Integrating machine learning and statistical analysis to examine the effect of gamification on biosignals, performance, and motivation in soccer

Rebecca Lennartz[*1], Maike Stoeve[1], Nandu Kumarampulakka[2], Karolina Attri[2], Lucas Wittmann[1], Matthias Witte[2], Bjoern M. Eskofier[1 3], Eva Dorschky[1]

[1] Machine Learning and Data Analytics Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-University Erlangen-Nürnberg, Erlangen, Germany
[2] Adidas Innovation Athlete Performance, adidas AG, Herzogenaurach, Germany
[3] Translational Digital Health Group, Institute of AI for Health, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

[*] rebecca.lennartz@fau.de

ORIGINAL ARTICLE

Submitted: April 15, 2025

Accepted: March 9, 2026

Published: March 30, 2026

Editor-in-Chief

Claudio R. Nigg, University of Bern, Switzerland

Guest Editors

Thorsten Stein, Karlsruhe Institute of Technology, Germany

Bernd Stetter, Karlsruhe Institute of Technology, Germany

ABSTRACT

Motivation, a complex construct that influences behavior, plays a critical role in an athlete's success and has been extensively researched in sports sciences. However, motivation is still primarily assessed through self-report motivational questionnaires, resulting in a lack of objective, continuous measurements during athletic performance. Biosignals, increasingly used to assess psychological processes, have gained relevance due to their integration into wearable sensors, allowing non-stationary and unobtrusive data collection. Therefore, we assessed performance data, cardiovascular signals, and eye-tracking metrics to investigate the psychological and physiological processes under two different conditions designed to induce different motivational states via gamification.

In this study, gamification elements, aligned with self-determination theory, were used to influence soccer players' motivation in an immersive space, with N = 42 participants completing a passing drill in both Gamified and Non-Gamified scenarios. Features were extracted from session recordings and wearable sensors to assess whether performance or biosignals differed between conditions using a combination of machine learning and conventional statistical analysis. While self-report questionnaires and performance metrics revealed no significant differences, the machine learning classifiers were able to distinguish between scenarios based on eye-tracking-related features. The best-performing model, a k-nearest neighbor classifier, reached a macro F1-score of 82.75 %. We identified by feature importance methods that blink behavior and pupil dynamics, indicative of visual attention, were the main contributors. This study contributes to a deeper understanding of

the value of integrating multimodal data and advanced evaluation methods to uncover implicit processes in applied sports contexts involving complex and heterogeneous data.

Keywords

gamification, motivation, performance, biosignals, machine learning

Citation:

Lennartz, R., Stoeve, M., Kumarampulakka, N., Attri, K., Wittmann, L., Witte, M., Eskofier, B. M., & Dorschky, E. (2026). Integrating machine learning and statistical analysis to examine the effect of gamification on biosignals, performance, and motivation in soccer . *Current Issues in Sport Science*, *11*(3), Article 003. https://doi.org/10.36950/2026.3ciss003

Introduction

The concept of motivation is defined as the hypothetical construct that influences the initiation, direction, intensity, persistence, quality, and continuation of behavior (Roberts & Treasure, 2012; Vallerand, 2007). In sports, motivation is a crucial factor for athletes' success and persistence (Vallerand, 2007). Despite being one of the most extensively researched topics in psychology, motivation lacks a single, universally accepted definition and is instead described through a continuum of various theories (Roberts & Treasure, 2012; Ryan et al., 2019). One of the most widely studied and popular theories is the concept of intrinsic motivation (IM) and self-determination theory (SDT) by Deci and Ryan (1985). According to this theory, motivation exists on a spectrum ranging from IM to extrinsic motivation (EM) and amotivation, with IM being regarded as having a higher quality and longer-lasting impact (Deci & Ryan, 1985). Motivational states occur at three levels: global (general orientation), contextual (domain-specific), and situational (momentary state during an activity) (Vallerand, 2007).

SDT identifies three fundamental psychological needs, relatedness (a sense of connection with others), competence (feeling capable and effective), and autonomy (a sense of personal agency), that must be fulfilled for individuals to experience IM on each of the levels (Vallerand, 2007). Multiple studies have demonstrated that experiencing higher levels of IM, driven by the fulfillment of the three needs, leads to higher participation in exercise, greater long-term adherence (Teixeira et al., 2012), and is linked to increased well-being and performance. Consequently, findings from motivation research in sports should be integrated into coaching, teaching, and the prevention of motivational issues (Roberts & Treasure, 2012).

Since motivation is a psychological construct rather than a directly observable variable, researchers assess it using indirect and measurable indicators, typically by evaluating its observable consequences in behavior, affect, and cognition (Vallerand & Losier, 1999). Higher levels of IM are associated with positive affective outcomes, such as reduced fatigue and lower perceived exertion, while also enhancing long-term satisfaction, interest, and enjoyment. Cognitive outcomes include concentration, task perception, learning, and the ability to recall situations and errors. Finally, behavioral outcomes encompass attendance, objective performance, effort, continued engagement, and persistence (Touré-Tillery & Fishbach, 2014).

Among these indirect methods, self-reported questionnaires are the most widely used instruments to capture motivation aligned with various motivational theories and adapted to different contexts such as edu-

cation, the workplace, or athletic performance at both contextual and situational levels (Clancy et al., 2017). However, self-reported questionnaires have several limitations, including a lack of objectivity and the inability to provide continuous, real-time measurements during sports activities. Motivation can only be assessed retrospectively, which may introduce response bias and limit the ability to monitor specific interventions in real time. Additionally, choosing the right questionnaire is further complicated by the wide range of psychological theories and questionnaires, requiring deep theoretical knowledge. Other limitations include low sensitivity to group or cultural differences, a lack of practical applicability, and issues with reliability and interpretability (Anshel & Brinthaupt, 2014; Vealey et al., 2019).

Biosignals offer an objective alternative to questionnaires for assessing psychological processes, with applications in stress research (Giannakakis et al., 2022), emotion recognition (Jerritta et al., 2011), and cognitive load assessment (Mutlu-Bayraktar et al., 2019). Advances in wearable sensors, such as chest-worn electrocardiograms (ECG), portable eye-tracking (ET) glasses, or inertial measurement units (IMUs), have made biosignal acquisition more accessible, even in complex environments. Despite this progress, the use of biosignals remains limited in motivational research. Herlambang et al. (2019) examined workload and fatigue caused by motivational changes using heart rate (HR) variability (HRV) and pupillometry. Similarly, affective responses were examined through emotion recognition via facial expressions and electrodermal activity to detect motivational joy and boredom (Korn & Rees, 2019).

While some correlations based on biosignals are well understood, expert knowledge remains limited in other contexts, particularly when dealing with heterogeneous data. Machine learning (ML) can help uncover complex patterns and extract relevant features. For instance, Vorberg et al. (2023) showed that ML-based regression models using HR(V) provided deeper insights into stress-coping strategies than traditional regression analysis. Stoeve et al. (2022) demonstrated that stress and non-stress conditions can be classified using ET data recorded in virtual reality sports scenarios with an accuracy of 87.3 %, while also investigating the influence of feature subsets. To investigate motivation and its impact on performance and cognitive processes, motivation must be systematically manipulated according to a suitable theoretical framework, in this case, SDT. A widely used approach involves modifying the environment through gamification, which Deterding et al. (2011) defined as "the use of game elements in non-game contexts". For example, positive feedback or progress indicators can support the need for competence, while customizable profiles and avatars address autonomy. Group interactions, on the other hand, enhance relatedness (Francisco-Aparicio et al., 2013; Seaborn & Fels, 2015).

The presented theoretical frameworks have been tested and evaluated in different applications and fields. While Mekler et al. (2017) found that points and leaderboards improved performance but not IM, Hanus et al. (2015) even reported a decrease in motivation and exam scores in educational settings. In contrast, gamification enhanced intrinsic need satisfaction in online communities (Xi & Hamari, 2019) and sports apps (Bitrián et al., 2020). Additionally, Sotos-Martínez et al. (2024) observed increased IM and need-satisfaction in gamified physical education. These inconsistencies highlight the need to explore how specific game elements influence motivation depending on context and individual differences.

The reviewed research highlights the importance of IM in sports, as it directly impacts performance and persistence. Additionally, the use of wearable sensors for capturing biosignals and behavioral data has demonstrated potential as an objective, real-time measurement tool in psychological research. To investigate whether wearable sensors can effectively assess motivation in high-intensity sports activities, we implemented a soccer drill within an immersive environment known as the Igloo (Igloo Vision Ltd., Shropshire, UK), a 360° projection system with a six-meter-diameter circular space covered in artificial turf, integrating virtual and real-world elements. Projectors create an

interactive environment simulating a passing drill, allowing participants to engage with a real soccer ball that can be passed against a rebound barrier positioned along the edge of the space. This setup was chosen primarily for its high level of experimental control and standardizability, enabling the precise implementation of gamification elements based on SDT to systematically influence the player's motivation. Although the drill addressed relevant soccer-related skills such as scanning behavior and decision making, while preserving natural movement patterns and realistic ball interaction, it was not intended to directly assess soccer performance.

In this study, we focused on how situational motivation influences both performance and psychophysiological responses. To investigate this, participants completed both a Non-Gamified and a Gamified scenario, with the latter systematically designed to enhance situational motivation based on self-determination theory. Therefore, the study explores the feasibility of measuring situational motivation using session recordings and wearable sensors, in addition to traditional questionnaires, serving as an established benchmark reference for evaluating the sensor-based methods.

Based on prior research, we hypothesize that the self-reported situational motivation will be higher in the Gamified scenario compared to the Non-Gamified scenario. Furthermore, we expect biosignals and performance metrics to differ between scenarios, consistent with the assumption that gamification-induced increases in motivation influence psychophysiological responses. While these hypotheses address expected differences in motivation and biosignals between scenarios, the feature identification and machine learning components were conducted in an exploratory manner to uncover additional patterns beyond the predefined expectations.

