Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/OJIM.2022.1234567

Real-Time Player Engagement Measurement Using Non-Intrusive Game Telemetry

AMMAR RASHED *, SHERVIN SHIRMOHAMMADI * (FELLOW, IEEE), AND MOHAMED HEFEEDA † (SENIOR MEMBER, IEEE)

[1]University of Ottawa, Ottawa, ON, K1N 6N5 Canada
[2]Simon Fraser University, Burnaby, BC, V5A 1S6 Canada

CORRESPONDING AUTHOR: Ammar Rashed (e-mail: ammarrashed54@gmail.com).

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada grant # ALLRP556311-20.

ABSTRACT Player engagement is crucial for the success of modern video games, yet its real-time measurement remains challenging due to the intrusive nature of traditional measurement methods. In this paper, we present a novel framework for non-intrusive, real-time and indirect measurement of engagement in multiplayer online games based on flow theory. Our approach combines Graph Convolutional Networks for modeling player interactions with Transformer networks for temporal processing, enabling indirect measurement of both player skill and game challenge, which in turn are used to classify player engagement. Using PlayerUnknown's Battlegrounds (PUBG) as a case study, we demonstrate that our framework can effectively measure phase-specific engagement using one minute of gameplay telemetry data. Our framework achieves 73% accuracy and 0.83 ROC-AUC in engagement classification, matching the performance of traditional survey-based methods while operating non-intrusively and in real-time. Further cross-domain validation of the frameowrk, as is and without transfer learning, with the games FIFA'23 and Street Fighter V, leads to 66% accuracy, demonstrating the model's stable performance despite the significant differences in the test domains. Interestingly, our results suggest that objective gameplay metrics may better reflect engagement than subjective player assessments, with skill estimates showing significant correlation with self-reports.

INDEX TERMS Engagement Measurement, Machine Learning-Assisted Measurement, Flow Theory, Game Telemetry

I. INTRODUCTION

THE video gaming industry has experienced exponential growth, now generating more revenue than the music and movie industries combined [1]. As games become increasingly complex and extensive, measuring and monitoring player engagement has become crucial for game developers and researchers alike [2]. Player engagement is a multidimensional construct that encompasses cognitive, emotional, and behavioral aspects of gameplay [3], requiring sophisticated instrumentation methods to capture its various dimensions [4], [5]. The importance of player engagement spans various domains, including entertainment, education [6], and business aspects such as churn prediction [7], [8] and game design improvements [9]–[11].

Precise measurement of player engagement in real-time faces multiple measurement challenges. First, player engagement is a complex, multi-dimensional construct requiring simultaneous monitoring of cognitive, behavioral, and emotional signals. Second, engagement manifests both as an instantaneous measurable state and as an evolving process over time, complicating the development of unified measurement models. Third, players' diverse preferences and gaming experiences necessitate adaptive measurement approaches across different game contexts.

Traditional methods for measuring player engagement rely on two main approaches: post-game surveys and physiological data collection [12], [13]. Post-game surveys, such as the Game Experience Questionnaire (GEQ) [14], provide

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

comprehensive measurement data and are straightforward to implement. However, they suffer from recall bias due to the time gap between gameplay and reporting [15], and cannot capture temporal fluctuations in engagement signals [16]. Conversely, physiological measurements, such as heart rate monitoring and eye tracking, offer continuous, objective signal collection during gameplay. While this approach provides high-resolution temporal data, it requires specialized sensing equipment, creates artificial measurement conditions, and typically involves complex signal processing that prevents real-time engagement detection.

To address these limitations, we propose a non-intrusive framework for real-time measurement of player engagement using game telemetry signals. Modern games routinely collect comprehensive telemetry data about player actions, performance metrics, and game states, providing an accessible and scalable measurement source. Our measurement approach is grounded in Flow Theory [17], which posits that optimal engagement occurs when a player's skill matches the game's challenge. But skill and challenge are also difficult to measure directly and non-invasively during gameplay.

To address this problem, we propose using easier-to-measure telemetry signals including combat statistics, movement patterns, resource management, and general match states, to then indirectly measure skill and challenge. To do so, we use Graph Convolutional Networks (GCN) that detect complex player interactions and spatial relationships, coupled with Transformer networks that process temporal sequences of game states. This hybrid architecture produces two proxy metrics that quantify skill and challenge, as described next.

