutes of the experience where players were at rest.

III. Dataset Analysis and Preprocessing

A. Game event analysis

The players' preferred activities were determined by examining their fun ratings in relation to game events. Fig. 2 shows the distribution of participants' fun ratings as a violin plot and a sample of the underlying points. The ratings were aggregated based on different activities and averaged by participants (e.g. a player who has rated the activity "Leaping" at an average of 20 is represented as a point in this distribution for the activity "Leaping"). This figure shows a large variance in the rating of each activity and that participants mostly rated the fun above zero. Most activities have their means around the average fun for all activities (35), meaning that the participant fun is not strongly linked to it. Exceptions to this are the "Conflict", "Beaming" and to a lesser extent the "investigated", "Slow walk" and "Known" activities because their median differ from the average. The "Conflict" activity is the most interesting one, since its occurrence is high and can be easily interpreted as participants having more fun during conflict. It also shows that the distribution is fairly even and that no clear groups emerge. It would therefore be difficult to differentiate different type of player based on this information alone.

B. Feature Extraction

The continuous biometric signal and rating of the fun by the participant were divided into epochs of fixed time length as this method is useful for many statistical analyses. The epoch's duration has been set empirically to 5 sec with no overlapping. The 5 sec epoch duration provided the best trade off between a good temporal resolution and higher information (entropy) from the signals, being longer than the time constant of the physiological signals.

From the array of input modalities, features have to be extracted to suit machine-learning techniques. A total of 244 features have been extracted from all the data sources, which can be grouped in two different categories: time dependent and time independent. Time dependent feature consists of all the different time series presented in the Table II. From the time dependent signal, statistical features are extracted for each epoch. Those statistical features were the mean, min, max, skewness, kurtosis and the trend. Spectral power density was also extracted in bands from 0 Hz to 10 Hz.

Time independent data are answers to the questionnaires. All of the questions were based on a rating from 1 to 5 and are used as features appended to each epoch of the same participant. While these answers were obtained after the experiment so that they cannot be used in real time, they are

TABLE I. GENERAL STATISTICS OF THE DATABASE, AVERAGE (STANDARD DEVIATION) ACROSS ALL PARTICIPANTS

	ACU - S5M3	ACU - S9M3	ACS - S4M1	ACS - S5M3
Completion time [min]	14.5 (2.2)	8.62 (2.0)	9.9 (2.2)	14.6 (2.0)
Rated fun [-100; 100]	35.6 (20.3)	38.7 (22.1)	32.8 (22.8)	34.8 (22.2)
Heart Rate [beats/minutes]	75.0 (13.6)	75.4 (13.1)	73.3 (10.7)	73.8 (10.4)
Respiration Rate [resp./minutes]	24.3 (1.8)	24.6 (1.8)	24.9 (2.0)	24.8 (1.6)
Pupil diameter [cm]	0.45 (0.07)	0.46 (0.07)	0.42 (0.07)	0.44 (0.07)