Specifically, the purposes of this work were to:

1. Examine the effect of gamification on situational subjective motivation using state-of-the-art motivation questionnaires.
2. Identify relevant features from session recordings and wearable sensors and determine whether performance metrics or biosignals differ between experimental conditions using a mixed approach involving ML and traditional, inferential statistical analysis.
3. Analyze how extracted features relate to motivational questionnaire results to explore the relationships among gamification, motivation, performance, and biosignals.

Methods

Participants

A total of $N = 42$ recreational soccer players (17 female, 25 male; age: $M = 27.7$, $SD = 7.8$ years) participated in this study, conducted between November 7, 2023, and January 9, 2024. The sample size reflects a trade-off between resource constraints and generalizability, while exceeding those of similar biomarker-based psychological studies (Sajno et al., 2023; Stoeve et al., 2022). Written informed consent was obtained from all participants, and the study was approved by the ethics committee of Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (Ethics ID: 22-37-B). All participants met specific health-related criteria, including the absence of current injuries or health risks. Eight participants were removed from the final analysis due to technical issues or methodological errors.

The final sample consisted of $N = 34$ participants (13 female, 21 male, age: $M = 28.70$, $SD = 8.00$ years). Of these, $n = 28$ were still actively playing soccer at the time of the study, while $n = 6$ had previously played but were no longer active. The participants had an average of $M = 16.9$ years ($SD = 10.6$) of soccer experience. Most identified as amateur players playing organized soccer, i.e., in clubs (60.6 %), followed by amateur recreational players (24.2 %), and semi-professional or professional players (15.2 %).

Participants were further questioned regarding their experience with gamified sports applications such as

Nintendo Wii. A total of n = 32 participants reported previous experience with such applications. Regarding video game frequency, 42.4 % played multiple times per month, 30.3 % had stopped playing, and 18.2 % had never played video games. Only 9.1 % reported playing weekly or more frequently. Participants were primarily motivated by their interest in soccer, the exploration of new training methods, and curiosity about virtual/extended reality.

Hardware

Data recording took place inside the Igloo environment, which was implemented and controlled via a custom Python-based script integrated into the Unity game engine (Unity Technologies, San Francisco, USA). While the projected environment was entirely virtual, participants interacted with a real soccer ball. Ball tracking was performed using a video camera (ELP 1080P webcam Full HD, China) and a Python-based tracking algorithm based on Ultralytics YOLOv8 (Ultralytics, Los Angeles, USA). This system enabled real-time detection of the ball's exact position, allowing automatic goal detection. To compensate for potential tracking delays, a 0.5-second grace period was applied.

Environmental data were controlled and recorded by a central WebSocket, which saved the video output of the ball-tracking algorithm, while all logged events were stored in a separate .txt file. A wearable sensor captured ECG, specifically, a one-channel ECG (Lead I of Einthoven's triangle), recording at 256 Hz (Portabiles GmbH, Erlangen, Germany), was worn on a chest strap while storing data internally for cardiac analysis. Additionally, Tobii Pro Glasses 3 (Tobii AB, Stockholm, Sweden) recorded binocular ET data at 100 Hz, including pupil diameter (PD) and gaze coordinates, as well as linear acceleration and angular velocity from an integrated IMU sensor. All ET data were recorded using iMotions software (iMotions A/S, Copenhagen, Denmark) and later exported as .csv files.

Questionnaires were implemented digitally using Qualtrics (Qualtrics LLC, Provo, USA), an online survey platform. The responses were collected and exported as .csv files and included the following questionnaires:

- General questionnaire covering demographics, soccer experience, familiarity with video games, and prior use of gamified sports applications,
- Behavioral Regulation in Sport Questionnaire (BRSQ) (Lonsdale et al., 2008), based on SDT to assess the general motivation towards soccer, rated on a 7-point Likert-Scale.
- Rating of Perceived Exertion (RPE) scale/Borg Scale (Borg, 1970) measuring perceived exertion from 6 (no exertion) to 20 (maximal exertion).
- User Experience Questionnaire (UEQ) (Laugwitz et al., 2008), evaluating user experience and usability.
- Intrinsic Motivation Inventory (IMI) (Ryan et al., 1983), specifically the Interest/Enjoyment subscale, was used as a self-report measure of activity-specific IM based on SDT on a 7-point Likert Scale.
- Player Experience of Need Satisfaction (PENS) questionnaire (Ryan et al., 2006) assessing perceived competence, autonomy, and relatedness, adapted from the IMI for video games, rated on a 7-point Likert Scale.
- Final questionnaire evaluating the overall Igloo experience, feedback on gamification elements, and participants' willingness to use the system in the future

Game design

The passing drill design consisted of nine small soccer goals, evenly projected onto the canvas. Goals opened sequentially in a fixed order, highlighted for three seconds before closing again. Participants were instructed to score as many goals as possible during the three-minute drill, which included a 15-second break halfway through, allowing a maximum of 60 goals per drill by shooting against the rim.

Figure 1 *Picture of the Gamified Drill Including Gamification Elements*

The yellow-framed goal is currently open, while the bar above the goal indicates the remaining opening time. The displayed word "GOAL" appears after a goal was successfully hit. Gamification elements from left to right: Leaderboard including a goal count, remaining time, streak counter, and badges.

Two scenarios were implemented, following the previously described logic: a Gamified and a Non-Gamified version. Both included visualizations of the current score and remaining time, as well as background sounds such as auditory cues for goal openings, successful hits, and misses. Additionally, ambient stadium noises, crowd sounds, and countdown signals were played before the drill started and in the final 10 seconds.

The Gamified scenario featured additional elements designed according to SDT to address the psychological needs of autonomy, competence, and relatedness (Francisco-Aparicio et al., 2013). These elements were identical for all participants and filled with preset data except for the current participant's score. Specifically, the gamification elements included:

- Leaderboard: Displayed five fictional players with preset scores, initially placing the participant at the bottom with zero points. As goals were scored, participants could move up the leaderboard, enhancing the feeling of competence.
- Team leaderboard: Similar to the individual leaderboard, it ranked the participant's team at the lowest rank out of four, fostering. competence, relatedness, and autonomy by allowing participants to choose a team before the drill.
- Streak counter: Activated when participants scored consecutive goals without missing, reinforcing competence. Missing a goal resets the streak.

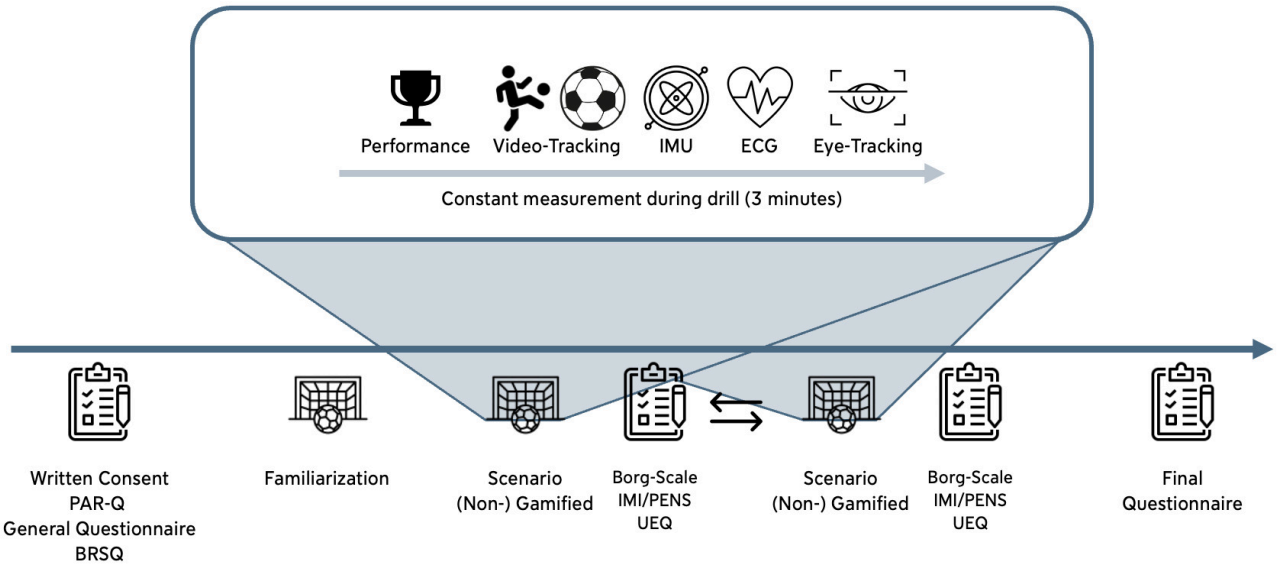

Figure 2 *Study Procedure, Including the Order of the Questionnaires, Measurements, and the Drill Scenarios. For simplicity, the baseline measurements were omitted.*

- Achievement badges: Awarded for specific in-game accomplishments. The "Around the World" badge was earned by hitting every goal at least once, while the "On Fire" badge was awarded for scoring ten consecutive goals, supporting competence.
- Enhanced sound effects: Additional supportive voice feedback to encourage and reinforce motivation.

During the break, the environment was darkened while the score, remaining time, and gamification status were displayed. Figure 1 displays an image of the inside of the Igloo environment, including the gamification elements.

Experimental design

Participants completed both experimental conditions, the Gamified and Non-Gamified versions, presented in randomized, counterbalanced order. After excluding invalid participants, n = 18 completed the Non-Gamified scenario first (7 female, 11 male), while n = 16 started with the Gamified scenario (6 female, 10 male).

Before data recording, all participants were welcomed, informed about the study procedure, and provided written informed consent. After confirming eligibility, they were equipped with the IMU and ECG sensors before completing the general questionnaire and the BRSQ. During questionnaire completion, a 2-minute baseline ECG measurement was recorded. Participants then underwent a brief familiarization phase, where they were introduced to the Igloo environment, measurement tools, safety protocol, and drill. They had the opportunity to test the setup; however, to avoid bias, no explanations regarding the gamification elements were given.