The first proxy metric is the player's ranking in the match, which indicates their skill level. Ranking is usually determined at match completion, but our framework detects likely match outcomes in real-time based on ongoing performance signals. Higher ranking indicates the player outperforming others, suggesting higher skill levels. We consider skill to be a continuous ordinal quantity normalized between 0 (lowest skill) and 1 (highest skill).

The second proxy metric is the total damage sustained from both enemy attacks and environmental hazards per game phase, which quantifies the challenge level. Higher damage indicates relatively more difficult game conditions during that phase. We consider challenge to be a continuous, positive, and unbounded ordinal quantity.

Finally, the measured skill and challenge are fed to an engagement measurement module - a binary classifier which uses an established baseline from player survey data to classify engagement as an ordinal quantity of either 0 (low engagement) or 1 (high engagement). The entire measurement pipeline operates in real-time, providing engagement estimates that precede actual gameplay outcomes by variable time intervals, depending on when the player is eliminated or the phase ends.

The instrumentation and measurement literature emphasizes the importance of user engagement mainly in medical measurement applications, such as recognizing the emotional dimension of engagement in rehabilitation [18], measuring engagement using a proprietary tool from Emotiv Inc. to evaluate the performance of an ADHD detection system [19], utilizing cognitive engagement for risk assessment in the use of medical devices [20], and predicting engagement in older adults with dementia [21]. An exception is [22], which uses physiological measures to assess driver engagement. While [19] and [21] also use games, none of the above works use game telemetry data to measure engagement in real time, which is a main novelty of our proposed framework. The primary contributions of our framework can therefore be summarized as:

- An instrumentation methodology for real-time measurement of engagement based on flow theory, transforming standard game telemetry signals into continuous skill and challenge measurements without gameplay interruption.
- A hybrid signal processing architecture combining GCN for player interactions with Transformers for temporal sequences, demonstrating superior measurement performance over single-architecture alternatives.
- A high-resolution engagement measurement methodology using survey-calibrated baselines, enabling detection of significant variations within inherently engaging game contexts.
- A practical realization of the measurement framework in PUBG, demonstrating its viability for complex multiplayer environments with diverse gameplay mechanics and player interactions.
- Cross-domain validation of the model, as is and without transfer learning, with FIFA'23, a sports game, and Street Fighter V, a fighting game, demonstrating that the approach is genre-agnostic, applicable to a wide variety of game types beyond combat-focused games, including sports games, racing games, strategy games, and more.

The real-time engagement metrics provided by our framework offer several practical applications for game developers. First, they enable dynamic difficulty adjustment, where the game can automatically modify challenge levels based on detected engagement states, preventing player frustration or boredom [23], [24]. Second, they facilitate targeted content delivery, allowing developers to introduce new gameplay elements or narrative sequences precisely when engagement begins to decline [25]. Third, they support personalized matchmaking systems that can maintain optimal skill-challenge balances across different player segments [26], [27]. Fourth, they enable intelligent resource allocation in cloud gaming environments, where streaming quality and latency significantly impact user engagement [28]. By identifying highly engaging gameplay moments, cloud gaming

providers can dynamically allocate more bandwidth and processing resources to maintain quality during critical periods, similar to how video providers optimize streaming quality to maximize user engagement [29]. Beyond individual player optimization, our metrics can reveal engagement patterns across different game features, scenarios, play sessions, and platforms, enabling developers to identify which specific game elements consistently drive or diminish engagement. The measurement approach we propose provides several key features: temporal granularity (detecting engagement fluctuations within individual matches), contextual awareness (understanding engagement in relation to specific gameplay contexts), scalability (processing data from thousands of concurrent players), and actionability (producing metrics that directly inform design decisions). Unlike post-hoc analysis, these real-time capabilities enable proactive interventions that can significantly impact player retention and satisfaction.

The rest of this paper is organized as follows. Section II covers the background science on engagement and Flow Theory, and also looks at the related work in player engagement measurement, while Section III describes the proposed framework and problem formulation. In Section IV we present the proposed solution, while in Section V we discuss the PUBG case study and analyze the performance evaluation results. The paper is concluded in Section VI.