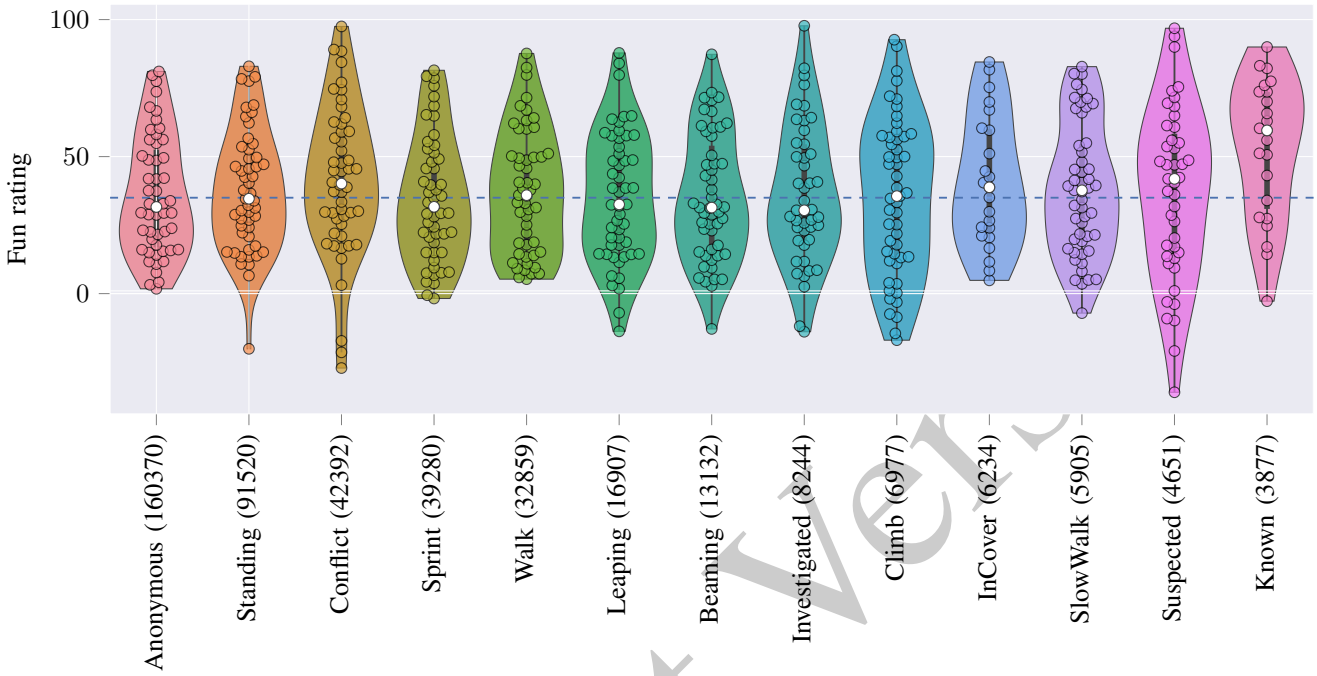

Fig. 2. Distribution of the average rating of the participants for each activity. The activities are arranged from most occurrences to least occurrences, with the occurrences in parentheses. The dotted blue line is the average of the fun for all the game sessions.

TABLE II. MODALITIES DESCRIPTION

ecg	Electrocardiogram and its derivative: heart rate and heart rate variability
emg	Electromyography of the right abductor pollicis longus (thumb abductor)
rsp	Respiration intensity (diameter of the thorax) and its derivative: respiration rate
eda	Electro dermal activity
pup	Pupil diameter (Smart Eye system)
eye	Eye information (Smart Eye system) such has position of the gaze on the screen, eye fixation, saccades and blinking
head	Pitch, Yaw, Roll of the head (Smart Eye system)
au	Facial Action Units (Noldus FaceReader)
lum	Screen luminosity, to capture interactions with pupil diameter
immrQ	Responses to the immersion questionnaire
nasaQ	Responses to the Nasa TLX questionnaire
acgame	Time played on previous entries in the Assassin's Creed series
difficulty	Self reported difficulty of the mission played
appreciation	Self reported appreciation of the game session
gender	Participant gender
age	Age of participant
spurious	A random value that will help in identifying truly useful features

representative of what time-independent features can bring to a classifier.

C. Data Preprocessing

The participants were divided into training and test set to the ratio of 3 training participants for 1 test participant. The test participants were kept for final analysis to evaluate accuracy on unseen data.

Missing values were filled in by the imputation of values using the average of the training set corresponding features. All features were then standardized so that the training features have a zero mean and unit standard deviation along feature type.

D. Labelling

It is possible to infer the level of fun on a linear scale by regression methods but those are subject to the limitations inherent to ratings. Indeed, this kind of rating is subjective and of limited use, as is, notably due to limitations such as interpersonal differences and non-linearity as reported in [29] and [30]. While the ratings in this experiment had a large numerical spectrum as opposed to more common rating-based questionnaires in which the participant is asked to choose his level of agreement on a scale of one to five, it entailed the same limitations.