Both drills followed the same procedure regardless of condition. First, the ET glasses were calibrated, followed by a 30-second baseline recording inside the Igloo under the same lighting conditions as during the drill. After completing a drill, participants took a short break while answering the IMI, PENS, UEQ, and Borg Scale regarding the scenario they had just completed. After completing both drills, they also filled out the final questionnaire before being debriefed and thanked for their participation. An overview of the study procedure is displayed in Figure 2.

Data processing

Questionnaire responses were extracted from the .csv files, cleaned, and scored according to the procedures outlined in the respective publications. For the BRSQ, the IM-General subscale was used for IM, and the amotivation subscale for amotivation. EM was calculated by averaging the subscales of external regulation and introjected regulation.

Task performance, including the number of scored goals, the highest streak count, the average streak count, and achieved badges, was extracted from the text file. Due to the WebSocket-based communication, events were not always logged in chronological order but rather in the order in which they were received. Therefore, the relevant events had to be sorted and matched with their corresponding counterparts (e.g., a successful goal with the related goal-opening event). Scored goals and streaks were checked for plausibility and corrected if necessary. Additionally, the highest achievement percentage for the badges was extracted. The "hit time", defined as the duration between a goal's opening and its successful completion, was derived from the sorted events, and the mean hit time was calculated while excluding goals that were not hit in time.

ECG data preprocessing was performed using BioPsyKit, an open-source Python package for biopsychological data analysis (Richer et al., 2021). First, HR was derived from RR intervals extracted after noise reduction using high-pass filtering with a 5th-order 0.5 Hz Butterworth filter, followed by powerline filtering (50 Hz) and QRS complex detection using the Neurokit2 library (Makowski et al., 2021). Artifacts in RR intervals were mitigated by removing physiological outliers (HR $\leqslant$ 45 bpm or $\geqslant$ 210 bpm), as well as statistical outliers in RR intervals ($\leqslant$ 2.576 σ) and differences in successive RR intervals ($\leqslant$ 1.96 σ). Removed RR intervals were replaced with the average value of the 10 preceding and 10 succeeding RR intervals.

Preprocessing and feature extraction of the ET-data followed the proposed pipeline by Stoeve et al. (2022). To mitigate noise in the PD data before feature extraction, an adapted preprocessing pipeline based on Kret and Sjak-Shie (2019) was used. First, blinks and samples marked as invalid by the eye tracker had to be removed. Blink detection was performed by first interpolating segments of data loss in the PD if the segment was shorter than 40 ms (Nyström et al., 2024) before marking missing data segments longer than 70 ms as blinks (Bafna et al., 2020). For the results on eye-metrics, participants who did not reach a threshold of 75 % valid samples, including blinks, were removed from the analysis.

To increase the robustness of blink detection, all samples identified as blinks in at least one eye were marked accordingly, and segments separated by less than 100 ms were merged (Nyström et al., 2024). Derived measures from blinks have been reported to have a relationship with cognitive load and task difficulty (Bafna et al., 2020). Therefore, the resulting blink-labeled samples were used to compute blink-related features, including the number of blinks, the blink frequency (Hz), the average blink duration (ms), and the average blink interval duration (ms), representing the mean time between consecutive blinks.

After blink detection, additional physiologically implausible PD values outside the range of 1.5 to 9 mm were discarded. This was followed by three filtering steps: Samples exceeding a dilation speed above a threshold based on the median absolute deviation (MAD) were removed, accounting for blinks and eyelid occlusions. The resulting gaps were interpolated, and the signal was smoothed to generate a trendline. Samples deviating from this trendline were iteratively removed until no further outliers were detected. A final sparsity filter marked temporally isolated samples near measurement gaps (> 40 ms) and excluded segments shorter than 50 ms. The two cleaned PD signals (one per eye) were then merged using piecewise cubic Hermite interpolating polynomial (Pchip) interpolation (Dan et al., 2020). Lastly, a baseline correction was applied by subtracting the median of a one-second window of continuous, valid samples recorded as close as possible to the beginning of the drill from the PD signal. This reduced the impact

of random pupil-size fluctuations and provided the relative change in PD compared to the baseline (Mathôt et al., 2018). Further, the rate of change in PD, referred to as the slope, was determined by fitting a linear least-squares function to the PD for each segment, respectively, each half-time, and the timeout (Baltaci & Gokcay, 2016).

Further features were the index of pupillary activity (IPA) and the low/high IPA (LHIPA), which reflect cognitive workload and mental effort. Following the pipeline proposed by Duchowski et al. (2020), samples were removed within 200 ms before the start and after the end of a blink, as identified in the previous steps. The cleaned raw PD signal of both eyes was then processed using a Daubechies-4 wavelet decomposition to compute IPA and LHIPA, which were subsequently averaged across eyes.

In addition to PD-related parameters, fixation-related features were derived, since they are associated with cognitive processing time (Falkmer et al., 2008). First, invalid samples and those marked as blinks were removed from the two gaze vectors recorded by the eye tracker, followed by a Pchip interpolation to fill the gaps in the data. The Gaze Intersection Point (GIP) relative to the head center was then calculated, and distances between the GIP and the head were used to exclude unrealistic values and distances larger than 7 m. Based on consecutive GIPs in the three-dimensional space, the instantaneous visual angle θ at each position was computed. Following Duchowski (2017), a 2-tap velocity filter using a threshold of 130°/s was applied to the Pchip interpolated visual angle for the saccade detection, which then served as the basis for identifying fixations. Fixations shorter than 90 ms were excluded to reduce noise. From the remaining fixations, three features, the mean fixation duration, the total fixation duration, and the total fixation count, were extracted.

Statistical analysis

Despite the familiarization trial, learning effects were expected between the first and second runs. To quantify these effects, the differences between the first and second drills, regardless of the scenario, were analyzed. Therefore, in the first case, the drill order (first vs. second) and in the second case, the scenario type (Gamified vs. Non-Gamified) were the independent variables. Measured and extracted performance, questionnaire, HR variables, and selected ET metrics served as dependent variables. Normality was tested using the Shapiro-Wilk test (Shapiro & Wilk, 1965). For normally distributed data, a paired Student's t-test was applied; otherwise, a Wilcoxon signed-rank test was used. Unpaired tests were performed using the unpaired Student's t-tests for parametric distributions, and the Mann-Whitney *U* test otherwise. The significance level was set to α = .05 for all tests, and the reported effect sizes are Hedges' *g* since it corrects for small sample sizes (Turner & Bernard, 2006). A Bonferroni correction was applied in both cases within each modality and questionnaire family to account for multiple comparisons.

Classification pipeline

For the machine-learning-based comparison between the Gamified and Non-Gamified scenarios, we applied a classification approach in combination with explainable artificial intelligence (XAI). XAI methods help to make model decisions transparent by quantifying how individual features contribute to a specific prediction, allowing us to identify the most relevant features for distinguishing between the two scenarios.

In addition to the previously extracted features, additional statistical descriptors were derived based on prior work. For the ET-based metrics, 13 supplementary statistical features were included for the PD, following Stoeve et al. (2022). Blink- and fixation-related features were extended by their standard deviations where applicable (Kardan & Conati, 2012; Shojaeizadeh et al., 2019). For HR, six additional statistical features were extracted alongside the normalized mean (Hasnul et al., 2021). Performance data were extended by the standard deviation of the hit time to ensure consistency across modalities. Table 1 provides a detailed overview of all 41 extracted features,

with features exclusively used for inferential statistics marked with an asterisk.

The classification models were trained on different subsets of modalities (performance, HR + ET, HR + ET, performance + ET, and ET only) to assess the contribution of each modality combination. For each feature combination, all feature sets were additionally extracted using several window sizes: fixed windows of 10 s, 20 s, and 50 s (each computed with a 50% overlap), as well as a bisected version of the time series and the full time series without windowing. For the windowed feature sets, all modalities were temporally aligned and trimmed to the shortest available segment to ensure consistency across data streams.

Table 1
Detailed Overview of the Extracted Features of the Performance Metrics and Biosignals

Metric	Extracted features
Change of pupil diameter	Mean*, median, standard deviation, variance, maximum value, minimum value, range, first quantile, third quantile, interquartile range, skewness, kurtosis, and the number of samples until the maximum value
Blinks	Total number of blinks* Blink frequency* Blink duration mean* and standard deviation Blink interval duration mean* and standard deviation
Additional gaze features	Fixation duration mean* and standard deviation Total fixation duration* Total fixation count*
Additional PD based features	Index of pupillary activity (IPA)* Low/high IPA (LHIPA)* Slope of PD during first half-time*, timeout*, and second half-time*
HR	Normalized mean*, median, standard deviation, variance, maximum value, minimum value, range
Performance	Number scored goals* Streaks maximum* Streaks average* Hit times mean* and standard deviation

The metrics include statistical features based on the change of pupil diameter (PD), blinks, gaze, and additional PD metrics for eye tracking, the heart rate-related features, and performance-based measures. Metrics marked with an asterisk (*) are also used for the inferential statistics.

For the classification between the Gamified and Non-Gamified scenarios, we selected six ML models that have shown strong performance in prior work on ET data in the context of comparable cognitive or psychological processes such as stress or task demand (Lim et al., 2022; Novák et al., 2024; Stoeve et al., 2022). Based on the literature, we selected a representative set of algorithms covering different model families and levels of complexity, thus implementing and comparing logistic regression (LR), k-nearest neighbor (kNN), decision tree (DT), random forest (RF), support vector machine (SVM), and a fully connected neural network (NN).