II. BACKGROUND & RELATED WORK

A. Background

1) Player Engagement

Player engagement is a multidimensional construct encompassing cognitive, emotional, and behavioral aspects of gameplay [3]. It extends beyond mere interaction to include the player's immersion, motivation, and overall satisfaction with the gaming experience [4], [5]. It can be understood as a mosaic of complementary aspects [30], as illustrated in Figure 1. The engagement process typically begins with motivation, which can be intrinsic (e.g., curiosity) or extrinsic (e.g., social pressure) [31], [32]. As players become involved with the game, they experience immersion - a state of cognitive absorption characterized by audiovisual stimulation, deep concentration, and spatio-temporal distortion. This immersion can lead to engrossment, marked by emotional attachment and a sense of presence in the game [33]–[35].

The pinnacle of engagement is often described as flow, a state of optimal experience where players' skills are well-matched with the game's challenges [36], [37]. Flow is characterized by clear goals, balanced challenge-skill ratio, and immediate feedback. This engaging experience can extend beyond a single session through the concept of endurability [38], [39], where positive experiences reinforce the desire to play again. Figure 1 illustrates these interconnected aspects of engagement as a continuous process. As Flow Theory is an important aspect of our design, we take a more detailed look at it in the next subsection.

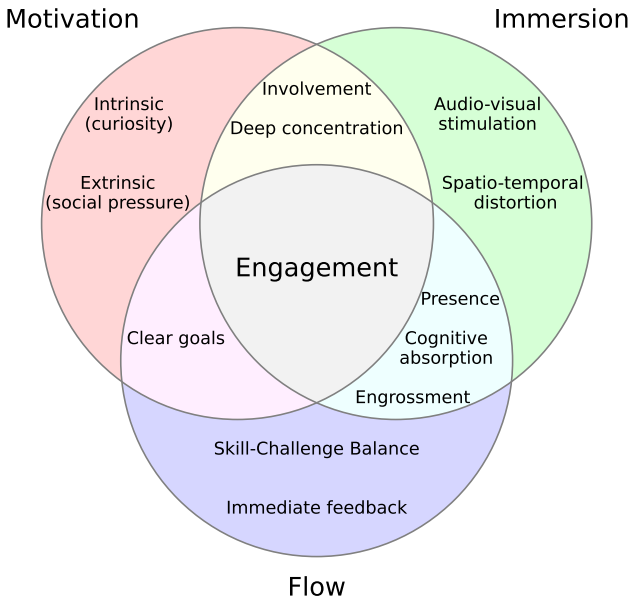


FIGURE 1: Conceptual framework of player engagement components and their interactions.

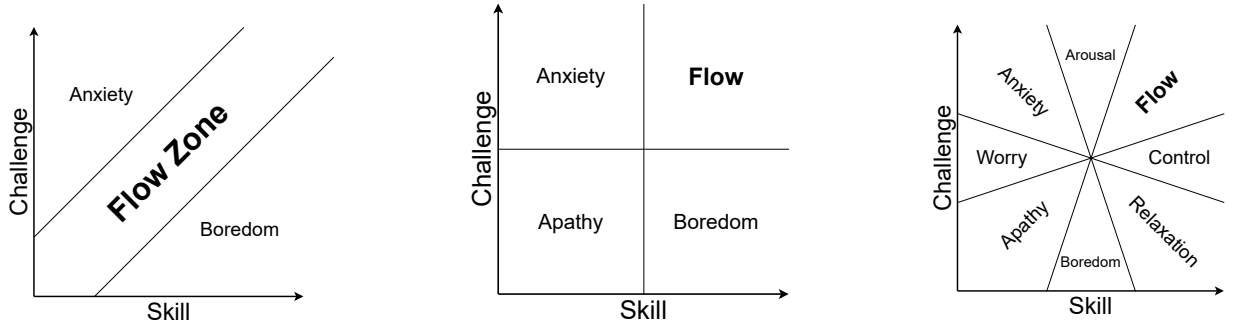
2) Flow Theory

Flow Theory, introduced by Csikszentmihalyi [17], states that optimal engagement occurs when a player's skill level is well-matched with the game's challenge. This concept has been widely adopted in game design and engagement studies [40]. Several models have been developed to apply Flow Theory in gaming contexts, including the Quadrant Model [41], the Experience Fluctuation Model (EFM) [42], and the Flow Channel Model. These models, illustrated in Figure 2, provide different perspectives on the relationship between skill, challenge, and engagement.