Two different methods for reducing label variance were investigated. First, a simple threshold classification of the fun, i.e. the fun is classified according to its relation to a threshold value. For example, it is placed in one class if higher than the threshold or in another if lower. The choice of the threshold values still remains somewhat arbitrary but can be chosen using relevant statistical method. An example of threshold is the

mean fun of the participant during the game session. For the current project, the thresholds were thus chosen in relation to the mean ($\overline{\mathbf{x}}$) and standard deviation ($\sigma_{\mathbf{x}}$) of each game session as

$$\text{class}(x_i) = \begin{cases} \text{low fun} & \text{if} \quad (x_i - \overline{\mathbf{x}}) < -\frac{1}{3}\sigma_{\mathbf{x}} \\ \text{neutral fun} & \text{if} \quad -\frac{1}{3}\sigma_{\mathbf{x}} \leq (x_i - \overline{\mathbf{x}}) \leq \frac{1}{3}\sigma_{\mathbf{x}} \\ \text{high fun} & \text{if} \quad \frac{1}{3}\sigma_{\mathbf{x}} < (x_i - \overline{\mathbf{x}}) \end{cases}, \tag{1}$$

where x_i is the fun rating at time i in the game session and $\mathbf{x} = [x_1, \ldots, x_i, \ldots, x_{\text{last}}]^T$.

Second, ranking was chosen as a method of classification. Since the absolute level of the fun is subjective and suffers from the non-linearity of reporting. The differentiation of the absolute level of fun might give a clearer indication of the increase or decrease in the level of fun. From the changes of the absolute level of fun, a change in the fun ranking can be inferred, that is, if the participant reports an increase in the absolute fun, a change from a lower ranking of fun to a higher ranking of fun should occur. It is mathematically expressed as

$$\text{rank}(x_i) = \begin{cases} \min(-1, \text{rank}(x_{i-1}) - 1) & \text{if } x_i - x_{i-1} < -T \\ \max(1, \text{rank}(x_{i-1}) + 1) & \text{if } x_i - x_{i-1} > T \\ x_i = x_{i-1} & \text{otherwise,} \end{cases} \tag{2}$$

where T is an adjustable threshold that has been set to balance the classes evenly. Drift issues caused by the differentiation of the fun signal are limited by constraining the rank to a maximum and minimum. This ranking still implies some subjectivity in the choice of the magnitude of change in the absolute fun level considered enough for a rank change (threshold).

IV. Modelling

This section presents the models developed and how they are trained. Three different models are presented: one based on a regression technique and two based on classification techniques with different labelling methods, namely the classification and the ranking method.

A. Grouping and Validation Scheme

In preliminary experiments [26], prediction accuracy of the fun of the player was found to be higher when the training and test sets came from a single participant (intra-participant), than from training on multiple participants before making prediction on untested participants. While the prediction accuracy was better in the intra-participant case, it was of less interest for finding a specific physiological signature of the fun during a video game play, since it requires previous knowledge of the player. For this reason, the current study focused on inter-participant predictions, meaning that a signature was to be found on a set of players and then tested on another set to see if it generalizes well.

The dataset is split in two, a train set and a test set. The train subset is used in a cross validation scheme for hyper-parameter tuning and model selection. The test set is used to report final model accuracy and was only used once to report on model accuracy.

TABLE III. Classification results on the cross validation folds on the class label

	F1 score (standard deviation)	Accuracy
K Nearest Neighbours	0.354 (0.005)	0.415
Support Vector Classifiers	0.340 (0.012)	0.334
XGBoost	0.380 (0.018)	0.410
Multi-layer Perceptron	0.36	0.373
Most basic classifier	0.331 (0.013)	0.340

Since a general physiological signature of the fun is sought, the game events, which are specific to the Assassin's Creed's game, were discarded as features for fun prediction even if they were related to the fun experienced by participants.