Table 2
Overview of the Search Spaces of the Model Hyperparameters and Settings Optimized in the Cross-Validation

Classifier	Hyperparameter	Search space
LR	Regularization	$\in \{L_1, L_2\},\ C \in \{0.0001, \ldots, 100\}$
	Class weight	$\in \{\text{None}, \text{balanced}\}$
kNN	Number of neighbors	$\in \{5, \ldots, 100 \text{ or maximum number of samples}\}$
	Weights	$\in \{\text{uniform}, \text{distance}\}$
	Metric for distance computation	$\in \{\text{euclidean}, \text{manhattan}, \text{minkowski}, \text{cosine}\}$
DT	Maximum depth	$\in \{1, \ldots, 32\}$
	Minimum samples per split	$\in \{2, \ldots, 32\}$
	Minimum samples per leaf	$\in \{1, \ldots, 32\}$
	Criterion	$\in \{\text{gini}, \text{entropy}\}$
	Split strategy	$\in \{\text{best}, \text{random}\}$
	Number of features for split	$\in \{\text{sqrt}, \text{log2}\}$
	Class weight	$\in \{\text{None}, \text{balanced}\}$
RF	Number of estimators	$\in \{50, \ldots, 300\}$
	Maximum depth	$\in \{10, \ldots, 50\}$
	Minimum samples per split	$\in \{2, \ldots, 15\}$
	Minimum samples per leaf	$\in \{1, \ldots, 6\}$
	Criterion	$\in \{\text{gini}, \text{entropy}\}$
	Number of features for split	$\in \{\text{sqrt}, \text{log2}\}$
	Class weight	$\in \{\text{None}, \text{balanced\_subsample}, \text{balanced}\}$
	Bootstrapping	$\in \{\text{True}, \text{False}\}$
SVM	Kernel	$\in \{\text{linear}, \text{polynomial}, \text{rbf}, \text{sigmoid}\}$
	Cost parameter	$\in \{0.1, \ldots, 1000\}$
	Degree for polynomial kernel	$\in \{2, \ldots, 4\}$
	Gamma for linear kernel	$\in \{\text{scale}\}$
	Gamma for remaining kernels	$\in \{0.0001, \ldots, 100\}$
	Class weight	$\in \{\text{None}, \text{balanced}\}$
NN	Number of hidden layers	$N \in \{1, 2, 3\}$
	Number of output features	$N \in \{16, \ldots, 64\}$
	Batch normalization	$\in \{0, 1\}$
	Optimizer	$\in \{\text{Adam}, \text{RMSprop}, \text{SGD}\}$
	Activation function	$\in \{\text{ReLU}, \text{Tanh}, \text{Sigmoid}, \text{LeakyReLU}\}$
	Dropout probability	$\in \{0, \ldots, 0.5\}$
	Learning rate	$\in \{1\text{e}{-4}, \ldots, 1\text{e}{-1}\}$

Model training and hyperparameter optimization were conducted using Scikit-learn (Pedregosa et al., 2011) and Optuna (Akiba et al., 2019). To ensure unbiased performance estimation and prevent data leakage, a nested cross-validation (CV) scheme was applied. In the inner loop, a grouped five-fold group CV was used, ensuring that data from the same participant remained strictly within either the training or the test set for hyperparameter optimization and feature selection. The objective function for all models was to maximize the nested cross-validated macro F1-score, making sure each class is treated equally, independent of occurrence. Efficient search was facilitated by Optuna's Tree-structured Parzen Estimator (TPE). Each model was trained for 100 epochs with early stopping applied if the validation loss did not improve over fifteen consecutive epochs. In the outer CV loop, we applied leave-one-subject-out (LOSO- CV) for the performance evaluation of the optimized models. The reported evaluation metrics are the macro F1-score and accuracy, both d across all trials. Table 2 lists the search space for the optimized hyperparameters for each classifier.

The NN architecture was optimized for the number of hidden layers and neurons per layer, including optional batch normalization, dropout, and activation function for each layer. To ensure numerical stability, the output layer produced logits. The optimizer and learning rate were also part of the hyperparameter search.

In addition, the scaling strategy (StandardScaler, MinMaxScaler, RobustScaler, or none), as well as feature selection and dimensionality reduction techniques (SelectKBest, recursive feature elimination (RFE), principal component analysis (PCA), or none), were treated as hyperparameters and optimized for each model individually. Classifier-specific requirements were considered, for example, that the SVM requires a scaling method or only uses SelectKBest for the NN. To address the sensitivity of the distance-based KNN to high-dimensional feature spaces, we additionally added linear discriminant analysis (LDA) as a supervised dimensionality-reduction method, projecting the data into a subspace that maximizes class separability. The combination of window size and feature subset achieving the highest average F1-score across all subjects was selected for each model. The overall best-performing model was additionally used to investigate the relationship between classification performance and motivational questionnaire responses. Non-numerical hyperparameters were preselected using a majority vote across each optimal configuration and subsequently fixed for training and saving a final model using all available samples. To interpret the contribution and characteristics of individual features, we applied SHapley Additive exPlanations (SHAP). As a local, model-agnostic XAI method, SHAP provides insight into how individual features influence specific predictions while also enabling the derivation of consistent global feature importance (Lundberg & Lee, 2017). SHAP was selected over alternative XAI approaches because it allows for an assessment of global feature importance to characterize overall model behavior while providing stable and consistent explanations (Bekler et al., 2024; Hasan, 2023).

For each classifier, the corresponding SHAP explainer was used. When dimensionality reduction methods such as LDA and PCA were applied, SHAP values were computed in the transformed feature space and subsequently back-projected onto the original feature space to preserve interpretability. The selected hyperparameters and the features accounting for 90 % of the cumulative importance are reported.

Results

Questionnaires

According to the BRSQ, participants reported high levels of IM toward soccer before the drill (M = 6.40, SD = 0.67), particularly driven by the desire to experience the pleasurable sensations associated with the activity. EM (M = 1.85, SD = 0.94) and amotivation values were low (M = 1.55, SD = 0.81), indicating that all participants experienced interest and enjoyment in the activity itself.

Investigating the effects of the order of drills, a small effect in perceived exhaustion was measured. Participants experienced the second scenario, independent of the Gamification, as significantly more exhausting than the first (M = 11.62, SD = 1.65 vs. M = 12.44, SD = 1.91), $t(33) = -4.538$, $p < .001$, $g = -0.456$. No other effects regarding the order were found.

An error in the IMI implementation led to the absence of a question ("I thought this activity was quite enjoyable"). No correction or compensation was applied for the missing values; however, their absence may impact the results and should be considered in the interpretation. Overall, participants experienced high IM (M = 6.66, SD = 0.64) on the Interest/Enjoyment subscale. Competence was perceived as the strongest of the three basic needs (competence: M = 5.44, SD = 1.52; autonomy: M = 4.94, SD = 1.47; relatedness: M = 4.25, SD = 1.65). The responses to the UEQ showed that the drill itself was rated excellent in the categories of attractiveness, perspicuity, efficiency, stimulation, and novelty. Only dependability was rated as good.

In Table 3, the detailed questionnaire values and the results of the statistical differences between the scenarios with and without Gamification are listed. Bonferroni corrections were applied within each separate family of tests, given that they measure distinct constructs: perceived exertion (Borg scale, 1 item), motivational measures (IMI and PENS, 4 items in total), and user experience (UEQ, 6 items). No significant differences were observed. Both scenarios were perceived as equally exhausting, indicating a light exertion. IM was rated slightly higher in the gamified scenario, while perceived autonomy was slightly lower in the Gamified version. Changes in perceived autonomy and relatedness were small. UEQ results showed slightly higher attractiveness and dependability values in the gamified version, while being slightly less clear and understandable. No differences could be observed in Efficiency, Stimulation, and Novelty.

Table 3
Questionnaire Responses of the Borg-Scale, IMI, PENS, and UEQ, Including Test Statistics Comparing the Non-Gamified vs. Gamified Scenarios

Questionnaire	Non-Gamified		Gamified		W	$t(33)$	p_{corr}	Hedges' g
	M	SD	M	SD				
Borg Scale	12.06	1.77	12.00	1.89	-	0.255	0.800	0.032
IMI	6.61	0.74	6.70	0.54	57.5	-	1.519	-0.150
PENS								
Competence	5.41	1.54	5.46	1.52	87.5	-	2.097	-0.032
Autonomy	5.04	1.35	4.84	1.60	136.5	-	0.525	0.131
Relatedness	4.20	1.68	4.29	1.64	117.0	-	2.126	0.532
UEQ								
Attractiveness	2.24	0.83	2.29	0.61	125.5		2.943	-0.068
Perspicuity	2.19	0.71	2.13	0.56	133.0		3.801	0.094
Efficiency	1.98	0.76	1.97	0.61	222.5		5.062	0.022
Dependability	1.39	0.88	1.52	0.65	164.0		3.295	-0.167
Stimulation	2.29	0.80	2.32	0.634	138.5		4.494	-0.042
Novelty	1.69	0.85	1.70	0.72	-	-0.151	5.285	-0.019

The paired Student's t-test and the Wilcoxon signed-rank test were used for parametric and non-parametric distributions, respectively. Bonferroni correction was applied within each test family (Borg-Scale, IMI and PENS combined, and UEQ).

When asked after both drills, most participants liked the Gamified scenario more (Non-Gamified: 12, Gamified: 22). The preferred use case would be to improve their soccer skills, followed by using it as an exercise tool, and for pure enjoyment. The most liked gamification elements were the visual feedback, the noise feedback, and the leaderboard. However, Badges were not noticed by eight participants, the Team Leaderboard and the Streak Counter by five participants, respectively, while being experienced as neutral. None of the gamification elements was experienced as distracting.

Statistical comparison of biosignals

Regarding potential learning effects between the first and second scenarios, participants' performance dropped by an average of 0.41 goals (M = 52.32, SD = 2.25 vs. M = 51.91, SD = 2.86). The maximum streak count slightly decreased by 0.68 (M = 21.53, SD = 7.51 vs. M = 20.85, SD = 9.02), while the average streak count increased by 0.79 (M = 8.75, SD = 3.80 vs. M = 9.54, SD = 4.95). Differences in average scoring times were minimal (M = 2.67 s, SD = 0.19 vs. M = 2.53 s, SD = 0.22). Overall, no significant differences were observed.