While each of the above models offers valuable insights, they have limitations for real-time engagement estimation in complex gaming environments. Our work draws inspiration from these existing models to develop a novel approach for engagement measurement. We focus particularly on the relationship between skill and challenge as key determinants of player engagement, aligning with the core principles of Flow Theory. By combining this theoretical foundation with machine learning-assisted measurement techniques and empirical data, our approach aims to capture the nuances of engagement in dynamic multiplayer online games while remaining computationally feasible for real-time estimation.

3) Engagement Across Gaming Scenarios

Player engagement manifests differently across various gaming scenarios, each requiring appropriate measurement considerations. In single-player games, engagement primarily derives from narrative immersion, progression systems, and the balance between preset challenges and player skill [43].



(a) Flow-channel model (b) Quadrant model (c) Experience fluctuation model

FIGURE 2: Conceptual models relating skill and challenge in Flow Theory.

Local multiplayer introduces social dimensions where engagement features include interpersonal dynamics and immediate social feedback [44]. Online multiplayer presents the most complex scenario, combining individual performance metrics with social engagement indicators [45].

Our framework, while validated primarily in online multiplayer environments, is designed with sufficient flexibility to address these varied contexts. In PUBG specifically, our approach captures engagement across multiple interactive dimensions: player-versus-player (competitive team combat), player-versus-environment (survival against shrinking play zones), and cooperative team dynamics (squad coordination). The telemetry signals we leverage, such as combat statistics, movement patterns, and resource management, have analogues in most gaming contexts—combat performance in single-player games, collaborative actions in local multiplayer, or competitive metrics in online play. The underlying Flow Theory principles regarding skill-challenge balance remain applicable across all gaming modalities, though the specific telemetry sources and proxies would need appropriate adaptation. Our emphasis on non-intrusive measurement makes the framework particularly valuable for online multiplayer environments where direct observation is impractical and interruption-based measures disrupt the experience.

B. Related Work

1) Skill and Challenge Estimation

Measuring player skill and game challenge is crucial for understanding player engagement. Previous works have explored various approaches to quantify skill and challenge. For instance, Aponte et al. [46] used reinforcement learning to train virtual agents and measure challenge based on the agents' pass rates. Wheat et al. [47] analyzed level characteristics in 2D games to model challenge. For skill estimation, Diah et al. [48] used heuristics like the number of enemies defeated in MOBA games. While these approaches provide valuable insights, they often lack generalizability and struggle to capture the dynamic nature of multiplayer online games. Our telemetry-based framework addresses these limitations by measuring skill through relative competitive performance and challenge through immediate survival threats. This generalizable approach enables real-time measurement across various competitive games, as we will demonstrate in section V with an actual use case in PUBG.

2) Engagement Estimation

Recent advancements in engagement estimation have explored various methodologies. Chen et al. [49] introduced facial expression-based models for non-intrusive engagement estimation. Fortin et al. [50] developed models using physiological measures and game events. However, these approaches often face limitations in real-world applications. Facial expression-based methods, for instance, require webcam footage, which is not always available or practical in gaming environments. Our work overcomes these limitations by focusing exclusively on game telemetry data, which is readily available and non-intrusive.

3) Telemetry Data Analysis in Games

Game telemetry data has been increasingly used to understand player behavior and experience. Melhart et al. [51] used in-game events to model player experience, while Reguera et al. [52] explored using gameplay session data as a proxy for engagement. A notable study by Melhart et al. [53] demonstrated the power of gameplay features in predicting engagement, albeit from a different perspective. Their work on PUBG used telemetry data to predict moment-to-moment viewer engagement on Twitch streams. By analyzing the relationship between in-game events and viewer chat frequency, they achieved prediction accuracies of up to 80% on average. This study underscores the potential of telemetry data as a powerful predictor of engagement. But not all games will have viewers or chats available, so our work extends telemetry data usage by deriving quantitative

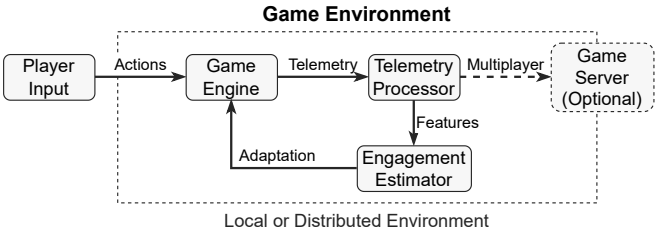


FIGURE 3: The considered model for gaming systems.

skill and challenge measures from gameplay data, mapping these to engagement states via Flow Theory.