B. Regression

The first machine learning technique evaluated was a regression technique, which, for each epoch, the average of the fun rating was to be predicted. Several regression algorithms from the Scikit-Learn library [31] were tested such as linear model (Elastic Net), Support Vector Machines (SVM) and Nearest Neighbours. An ensemble method, the optimized distributed gradient from the XGBoost library, [32] has also been tested. None of those regression algorithms were able to predict the fun state with accuracy. None showed a linear correlation between their predictions and the truth labels as the Pearson correlation coefficient stayed at zero. Noisy labels were assumed to be the main culprit of this poor performance. This is why a classification method with a limited number of classes might perform better by aggregating similar ratings and therefore limit the variance of the labels.

C. Classification

Classification techniques were then evaluated to learn to predict distinct state of fun for the player. To translate the rated fun into distinct state of fun, the method shown in (1) was used for each player. The number of classes was chosen as to maintain a good accuracy but also to keep a meaningful difference between classes. The three classes were interpreted as low fun, neutral and high fun. Those classes are therefore not absolute and are relative to the game session. Several classifiers, available in the Scikit-Learn library, have been tested for classification such as Nearest Neighbours, Support Vector Classifiers (SVM), Random Forest, logistic regression and Adaboost. An optimized distributed gradient boosting library (XGBoost [32]) was also tested. Hyper-parameters for each of the algorithm have been tuned by random searches in a threefold cross validation scheme on the training set. A multi-layer perceptron was also implemented using the Keras deep-learning library. Only a subset of those classifier is presented here which correspond to the best among their family of classifier. Finally, a most basic classifier that predicts at random following the probability distribution of each class was also implemented as a point of comparison to chance level accuracy. The F1 score was chosen as a scoring metric. This scoring method is a weighted average of the precision and recall globally across the total true positives, false negatives and false positives. The average F1 scores across three folds are presented in Table III for each classifier including the most basic classifier which indicates the score corresponding to a random guess.

TABLE IV. FEATURES RANKING FROM THE XGBOOST CLASSIFIER

	Classification		Ranking	
rank	modality	score	modality	score
1	rsp	0.164	rsp	0.397
2	ecg	0.141	eye	0.139
3	eye	0.131	au	0.118
4	head	0.094	ecg	0.093
5	immrQ	0.080	head	0.081
6	emg	0.080	lum	0.038
7	nasaQ	0.080	emg	0.036
8	au	0.059	immrQ	0.029
9	pup	0.052	eda	0.022
10	eda	0.034	pup	0.021
11	ACgame	0.025	nasaQ	0.013
12	lum	0.024	age	0.006
13	age	0.008	ACgame	0.005
14	difficulty	0.004	spurious	0.002
15	appreciation	0.003	appreciation	0.001
16	spurious	0.002	difficulty	0.001
17	gender	0.001	game	0.000
18	game	0.000	gender	0.000

TABLE V. CONFUSION MATRIX OF THE XGBOOST CLASSIFIER ON THE TEST SET

	predicted low fun state	predicted neutral state	predicted high fun state
actual low fun state	1091	255	1721
actual neutral state	712	218	1535
actual high fun state	737	332	2304

From these results, it appears that the XGBoost classifier reached better performance for this task with respect to the F1 score. This classifier also gave an indication of features importance, which is explained in detail in [33]. The importance of each feature as computed by the XGBoost classifier on the training set is presented in Table IV. Features from the same modality (signal) are grouped to give a preview of the importance of each modality. The spurious feature, a random value added to each feature sample, can help determine a threshold score indicating unimportant modalities.

Features extracted from the respiratory activity were the most used, followed by features coming from the electrocardiogram, head tracking of the Smart Eye Pro system, questionnaires and electromyography. Facial action units and pupil size also contributed to the inference but to a lesser degree as did the electro-dermal activity. The other modalities such as the previous experience with the Assassin's Creed series's games, the perceived difficulty, the age, the general appreciation, the game played (Unity or Syndicate) and the gender are not or only marginally contributing to the inference. They indeed have a similar or lower score than that of the spurious feature. It is important to note that the distribution of participants was heavily skewed towards male participants (184 males against 9 females) and therefore the gender cannot be discounted as an important feature for further research.