Table 4 provides an overview of the performance metrics and their statistical comparison between the Gamified and Non-Gamified scenarios. While the number of scored goals and hit times remained similar across scenarios, both the average and maximum streak counts were higher in the Gamified scenario. All 34 participants earned the "Around the World" badge, while 31 achieved the "On Fire" badge.

Participants who first performed in the Non-Gamified scenario improved in the subsequent Gamified scenario by M = 3.17, SD = 12.58 maximum streak counts, and M = 2.04, SD = 7.10 average streak counts, but decreased their goal count by M = 1.11, SD = 3.45 goals. In contrast, those who started in the Gamified scenario improved by M = 0.38, SD = 2.22 goals but decreased by M = 5.00, SD = 7.54 in maximum streak count, and M = 0.62, SD = 3.49 in average streak count when switching to the Non-Gamified scenario. All statistical comparisons remained non-significant after the roni correction (four comparisons).

Table 4
Performance Results Including Test Statistics for the Non-Gamified vs. Gamified Scenario

Performance metrics	Non-Gamified		Gamified		*W*	*t*(33)	p_{corr}	Hedges' *g*
	M	*SD*	*M*	*SD*				
Scored goals	52.50	2.60	51.74	2.51	135.5	-	0.497	-0.296
Streaks maximum	19.18	7.00	23.21	8.98	-	2.259	0.123	0.495
Streaks average	8.46	3.79	9.83	4.90	241.0	-	1.353	0.309
Hit times (in s)	2.56	0.20	2.54	0.21		-1.450	0.625	-0.140

The paired Student's t-test and the Wilcoxon signed-rank test were used for parametric and non-parametric distributions, respectively. Bonferroni correction was set to 4.

ECG data from n = 32 participants were available and used to compute HR. The HR increased during the drill compared to the baseline and was significantly higher during the second scenario (M = 50.19%, SD = 27.11 vs. M = 61.50%, SD = 37.16), $t(31) = -2.066$, p = .047, g = 0.344. The increase in HR was on average higher in the Non-Gamified scenario (M = 61.03%, SD = 36.77) compared to the Gamified scenario (M = 50.66%, SD = 27.83). No significant differences were found. HRV features could not be calculated due to the heavy movement of the participants and the resulting noisy ECG signals. The extracted metrics were not within

the range of physiologically reasonable HRV metrics (Laborde et al., 2017).

Eye-tracking data from *n* = 29 participants were available, and after preprocessing, 23 participants were used to calculate the ET features. The detailed results are displayed in Table 5.

Table 5
Eye Tracking Results Including Test Statistics for the Non-Gamified vs. Gamified Scenario

ET metrics	Non-Gamified		Gamified		*W*	*t*(22)	p_{corr}	Hedges' *g*
	M	*SD*	*M*	*SD*				
Mean delta PD (mm)	0.77	0.79	0.60	0.73	100		3.373	0.220
Slope of PD (10^{-3} mm/s)								
1st half-time	2.05	2.58	1.74	1.92		0.639	6.883	0.135
2nd half-time	1.15	2.26	1.61	3.21	133		11.612	-0.164
Timeout	8.47	9.26	10.89	10.36	94		2.464	-0.242
IPA (10^{-3})	158.35	63.59	165.21	61.13		-0.714	6.276	-0.108
LHIPA (10^{-3})	15.69	7.86	13.74	5.67	111		5.556	0.279
Blinks								
Count	115.48	82.07	100.57	65.27	132		11.272	0.198
Frequency (Hz)	0.57	0.40	0.47	0.31	121		8.087	0.249
Mean duration (ms)	245.26	85.98	209.40	71.72		1.987	0.773	0.445
Mean interval duration (ms)	4122.71	7126.40	2947.49	2177.60	113		6.024	0.219
Fixations								
Mean duration (ms)	81.25	36.83	79.30	34.51	136		12.536	0.054
Total duration (s)	100.23	28.51	102.67	31.08		0.460	8.455	-0.080
Count	1295.48	166.46	1356.30	219.77	97		2.894	-0.307

Metrics were averaged across all participants. The paired Student's t-test and the Wilcoxon signed-rank test were used for parametric and non-parametric distributions, respectively. Bonferroni correction was set to 13.

Machine learning-based analysis

Table 6 provides an overview of the classification results comparing the Gamified and Non-Gamified scenarios. The kNN-classifier achieved the highest mean F1-score (*M* = 82.75%, *SD* = 20.42) and a mean accuracy (*M* = 83.70%, *SD* = 19.38), using the bisected time series and the ET feature subset only. The remaining classifiers achieved mean F1-scores between 54.34 % for the DT to 79.71 % for the SVM. Three classifiers performed best when using the bisected time series in combination with the ET features only. In contrast, the DT and RF achieved their best performance using the 10-second windows, i.e., with a larger amount of training data, while the NN performed best using features extracted from the full time series. Both the DT and NN additionally used performance features alongside ET data, whereas the RF was the only classifier achieving its best results when including ECG-based features.

Table 6
Mean and Standard Deviation of F1-Score and Accuracy in Percent of the Best-Performing Model and the Used Window Size in Samples

	LR	kNN	DT	RF	SVM	NN
F1-Score	74.78 ± 31.65	**82.75 ± 20.42**	54.34 ± 21.36	57.74 ± 22.54	79.71 ± 22.34	59.42 ± 37.13
Accuracy	77.17 ± 29.11	**83.70 ± 19.38**	59.41 ± 18.35	60.40 ± 20.53	82.61 ± 17.57	67.39 ± 32.00
Feature subset	only ET	only ET	Performance + ET	ECG + ET	only ET	Performance + ET
Window size	bisected time series	bisected time series	10 seconds	10 seconds	bisected time series	full time series

Values are presented as *M* ± *SD*. ET = eye tracking; ECG = electrocardiogram; LR = logistic regression; kNN = k-nearest neighbor; DT = decision tree; RF = random forest; SVM = support vector machine; NN = neural network.

The best results are highlighted in bold.

The confusion matrix of the best-performing model, the kNN classifier (Figure 3), indicates a balanced classification performance across the two classes. Based on the results of the outer cross-validation folds, corresponding to subject-wise performance, participants were divided into three groups. Eleven out of the 23 subjects were classified perfectly, achieving an F1-score of 100 % and forming Group 1. Two subjects achieved low F1-scores of 20 % and 50 %, respectively, and were assigned to Group 2. The remaining subjects were grouped into Group 3, with average F1-scores of 73 %.

To address the question of how the classifications relate to motivational questionnaire outcomes, subject-wise classification performance was compared with self-reported motivation for each group. Group 1 showed slightly lower overall motivation in the Gamified scenario; however, all three PENS dimensions were rated marginally higher compared to the Non-Gamified condition. In contrast, subjects in Group 2 reported higher overall motivation in the Gamified scenario, with an increase of approximately one scale point, while autonomy and relatedness were rated lower. Similarly, subjects in Group 3 indicated higher motivation in the Gamified scenario and increased relatedness, whereas competence and autonomy were rated slightly lower.

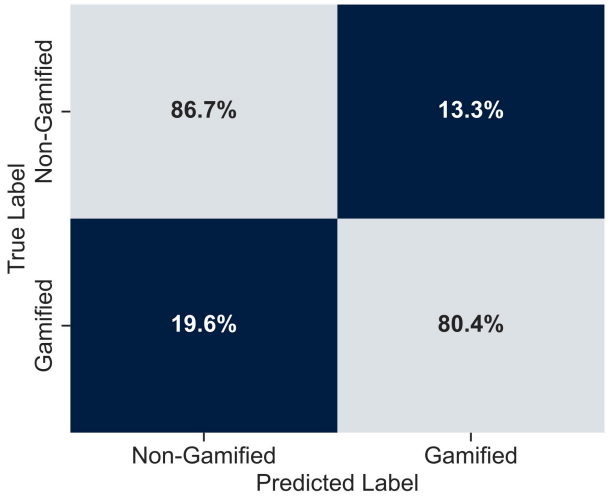

Figure 3 *Confusion Matrix of the Gamified vs. Non-Gamified Classification for the kNN Classifier Using the Bisected Time Series and The ET Data*

Table 7 displays the selected non-numerical hyperparameters and the subset of features contributing to 90 % cumulative importance for each final model. Across all classifiers, the most prominent features were related to ET metrics, specifically, blink-related measures and PD statistics appeared consistently as top contributors. For models employing dimensionality reduction, the primary components were heavily loaded with these metrics: the LDA component in kNN consisted mainly of blink and PD features, while the

first principal component (PC1) in the Decision Tree was defined by blink interval duration.

Table 7

The Selected Non-Numerical Hyperparameters for the Final Model, Including the Most Important Features Based on the SHAP Analysis, Are Listed for Each Classifier. The listed features account for 90% of the cumulative importance.

	Selected hyperparameters	Features
LR	No feature selection MinMaxScaler Penalty = L1 Class weight = balanced	Blink frequency mean Total number of blinks Delta PD median Delta PD 1st quantile
kNN	Feature selection LDA RobustScaler Uniform weights Metric for distance computation = minowski	LDA containing Total number of blinks Blink frequency mean Delta PD median Delta PD 3rd quantile Delta PD 1st quantile Delta PD mean Delta PD interquartile range
DT	Feature selection PCA No scaler Criterion = gini Number of features for split = sqrt Random splitter No class weight	PC1 containing Blink interval duration mean Blink interval duration standard deviation
RF	Feature selection RFE MinMaxScaler Criterion = gini Number of features for split = sqrt Class weight = balanced subsample Bootstraping True	Blink interval duration mean, standard deviation, Normalized HR mean, median, minimum value, maximum value Blink frequency mean Blink duration mean, standard deviation Delta PD median, mean, 3rd quantile, 1st quantile, maximum value, range, range, minimum value Total number of blinks Fixation duration standard deviation, mean, Total fixation duration
SVM	Feature selection RFE StandardScaler Linear kernel No class weight	Total number of blinks Blink frequency mean Delta PD mean Blink interval duration standard deviation
NN	No feature selection StandardScaler Optimizer RMSprop Number of layers = 1	Total fixation count, duration Delta PD number of samples until max, kurtosis, skewness LHIPA Blink duration mean, standard deviation Streaks maximum Slope of PD during timeout, first half-time Scorred goals Hit times mean, standard deviation Blink interval duration standard deviation Fixation duration standrad deviation

While linear and distance-based models relied primarily on eye-tracking data, the RF and NN models included the additional feature modalities. The RF model was the only classifier to select the ECG feature set within the top 90 % importance threshold. The NN utilized a distinct feature set, incorporating game performance metrics and higher-order pupil metrics alongside fixation and blink durations.