4) Ground Truth Establishment

Establishing reliable ground truth for engagement estimation remains a significant challenge in the field, particularly in the context of multiplayer online games where the difficulty dynamically changes based on competing opponents. Various methods have been employed in previous studies, including self-report questionnaires [54], continuous annotations [55], observational methods [56], and proxy measures such as conation (the desire to continue playing) [57], [58]. Each of these methods has its strengths and limitations, often facing issues such as recall bias in post-game surveys [15] or potential disruption of gameplay in real-time reporting.

Our work addresses these challenges through a hierarchical validation framework that systematically evaluates each component of the engagement estimation pipeline. We first establish an empirical upper bound by analyzing self-reported skill-challenge-engagement relationships. Then, we validate our telemetry-based proxies against these self-reports to ensure alignment with player perceptions. Finally, we evaluate our real-time estimation system end-to-end by comparing its predictions against post-game survey responses. This enables systematic validation from raw telemetry data to final engagement predictions.

III. SYSTEM MODEL & PROBLEM DEFINITION

In this section, we specify the considered model for gaming systems and formally define the player engagement problem.

A. System Model

Our engagement estimation framework operates within a general gaming environment, as illustrated in Figure 3. The Game Server authenticates players and matches them in games. Players send actions to the Game Engine, which renders the game and maintains the game state. The Game Engine also logs telemetry events containing details about shots fired, weapons used, locations of the attacker and victim, etc. The Telemetry Processor aggregates these raw events into meaningful features. Consider, for example, a combat scenario where a player attacks an opponent. The Telemetry Processor converts combat events into metrics like damage dealt and accuracy rates and movement events into distance traveled. It also converts inventory events into resource utilization patterns. These processed features capture both player skill (through combat performance and resource management) and challenge levels (through damage taken and threat proximity). Finally, the Engagement Estimator analyzes the high-level telemetry features to measure player engagement. It can be integrated directly within the Game Engine for immediate state updates or implemented as an external module when handling complex multiplayer scenarios requiring additional processing capacity.

To illustrate with a practical example from our PUBG case study: when a player engages in combat, the Game Engine logs raw events such as ”Player A fired a weapon,” ”Player A hit Player B,” and ”Player B took X damage.” The Telemetry Processor aggregates these into meaningful features including accuracy (hits/shots), damage per minute, and combat efficiency (damage dealt/damage taken). Simultaneously, movement events like ”Player A moved to position (x,y,z)” are transformed into metrics such as distance traveled, rotation frequency, and positioning relative to safe zones. Consider a specific in-game scenario where a player engages an opponent at medium range using an assault rifle. The Telemetry Processor would capture combat performance (e.g., 60% accuracy, 90 damage dealt), positioning context (e.g., partial cover utilization, high ground advantage), resource management (e.g., ammunition consumption rate, healing item usage), and threat assessment (e.g., proximity to other teams, position relative to play zone boundary). These processed features would then feed into the Engagement Estimator to determine the player’s current skill expression and challenge level during this combat interaction.

This flexible architecture supports various deployment scenarios, from single-player games with local processing to distributed multiplayer environments. In online multiplayer games, clients connect to game servers that aggregate player interactions and state updates, allowing the engagement estimator to operate at the server level to account for inter-player dynamics. The system assumes reliable telemetry data collection and low-latency processing capabilities to enable real-time engagement estimation.

To capture meaningful interactions in games, we represent the game state at time t as a dynamic graph $G(t) = (V(t), A(t))$, where $A(t)$ represents the adjacency structure. The vertices $V(t)$ represent game entities (players, non-player or AI characters, interactive objects, or environmental elements), while the adjacency structure encodes relevant relationships between these entities. These relationships can be defined flexibly based on game-specific interaction metrics such as spatial proximity, direct interaction, strategic relevance, or causal relationships.