To confirm the generalization of the learning from the XGBoost classifier, it was tested on the participants from the test set. A confusion matrix of the result is presented in Table V, along with a matrix presenting precision, recall and F1 score of the classifier in Table VI.

These results showed that the classifier leaned toward predicting a high fun state, indeed it identified 62% of the occurrence of a high fun state. It was unable to predict a neutral fun state, but fairly capable of predicting a low fun state. One hypothesis explaining this might be the fact that

TABLE VI. PRECISION RECALL AND F1 SCORE OF THE XGBOOST CLASSIFIER

	precision	recall	F1-score	support
low fun state	0.43	0.36	0.39	3067
neutral state	0.27	0.09	0.13	2465
high fun state	0.41	0.68	0.52	3373
avg / total	0.38	0.41	0.38	8905

TABLE VII. CLASSIFICATION RESULTS ON THE CROSS VALIDATION FOLDS ON THE RANK LABEL

	F1 score (standard deviation)	Accuracy
K Nearest Neighbours	0.33 (0.012)	0.415
Support Vector Classifiers	0.302 (0.008)	0.311
XGBoost	0.351 (0.022)	0.360
Multi-layer Perceptron	0.347	0.344
Basic classifier	0.344 (0.011)	0.341

players were playing a game they never played before (a recruitment criterion) and, thus, were mostly in a state of fairly high fun during the whole session. This could limit the difference between the low and the high fun state increasing the classification difficulty.

D. Ranking

As explained in the Section III-D, simple classification of rating from the player entails inherent limitations. To help circumvent some of these limits, classification based on a ranking was conducted to test if a better accuracy could be achieved. The same procedure as for the classification of the fun was applied here, with the difference that instead of predicting the average of the fun rating during the epoch, the average of the fun ranking, shown in (2), was to be predicted. Results from the same classifiers as before, retrained for ranking is shown in Table VII. It can be seen from a comparison between the two methods that the ranking method did not help classification. It seems that instead of reducing label noise, it increased it.

Features importance from the XGBoost classifier is also presented in Table IV, which shows that the modalities were ranked similarly in both classification and ranking, indicating a certain robustness to the features' rank.

V. DISCUSSION

The goal of this study was to find a physiological signature of the player's level of fun during a video game session by converging multiple sources of data, namely the physiological signals and questionnaire answers. Those sources of data served in the prediction of the fun factor, which was rated by the participant while watching a playback of his/her game session. The results of the different classifiers showed that the best classifier was better at predicting the player's level of fun than the most basic classifier (chance) by improving the F1 score by 15%, 0.38 against 0.331. One hypothesis for this limited improvement is due to noisy labels, which is a direct effect of inter-individual variability [7], i.e. differences in the subjective rating of the fun by each participant. This fact was also reported at an earlier stage of this project [26]. Indeed, accuracy was much higher in intra-participant prediction as opposed to predictions on an unseen set of participants. The addition of more participants, facial features and their responses to questionnaires has improved inter-participant prediction, but not by a large factor. There is therefore a need to first

categorize a player by their way of rating the fun. The method for ranking the fun presented in this paper still falls short of removing the impact of inter-individual variability.

With a goal of real-time inference of the fun and in light of the feature importance ranking, some type of modalities might be more useful than others like the electrocardiography, respiration and eye and head tracking. While head and eye movements are not intrusive measures, as they were acquired by cameras, an electrocardiogram and a respiration transducer are currently more intrusive for the player. Those are important considerations if such inference is to be deployed at larger scale. Questionnaires bring a small amount of information and are not intrusive during game play, but require additional time either before, or after a play session.

While the accuracy remains modest at 41% amongst 3 classes, it followed expectations as the fun rating is inherently subjective and suffers from non-linearity of reporting and inter-individual variability. This accuracy should nonetheless be useful to create a statistically significant profile of a player given many samples of similar events in a game session. Indeed, taking the conflict state as an example, it is occurring an average of 50 times in 5 sec epochs during a game session. By predicting the fun level with an accuracy of 41% each time, the mean fun level of the predictions should have a relatively low variance, which gives a good indication of the player's appreciation of conflicts.