Figure 4 presents the SHAP summary plot for the best-performing kNN configuration. The analysis identified the total number of blinks (36.86 %) and the mean blink frequency (35,54 %) as the dominant predictors for the model predictions, while PD-related features contributed approximately 20 %. A visual inspection of the SHAP value distributions reveals distinct directional patterns. For blink frequency, lower feature values (represented by blue dots) are associated with negative SHAP values, thereby shifting the prediction toward the Gamified scenario. In contrast, the total number of blinks shows an inverse pattern, where lower values are partially associated with positive SHAP values, corresponding to predictions of the Non-Gamified class.

Discussion

The purpose of this work was to explore the effect of gamification on soccer players' situational motivation during a controlled, high-intensity passing drill performed in an immersive environment, and to assess whether motivation-related changes can be captured beyond traditional, state-of-the-art questionnaires using behavioral and physiological data from wearable sensors. To achieve this, we analyzed these data using a combination of inferential statistical methods and an exploratory ML approach.

To address the first purpose, namely to examine the effect of gamification on situational motivation using motivational questionnaires, questionnaire data were analyzed to compare the Gamified and Non-Gamified scenarios. Using these metrics, no significant differences were found between the two conditions. This was consistent across both the motivational measures (IMI) and the complementary questionnaires immediately completed after each drill. However, the absolute motivation scores were comparatively high and exceeded those reported in related studies using immersive sport environments (Cuthbert et al., 2019; Ijaz et al., 2020), indicating a generally high level of need satisfaction and IM caused by the implemented drill, independent of the experimental condition.

One possible explanation could be the novelty and complexity of the experimental setup. Participants reported feeling overwhelmed by the unfamiliar, immersive environment, the demanding task, and the short duration of the drill. Some participants reported either missing some gamification elements or not perceiving a clear difference between scenarios, which might lead to an underestimation of the potential effectiveness of the manipulation (Sailer et al., 2017). Others stated they only noticed them relatively late in the drill, mostly after the 15-second mid-drill break, which was intended to mitigate this issue.

This represents the key limitation of the present study and suggests that future studies should include an extended familiarization or tutorial phase to ensure that participants fully understand the task and the implemented gamification elements before data collection. In the current setup, participants' ability to detect differences during the drills may have been limited. A longer exposure phase may help reduce cognitive overload and allow the motivational effect to unfold more clearly. Additionally, this study investigated the overall effect of gamification rather than its specific components. Therefore, it was not possible to retrospectively exclude participants unaware of certain gamification elements. Future research should therefore include a more in-depth analysis of individual responses, accounting for element awareness and perception.

Despite the absence of significant questionnaire differences, a clear majority of participants retrospectively preferred the Gamified scenario, suggesting that a perceptible difference between conditions was present. Overall, participants reported high IMI scores, indicat-

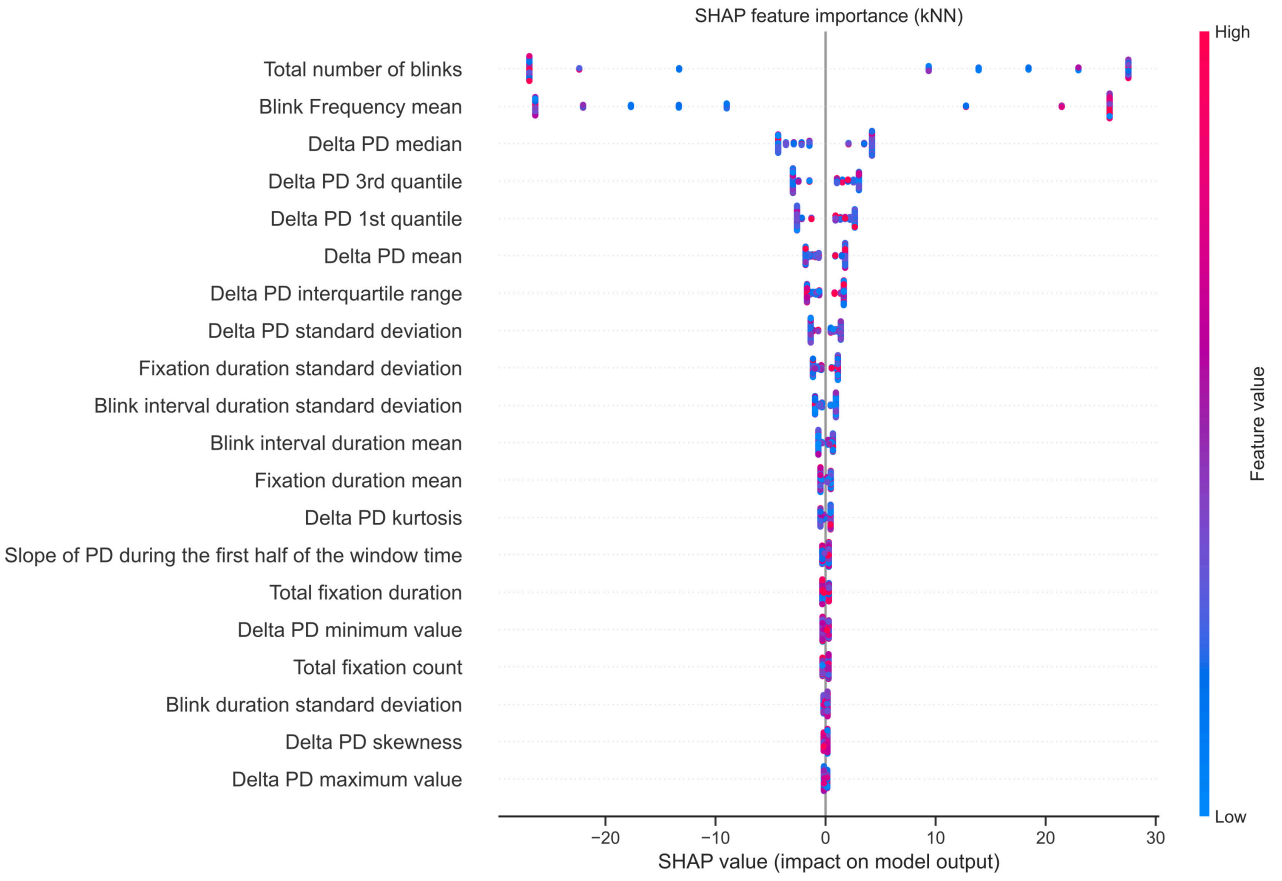

Figure 4 *SHAP Feature Importance Plot for the kNN Classifier (Gamified vs. Non-Gamified), Including the Top 20 Features Corresponding to 90 % Importance*

As the model utilized LDA for feature selection, SHAP values were back-projected from the latent LDA component space to the original features. Consequently, the importance of highly correlated metrics (e.g., Total number of blinks and Blink frequency mean) is distributed according to their loadings on the discriminant component. Each dot represents a single sample, with color indicating the feature value (red = high, blue = low). Positive SHAP values indicate a higher probability of the Non-Gamified condition, while negative values shift the prediction towards the Gamified condition.

ing a generally high level of IM during the drill. They expressed enthusiasm for the setup and game dynamics of the Igloo environment and interest in returning for training and enjoyment. However, this overall positive response should be interpreted with caution, as media research emphasizes the existence of a novelty effect, which can fade over time (Clark, 1983). As this study focused on immediate effects, a longitudinal investigation with repeated sessions is needed to assess the system's potential for sustained motivation, performance improvements, and behavioral change.

The second purpose was to identify relevant features from session recordings and wearable sensors and to determine whether performance metrics or biosignals differ between conditions using a mixed ML and statistical approach.

Performance metrics remained relatively stable across the two runs, independent of the order, indicating that

adaptation effects were negligibly small. When analyzing performance with respect to gamification, no significant differences were found. Nevertheless, a trend toward higher streaks in the Gamified scenario was observed, while the total number of scored goals was slightly higher in the Non-Gamified scenario. The order of scenarios further influenced performance patterns: When switching from the Non-Gamified to the Gamified scenario, streak counts increased while the number of scored goals decreased. In contrast, switching from the Gamified to the Non-Gamified scenario led to a pronounced decrease in streak counts and a slight increase in total goals. Additionally, stronger learning effects were observed when the Gamified scenario was presented second, indicating a subtle influence of the gamification. One possible explanation is the absence of a streak counter in the Non-Gamified scenario, with its introduction in the Gamified condition acting as an additional performance target. More generally, different gamification elements may encourage distinct behaviors and strategies rather than uniformly improving overall performance outcomes, depending on the players' subjective focus (Seaborn & Fels, 2015).

Analyzing the ECG signal, the HR increase relative to baseline was significantly higher in the second drill compared to the first scenario, suggesting that the break between scenarios may not have been sufficient for full physiological recovery. Across conditions, the mean HR increase compared to baseline was slightly lower in the Gamified condition. Assuming that a lower HR response is linked to cognitive and psychological processes, the players might have experienced lower mental stress (Taelman et al., 2009), reduced cognitive load (Solhjoo et al., 2019), or, in the context of game-based tasks, reduced perceived competitiveness (Cregan et al., 2025). However, the interpretation of parasympathetic activity is limited, as physiologically valid HRV metrics could not be reliably extracted due to the intense movement (Laborde et al., 2017). Furthermore, HR recovery was not analyzed, although it is closely related to autonomic regulation and emotional processes and could provide additional insights in future studies (Bunn et al., 2017).