For each vertex $v \in V(t)$, we maintain a feature vector $\mathbf{x}_v(t)$ that captures its current state. This graph representation is versatile and can model various game scenarios: competitive or cooperative interactions between human players,

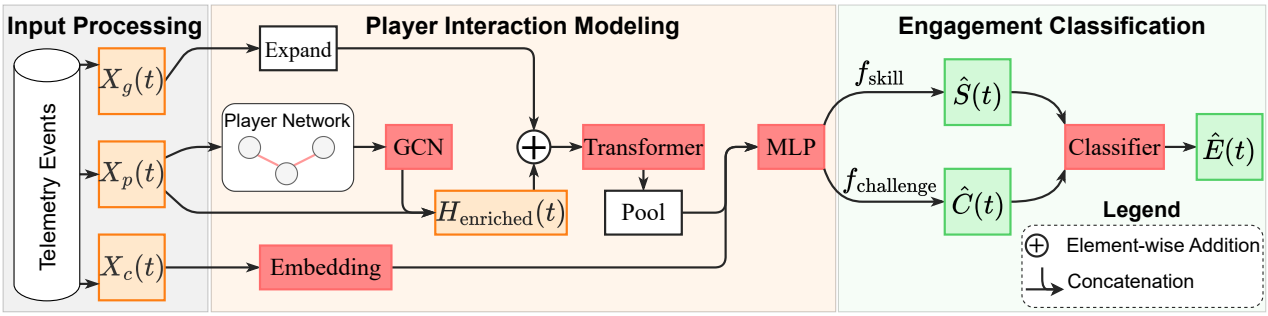


FIGURE 4: Overview of the proposed engagement estimation framework. Game telemetry data flows through the GCN and Transformer networks to predict the skill and challenge metrics, which are used to estimate engagement $\hat{E}(t)$.

interactions with AI-controlled enemies or environmental challenges, relationships between team members, or interactions between a player and game-generated entities in single-player games.

This system model is adaptable to various gaming scenarios. In single-player games, the graph representation simplifies as interactions occur primarily between the player and game-generated entities (environment, AI characters, objectives). Local multiplayer scenarios can be modeled with stronger emphasis on direct player-to-player interactions, often with richer adjacency structures reflecting physical proximity and shared interfaces. In online multiplayer contexts, as demonstrated in our PUBG case study, the model captures complex player-to-player interactions across potentially large networks, team-based dynamics, and player-environment interactions. The graph structure can flexibly represent competitive relationships (as negative edges or repulsive forces), cooperative alliances (as positive edges or attractive forces), or neutral interactions based on proximity or shared objectives. This flexibility enables our engagement estimation framework to accommodate diverse gaming modalities while maintaining a consistent mathematical formulation and measurement approach.

B. Problem Definition

Given a multiplayer online game environment, as described in the above system model, we consider the problem of estimating player engagement in real time.

Formally, at any time t during gameplay, for each player p, we aim to estimate their current engagement level $E_p(t)$ based on the historical telemetry data available up to time t, denoted as $\mathcal{H}_{\text{tele}}(t)$. This telemetry includes player actions, game states, and interaction patterns captured through standard game logs.

To concretize this problem definition, consider a player in a PUBG match at time t = 5 minutes into the game. The historical telemetry data $\mathcal{H}_{\text{tele}}(t)$ would include all player actions and game states up to that moment, such as the player's weapon acquisition sequence, early-game positioning decisions, initial resource gathering efficiency, and any early combat encounters. Our framework aims to estimate their current engagement level $E_p(t)$ based on these observable telemetry patterns before key gameplay outcomes materialize. For instance, the framework might detect declining engagement when a player's movement patterns become erratic after failing to find adequate equipment, allowing for potential interventions (such as nearby loot spawns) before the player becomes fully disengaged. Importantly, this estimation occurs without requiring any explicit feedback from the player, relying solely on behavioral signals captured through standard game logs.

This formulation advances beyond traditional engagement estimation approaches by emphasizing predictive capabilities - estimating engagement before critical gameplay moments materialize, rather than retroactively analyzing completed sessions. It also acknowledges the temporal nature of engagement, treating it as a dynamic measure that evolves throughout gameplay rather than a static post-game metric. While this real-time constraint introduces additional complexity, it enables practical applications in dynamic game adaptation.