VI. Conclusion

This paper presented a classifier capable of predicting the fun rating that could be a major step in the development of adaptive gaming. Indeed, by inferring the level of fun over multiple events, its noisy nature should get averaged out to give a more accurate representation of the likes and dislikes of a player. The method presented in this paper allowed the evaluation of the importance of each source of data. This should help in sensor selection for further research by favouring a heart rate monitor, eye/head tracker and respiration belt transducer. Future works will consist in identifying which modalities are less prone to affect gameplay and better performing at predicting the fun in real time. Also works in collaboration with game designers could include game events to the set of modalities used for prediction. Since an adaptive game has a direct impact on those game events, a careful integration of those events to the features is necessary as they close the information loop. Profiling the player to help better predict fun during gameplay is also considered as a way to increase prediction accuracy by reducing inter participant variability. Further development, which is one of the main goals of the FUN*ii* project, is the development of an adaptive game prototype that will take advantage of the predicted fun to adjust itself in a way that optimizes the gaming experience. Finally, even if this paper focuses on the fun in gaming, its conclusion should be applicable to a wide range of intelligent systems that uses physiological readings has proxy for other psychological states such as stress, workload and engagement, for example. It should help build adaptive systems that might maximize health, performance or security of workers or patients.

Future works include the creation of an adaptive game based on the fun prediction. Indeed, by collecting the fun predictions over time and by associating them with game events, an appreciation of each game event can be inferred. With that appreciation profile the adaptive game will tailor itself to the player preferences by modifying the game scenarios in real time.

Acknowledgment

This project was funded by NSERC-CRSNG, Ubisoft Québec and Prompt. Additional thanks to Nvidia for providing a video card for deep learning analysis through their GPU Grant Program.

References

[1] I. Granic, A. Lobel, and R. C. M. E. Engels, "The benefits of playing video games," *American Psychologist*, vol. 69, no. 1, pp. 66–78, 2014.

[2] D. Djaouti, J. Alvarez, and J.-P. Jessel, "Classifying serious games: The G/P/S model," *Handbook of research on improving learning and motivation through educational games: Multidisciplinary approaches*, no. 2005, pp. 118–136, 2011.

[3] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, "A systematic literature review of empirical evidence on computer games and serious games," *Computers and Education*, vol. 59, no. 2, pp. 661–686, 2012.

[4] Entertainment Software Association, "Essential Facts: About the computer and video game industry," *Entertainment Software Association*, p. 11, 2016.

[5] K. Bantinaki, "The paradox of horror: Fear as a positive emotion," *Journal of Aesthetics and Art Criticism*, vol. 70, no. 4, pp. 383–392, 2012.

[6] W. van den Hoogen, K. Poels, W. a. IJsselsteijn, and Y. a. W. de Kort, "Between Challenge and Defeat: Repeated Player-Death and Game Enjoyment," *Media Psychology*, vol. 15, no. 4, pp. 443–459, 2012.

[7] R. L. Mandryk, K. M. Inkpen, and T. W. Calvert, "Using psychophysiological techniques to measure user experience with entertainment technologies," *Behaviour & Information Technology*, vol. 25, no. 2, pp. 141–158, 2006.

[8] A. E. Zook and M. O. Riedl, "A temporal data-driven player model for dynamic difficulty adjustment," *Proceedings of the 8th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2012*, pp. 93–98, 2012.

[9] W. Wirth, F. Ryffel, T. Von Pape, and V. Karnowski, "The Development of Video Game Enjoyment in a Role Playing Game," *Cyberpsychology, behavior and social networking*, vol. 16, no. 4, pp. 260–4, 2013.

[10] P. Desmet, "Measuring Emotion: Development and Application of an Instrument to Measure Emotional Responses to Products," in *Funology: From usability to enjoyment*, 2003, pp. 111–123.

[11] R. a. Bartle, "Players Who Suit MUDs," *Mud*, p. 1, 1999.