Although no statistically significant differences were observed for the eye-tracking metrics, several descriptive trends were apparent. Specifically, fewer blinks and a lower blink frequency were reported in the Gamified scenario, accompanied by smaller pupil diameters. At first glance, these findings appear contradictory, as higher cognitive processing is associated with reduced blink rates but increased PD, reflecting the intention not to miss task-relevant information due to visual occlusion (Chen & Epps, 2014; Ledger, 2013). Although these effects did not reach statistical significance, they may indicate subtle differences in cognitive or visual attention processing between conditions and should be interpreted cautiously.

In contrast, the ML classifiers confirmed the discriminative capability of the recorded features for classifying the two scenarios. The kNN classifier achieved the highest macro F1-score of 82.75 %. The three best-performing models (kNN, SVM, LR) relied exclusively on ET data and reached F1-scores above 70 % F1-score, whereas the tree-based models (RF, DT) and the NN, which additionally incorporated performance or ECG data, achieved lower scores between 50% and 60%. Overall, our results are comparable to classification accuracies reported in related ET-based studies (Lim et al., 2022). These findings indicate that ET-derived features are particularly sensitive to scenario-related differences.

Across all classifiers, SHAP analysis consistently identified ET-derived features as the most influential contributors, with blink-related measures and PD statistics ranking highest. Specifically, a higher total number of blinks combined with a lower blink frequency shifted the kNN model's prediction toward the Gamified scenario, together accounting for 72.40 % of the cumulative feature importance. At first glance, this pattern appears physiologically contradictory. However, as shown in Table 5, the SHAP pattern for blink frequency accurately reflects the measured ground truth. The inverse pattern observed for the total number of blinks is therefore identified as a methodological artifact rather than a biological finding. This mirror effect is attributable to the LDA used for feature selection. In

the presence of high multicollinearity between frequency and count, the LDA algorithm assigns opposing weights to these redundant features to maximize the decision boundary margin (Hastie et al., 2009; Molnar, 2022). As a result, while both features contribute mathematically to the classification, blink frequency represents the physiologically interpretable predictor for the classification.

From a methodological perspective, this highlights a known limitation when interpreting SHAP values that are back-projected from reduced feature spaces such as LDA or PCA. When correlated features jointly contribute to a discriminant axis or principal component, SHAP distributes importance across these features, reflecting the underlying shared pattern rather than isolating a single independent sensor metric. Consequently, SHAP explanations in this context should be interpreted as global patterns across samples rather than as evidence for isolated causal feature effects (Bekler et al., 2024).

To address the third purpose, namely, analyzing how the extracted features relate to the motivational questionnaire results, subject-wise classification performance was aligned with the self-reported motivational measures. Interestingly, participants for whom the classifier failed to reliably distinguish the scenarios (i.e., low F1-scores) reported the largest motivational difference between the conditions, specifically, higher IM in the Gamified scenario. In contrast, a majority of participants who were classified perfectly reported no apparent motivational differences between scenarios based on the questionnaires. Overall, no consistent relationship between classification accuracy and reported IM or perceived need fulfillment was observed.

While no statistically significant differences were found in the questionnaire data alone, the combined consideration of performance measures and biosignals suggests scenario-related differences in behavior and perception associated with gamification. In particular, high classification accuracies demonstrate a clear separability of the recorded biosignals and behavioral features between the scenarios, indicating that gamification induced measurable changes beyond self-report. However, the mixed alignment between classification outcomes and questionnaire measures limits conclusions about the underlying psychological and physiological mechanisms. Specifically, the findings do not allow for a definitive mapping of gamification effects to specific motivational processes.

These findings suggest that gamification may induce subconscious or implicit effects that are not explicitly perceived by participants or accessible via self-reports. To better understand these dynamics, future research should more explicitly account for the individual nature of motivation by analyzing different response types more granularly. Further, individual differences such as skill level and experience, which are known to influence aspects such as gaze behavior (Krzepota et al., 2016), were not considered in this study. Additionally, the final sample size of n=23 limits the generalizability of the trained ML models. While the subject-wise cross-validation suggests robust discriminability within this cohort, validation on larger datasets is necessary to confirm the stability of the identified feature patterns across a broader population.

Conclusion

This work explored the feasibility of assessing situational motivation using session recordings and wearable sensors in combination with traditional questionnaire-based methods, in a gamified, immersive soccer-based drill. While no significant differences were observed in self-reported motivation or performance metrics, the robust ML classification performance, achieved primarily through ET-related features, indicates that gamification influenced players at a physiological level, reflecting heightened visual attention and engagement.

The observed dissociation between the clear separability of biosignals and the unaltered questionnaire response suggests that gamification effects may occur implicitly, becoming detectable through objective sensor data even when participants do not consciously perceive them. These findings highlight both the com-

plexity of measuring motivation and the value of integrating multivariate sensor data and machine learning approaches. Future research should explore how these physiological markers can be used in adaptive, immersive systems and motivational research, especially when dealing with complex and heterogeneous data, while emphasizing the need for individual approaches considering differences in perception, experience, and skill level.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Anshel, M. H., & Brinthaupt, T. M. (2014). Best practices for the use of inventories in sport psychology consulting. *Journal of Clinical Sport Psychology*, *8*(4), 400–420. https://doi.org/10.1123/jcsp.2014-0045

Bafna, T., Hansen, J. P. P., & Baekgaard, P. (2020). Cognitive load during eye-typing. *ACM Symposium on Eye Tracking Research and Applications, ETRA '20 Full Papers*, 1–8. https://doi.org/10.1145/3379155.3391333

Baltaci, S., & Gokcay, D. (2016). Stress detection in human–computer interaction: Fusion of pupil dilation and facial temperature features. *International Journal of Human–Computer Interaction*, *32*(12), 956–966. https://doi.org/10.1080/10447318.2016.1220069

Bekler, M., Yilmaz, M., & Ilgın, H. E. (2024). Assessing feature importance in eye-tracking data within virtual reality using explainable artificial intelligence techniques. *Applied Sciences*, *14*(14), 6042. https://doi.org/10.3390/app14146042

Bitrián, P., Buil, I., & Catalán, S. (2020). Gamification in sport apps: The determinants of users' motivation. *European Journal of Management and Business Economics*, *29*(3), 365–381. https://doi.org/10.1108/EJMBE-09-2019-0163

Borg, G. (1970). Perceived exertion as an indicator of somatic stress. *Journal of Rehabilitation Medicine*, *2*(2), 92–98. https://doi.org/10.2340/1650197719702239298

Bunn, J., Manor, J., Wells, E., Catanzarito, B., Kincer, B., & Eschbach, L. C. (2017). Physiological and emotional influence on heart rate recovery after submaximal exercise. *Journal of Human Sport and Exercise*, *12*(2), 349–357. https://doi.org/10.14198/jhse.2017.122.11

Chen, S., & Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, *29*(4), 390–413. https://doi.org/10.1080/07370024.2014.892428

Clancy, R. B., Herring, M. P., & Campbell, M. J. (2017). Motivation measures in sport: A critical review and bibliometric analysis. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00348

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research*, *53*(4), 445–459. https://doi.org/10.3102/00346543053004445

Cregan, S. C., Toth, A. J., & Campbell, M. J. (2025). Comparing the play of sport and action-adventure game genres on heart rate and heart rate variability. *Computers in Human Behavior Reports*, *17*, 100567. https://doi.org/10.1016/j.chbr.2024.100567

Cuthbert, R., Turkay, S., & Brown, R. (2019). The effects of customisation on player experiences and motivation in a virtual reality game. *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, 221–232. https://doi.org/10.1145/3369457.3369475

Dan, E. L., Dînşoreanu, M., & Mureşan, R. C. (2020). Accuracy of six interpolation methods applied on pupil diameter data. *2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*, 1–5. https://doi.org/10.1109/AQTR49680.2020.9129915

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, *19*(2), 109–134. https://doi.org/10.1016/0092-6566(85)90023-6

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification." *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, *MindTrek '11*, 9–15. https://doi.org/10.1145/2181037.2181040

Duchowski, A. T. (2017). *Eye tracking methodology: Theory and practice* (3rd ed.). Springer. https://doi.org/10.1007/978-3-319-57883-5

Duchowski, A. T., Krejtz, K., Gehrer, N. A., Bafna, T., & Baekgaard, P. (2020). The low/high index of pupillary activity. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, 1–12. https://doi.org/10.1145/3313831.3376394

Falkmer, T., Dahlman, J., Dukic, T., Bjällmark, A., & Larsson, M. (2008). Fixation identification in centroid versus start-point modes using eye-tracking data. *Perceptual and Motor Skills*, *106*(3), 710–724. https://doi.org/10.2466/pms.106.3.710-724

Francisco-Aparicio, A., Gutiérrez-Vela, F. L., Isla-Montes, J. L., & Sanchez, J. L. G. (2013). Gamification: Analysis and application. In V. M. R. Penichet, A. Peñalver, & J. A. Gallud (Eds.), *New trends in interaction, virtual reality and modeling* (pp. 113–126). Springer. https://doi.org/10.1007/978-1-4471-5445-7_9

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2022). Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, *13*(1), 440–460. https://doi.org/10.1109/TAFFC.2019.2927337

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education*, *80*, 152–161. https://doi.org/10.1016/j.compedu.2014.08.019

Hasan, M. M. (2023). Understanding model predictions: A comparative analysis of SHAP and LIME on various ML algorithms. *Journal of Scientific and Technological Research*, *5*(1), 17–26. https://doi.org/10.59738/jstr.v5i1.23(17-26).eaqr5800

Hasnul, M. A., Aziz, N. A. A., Alelyani, S., Mohana, M., & Aziz, A. A. (2021). Electrocardiogram-based emotion recognition systems and their applications in healthcare—A review. *Sensors*, *21*(15), 5015. https://doi.org/10.3390/s21155015

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Herlambang, M. B., Taatgen, N. A., & Cnossen, F. (2019). The role of motivation as a factor in mental fatigue. *Human Factors*, *61*(7), 1171–1185. https://doi.org/10.1177/0018720819828569

Ijaz, K., Ahmadpour, N., Wang, Y., & Calvo, R. A. (2020). Player experience of needs satisfaction (PENS) in an immersive virtual reality exercise platform describes motivation and enjoyment. *International Journal of Human–Computer Interaction*, *36*(13), 1195–1204. https://doi.org/10.1080/10447318.2020.1726107

Jerritta, S., Murugappan, M., Nagarajan, R., & Wan, K. (2011). Physiological signals based human emotion recognition: A review. *2011 IEEE 7th International Colloquium on Signal Processing and Its Applications*, 410–415. https://doi.org/10.1109/CSPA.2011.5759912

Kardan, S., & Conati, C. (2012). Exploring gaze data for determining user learning with an interactive simulation. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *User modeling, adaptation, and personalization* (Vol. 7379, pp. 126–138). Springer. https://doi.org/10.1007/978-3-642-31454-4_11

Korn, O., & Rees, A. (2019). Affective effects of gamification: Using biosignals to measure the effects on working and learning users. *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA '19*, 1–10. https://doi.org/10.1145/3316782.3316783

Kret, M. E., & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods*, *51*(3), 1336–1342. https://doi.org/10.3758/s13428-018-1075-y

Krzepota, J., Stępiński, M., & Zwierko, T. (2016). Gaze control in one versus one defensive situations in soccer players with various levels of expertise. *Perceptual and Motor Skills*, *123*(3), 769–783. https://doi.org/10.1177/0031512516664903

Laborde, S., Mosley, E., & Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research – Recommendations for experiment planning, data analysis, and data reporting. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00213

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In A. Holzinger (Ed.), *HCI and usability for education and work* (pp. 63–76). Springer. https://doi.org/10.1007/978-3-540-89350-9_6

Ledger, H. (2013). The effect cognitive load has on eye blinking. *The Plymouth Student Scientist*, *6*(1), 206–223. https://doi.org/10.24382/1bmd-pq94

Lim, J. Z., Mountstephens, J., & Teo, J. (2022). Eye-tracking feature extraction for biometric machine learning. *Frontiers in Neurorobotics*, *15*. https://doi.org/10.3389/fnbot.2021.796895

Lonsdale, C., Hodge, K., & Rose, E. A. (2008). The behavioral regulation in sport questionnaire (BRSQ): Instrument development and initial validity evidence. *Journal of Sport and Exercise Psychology*, *30*(3), 323–355. https://doi.org/10.1123/jsep.30.3.323

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *NIPS'17, 30*, 4768–4777.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. A. (2021). NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, *53*(4), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, *50*(1), 94–106. https://doi.org/10.3758/s13428-017-1007-2

Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, *71*, 525–534. https://doi.org/10.1016/j.chb.2015.08.048

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (Second). Christoph Molnar.

Mutlu-Bayraktar, D., Cosgun, V., & Altan, T. (2019). Cognitive load in multimedia learning environments: A systematic review. *Computers & Education*, *141*, 103618. https://doi.org/10.1016/j.compedu.2019.103618

Novák, J. Š., Masner, J., Benda, P., Šimek, P., & Merunka, V. (2024). Eye tracking, usability, and user experience: A systematic review. *International Journal of Human–Computer Interaction*, *40*(17), 4484–4500. https://doi.org/10.1080/10447318.2023.2221600

Nyström, M., Andersson, R., Niehorster, D. C., Hessels, R. S., & Hooge, I. T. C. (2024). What is a blink? Classifying and characterizing blinks in eye openness signals. *Behavior Research Methods*, *56*(4), 3280–3299. https://doi.org/10.3758/s13428-023-02333-9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res*, *12*, 2825–2830.

Richer, R., Küderle, A., Ullrich, M., Rohleder, N., & Eskofier, B. (2021). BioPsyKit: A python package for the analysis of biopsychological data. *Journal of Open Source Software*, *6*(66), 3702. https://doi.org/10.21105/joss.03702

Roberts, G. C., & Treasure, D. C. (Eds.). (2012). *Advances in motivation in sport and exercise* (3rd ed.). Human Kinetics.

Ryan, R. M., Bradshaw, E. L., & Deci, E. L. (2019). Motivation. In R. J. Sternberg & W. E. Pickren (Eds.), *The Cambridge handbook of the intellectual history of psychology* (pp. 391–411). Cambridge University Press. https://doi.org/10.1017/9781108290876.016

Ryan, R. M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology*, *45*(4), 736–750. https://doi.org/10.1037/0022-3514.45.4.736

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*(4), 344–360. https://doi.org/10.1007/s11031-006-9051-8

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*, 371–380. https://doi.org/10.1016/j.chb.2016.12.033

Sajno, E., Bartolotta, S., Tuena, C., Cipresso, P., Pedroli, E., & Riva, G. (2023). Machine learning in biosignals processing for mental health: A narrative review. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1066317

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, *74*, 14–31. https://doi.org/10.1016/j.ijhcs.2014.09.006

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591. https://doi.org/10.2307/2333709

Shojaeizadeh, M., Djamasbi, S., Paffenroth, R. C., & Trapp, A. C. (2019). Detecting task demand via an eye tracking machine learning system. *Decision Support Systems*, *116*, 91–101. https://doi.org/10.1016/j.dss.2018.10.012

Solhjoo, S., Haigney, M. C., McBee, E., van Merrienboer, J. J. G., Schuwirth, L., Artino, A. R., Battista, A., Ratcliffe, T. A., Lee, H. D., & Durning, S. J. (2019). Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Scientific Reports*, *9*(1), 14668. https://doi.org/10.1038/s41598-019-50280-3

Sotos-Martínez, V. J., Ferriz-Valero, A., García-Martínez, S., & Tortosa-Martínez, J. (2024). The effects of gamification on the motivation and basic psychological needs of secondary school physical education students. *Physical Education and Sport Pedagogy*, *29*(2), 160–176. https://doi.org/10.1080/17408989.2022.2039611

Stoeve, M., Wirth, M., Farlock, R., Antunovic, A., Müller, V., & Eskofier, B. M. (2022). Eye tracking-based stress classification of athletes in virtual reality. *Proc. ACM Comput. Graph. Interact. Tech*, *5*(2). https://doi.org/10.1145/3530796

Taelman, J., Vandeput, S., Spaepen, A., & Van Huffel, S. (2009). Influence of mental stress on heart rate and heart rate variability. In J. Vander Sloten, P. Verdonck, M. Nyssen, & J. Haueisen (Eds.), *4th European conference of the International Federation for Medical and Biological Engineering* (pp. 1366–1369). https://doi.org/10.1007/978-3-540-89208-3_324

Teixeira, P. J., Carraça, E. V., Markland, D., Silva, M. N., & Ryan, R. M. (2012). Exercise, physical activity, and self-determination theory: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *9*(1), 78. https://doi.org/10.1186/1479-5868-9-78

Touré-Tillery, M., & Fishbach, A. (2014). How to measure motivation: A guide for the experimental social psychologist. *Social and Personality Psychology Compass*, *8*(7), 328–341. https://doi.org/10.1111/spc3.12110

Turner, H. M., III, & Bernard, R. M. (2006). Calculating and synthesizing effect sizes. *Contemporary Issues in Communication Science and Disorders*, *33*(Spring), 42–55. https://doi.org/10.1044/cicsd_33_S_42

Vallerand, R. J. (2007). Intrinsic and extrinsic motivation in sport and physical activity: A review and a look at the future. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (1st ed., pp. 59–83). Wiley. https://doi.org/10.1002/9781118270011.ch3

Vallerand, R. J., & Losier, G. F. (1999). An integrative analysis of intrinsic and extrinsic motivation in sport. *Journal of Applied Sport Psychology*, *11*(1), 142–169. https://doi.org/10.1080/10413209908402956

Vealey, R. S., Cooley, R., Nilsson, E., Block, C., & Galli, N. (2019). Assessment and the use of questionnaires in sport psychology consulting: An analysis of practices and attitudes from 2003 to 2017. *Journal of Clinical Sport Psychology*, *13*(4), 505–523. https://doi.org/10.1123/jcsp.2019-0012

Vorberg, L., Pflueger, S., Richer, R., Jaeger, K. M., Küderle, A., Rohleder, N., & Eskofier, B. M. (2023). Prediction of stress coping capabilities from nightly heart rate patterns using machine learning. *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 1–4. https://doi.org/10.1109/BHI58575.2023.10313401

Xi, N., & Hamari, J. (2019). Does gamification satisfy needs? A study on the relationship between gamification features and intrinsic need satisfaction. *International Journal of Information Management*, *46*, 210–221. https://doi.org/10.1016/j.ijinfomgt.2018.12.002

Acknowledgements

During the preparation of this work, AI technologies were used to assist in the writing process. Specifically, Grammarly (Grammarly, Inc., San Francisco, CA, USA) was used to check for grammar and style consistency, and ChatGPT (GPT-4) (OpenAI, San Francisco, CA, USA) was used to assist with rephrasing and improving readability. After using these tools, the manuscript was carefully reviewed, and the content was edited as needed. No tools or services were used for content generation. The content of this work was entirely created by the authors, who ensured that all material presented reflected their original work and research. The authors take full responsibility for the content of the publication.

Funding

This work was conducted in the scope of the BISS research project supported by VDI/VDE-IT.

Competing interests

The authors have declared that no competing interests exist.

Data availability statement

All relevant data are within the paper.

Authors' contributions

Conceptualization: RL, MS, NK, KA and ED; Data curation: RL, MS, NK and KA; Formal analysis: RL; Funding acquisition: MW and BE; Investigation: RL, MS, NK, KA and LW; Methodology: RL, NK, KA and ED; Software: RL, MS and LW; Supervision: MW, BE and ED; Visualization: RL; Writing - Original Draft: RL; Writing - Review & Editing: NK, KA, MW, BE and ED. All authors read and approved the final manuscript.