[12] N. Yee, "Motivations for Play in Online Games," *CyberPsychology & Behavior*, vol. 9, no. 6, pp. 772–775, 2006.

[13] G. N. Yannakakis and J. Hallam, "Real-time game adaptation for optimizing player satisfaction," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 1, no. 2, pp. 121–133, 2009.

[14] C. Pedersen, "Modeling Player Experience through Super Mario Bros Supervisor Georgios Yannakakis," *Technology*, no. August, pp. 132–139, 2009.

[15] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with Computers*, vol. 21, no. 1-2, pp. 133–145, 2009.

[16] J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, "Psychophysiological Science: Interdisciplinary Approaches to Classic Questions About the Mind," in *Handbook of Psychophysiology*, 2000, pp. 3–22.

[17] M. D. Robinson and G. L. Clore, "Belief and feeling: evidence for an accessibility model of emotional self-report." *Psychological bulletin*, vol. 128, no. 6, pp. 934–960, 2002.

[18] L. E. Nacke, "An Introduction to Physiological Player Metrics for Evaluating Games," in *Game Analytics*, M. Seif El-Nasr, A. Drachen, and A. Canossa, Eds. London: Springer London, 2013, pp. 585–619.

[19] G. Durantin, J. F. Gagnon, S. Tremblay, and F. Dehais, "Using near infrared spectroscopy and heart rate variability to detect mental overload," *Behavioural Brain Research*, vol. 259, pp. 16–23, 2014.

[20] F. Dehais, M. Causse, F. Vachon, and S. Tremblay, "Cognitive conflict in human–automation interactions: A psychophysiological study," *Applied Ergonomics*, vol. 43, no. 3, pp. 588–595, may 2012.

[21] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio, "Basic emotions are associated with distinct patterns of cardiorespiratory activity," *International Journal of Psychophysiology*, vol. 61, no. 1, pp. 5–18, 2006.

[22] E.-H. Jang, B.-J. Park, M.-S. Park, S.-H. Kim, and J.-H. Sohn, "Analysis of physiological signals for recognition of boredom, pain, and surprise emotions." *Journal of Physiological Anthropology*, vol. 34, pp. 1–12, 2015.

[23] A. Dekker and E. Champion, "Please Biofeed the Zombies: Enhancing the Gameplay and Display of a Horror Game Using Biofeedback," in *Proc. of DiGRA*, 2007, pp. 550–558.

[24] D. Emmen and G. Lampropoulos, "BioPong: Adaptive Gaming Using Biofeedback," *Creating the Difference: Proceedings of the Chi Sparks 2014 Conference*, no. 1, pp. 100–103, 2014.

[25] C. Chamberland, M. Grégoire, P.-e. Michon, J.-c. Gagnon, and L. Philip, "A Cognitive and Affective Neuroergonomics Approach to Game Design," *59th Annual Meeting of the Human Factors and Ergonomics Society*, no. 2007, pp. 1075–1079, 2015.

[26] A. Clerico, C. Chamberland, M. Parent, P.-e. Michon, S. Tremblay, T. H. Falk, J.-C. Gagnon, and P. Jackson, "Biometrics and classifier fusion to predict the fun-factor in video gaming," *IEEE Conference on Computational Intelligence and Games (CIG'16)*, pp. 233–240, 2016.

[27] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton, "Measuring and defining the experience of immersion in games," *International Journal of Human Computer Studies*, vol. 66, no. 9, pp. 641–661, 2008.

[28] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Advances in Psychology*, vol. 52, no. C, pp. 139–183, 1988.

[29] G. N. Yannakakis and H. P. Martínez, "Ratings are Overrated!" *Frontiers in ICT*, vol. 2, no. July, p. 5, 2015.

[30] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't Classify Ratings of Affect; Rank Them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 314–326, jul 2014.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in {P}ython," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[32] T. Chen and C. Guestrin, "XGBoost : Reliable Large-scale Tree Boosting System," *arXiv*, pp. 1–6, 2016.

[33] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," *Elements*, vol. 1, pp. 337–387, 2009.