In addition, our formulation does not require any intrusive physiological measurements or post-game questionnaires. It only utilizes standard telemetry data, which ensures broader applicability across existing game infrastructures and maintains non-intrusive monitoring of player experiences.

IV. PROPOSED SOLUTION

This section first specifies the design goals of the proposed framework and presents an overview of how it functions. It then describes the details of each component.

A. Design Goals and Solution Overview

The proposed framework is designed to achieve the following goals.

- **Relative Engagement Scale**: By comparing current engagement levels to an established baseline $\bar{E}$, we enable the detection of meaningful engagement variations while maintaining computational efficiency.
- **Flow Theory-Based Quantification**: Engagement is quantified through the relationship between player skill

and game challenge, requiring only standard telemetry data while maintaining theoretical grounding.
- **Non-Intrusive Instrumentation**: The framework exclusively utilizes game telemetry data available through standard logging systems, enabling scalable deployment without additional hardware requirements.
- **Hybrid Measurement Approach**: The framework combines theoretical foundations with machine learning-assisted measurement techniques, enabling both interpretable and accurate engagement measurement while maintaining flexibility for different game genres and contexts.

Figure 4 presents an overview of the proposed framework for real-time player engagement estimation. The framework comprises three main components: input processing, player interaction modeling, and engagement classification. At a high level, telemetry data is first processed to extract representative features characterizing both player behavior and game state. A dynamic player interaction network is constructed, where edges can represent various types of relationships such as spatial proximity, direct interactions, team affiliations, or shared object interactions. This network, along with the processed features, is then enriched through a hybrid neural architecture combining GCNs for modeling player interactions and Transformer networks for capturing temporal patterns. The enriched features are used to measure player skill and game challenge levels, which are then mapped to binary engagement states using a Random Forest classifier.

We provide the details of each component in the following subsections. First, we describe the input processing module that handles various types of telemetry features and constructs the player interaction network. Then, we elaborate on the player interaction modeling component that enriches these features through graph-based and temporal processing. Finally, we present the engagement classification module that estimates skill and challenge levels to determine player engagement states.

B. Input Processing

Building upon the telemetry processing system described in Section A, our framework transforms the real-time stream of raw telemetry events into three types of high-level features. These features are computed and sampled at fixed intervals (e.g., every 10 seconds) to create consistent temporal snapshots of the gameplay state.

- **Player Features** ($X_p(t) \in \mathbb{R}^{T \times N \times d_p}$): These features characterize individual player performance and behavior patterns. For each player, we aggregate telemetry events into meaningful metrics capturing combat performance (e.g., accuracy, damage dealt), mobility (e.g., distance traveled, position changes), resource utilization (e.g., item usage, inventory management), and spatial awareness (e.g., proximity to threats, zone positioning). Here, T represents the sequence length, N is the number of players, and d_p is the feature dimension. These metrics serve as proxies for player skill and adaptability.
- **Game State Features** ($X_g(t) \in \mathbb{R}^{T \times d_g}$): These features capture the evolving match context and environmental conditions that affect all players. By tracking match progression metrics such as elapsed time, remaining players, and zone states, we can contextualize individual player behaviors and better estimate the current challenge level. The dimension d_g represents the game state feature space.
- **Categorical Features** ($X_c(t)$): Discrete contextual information such as game mode and phase information is encoded through embedding layers. These features provide essential context for interpreting player behaviors and performance metrics, as similar actions may have different implications across different game modes or phases.

C. Player Interaction Modeling

The player interaction modeling component processes the input features through a hybrid neural architecture designed to capture both spatial and temporal relationships in gameplay. Following the graph-based game state representation defined in subsection A, we first construct a dynamic player network where nodes represent players and edges represent their relationships. The edge weights w_{ij} between players i and j can encode various types of interactions such as spatial proximity, direct combat engagement, or team-based cooperation.

This player network is processed through a GCN consisting of multiple layers with decreasing dimensionality to learn compact representations that capture the structural properties of player interactions. The GCN architecture employs skip connections between layers to preserve individual player features while learning interaction-based representations. The GCN outputs are then combined with the original player features through concatenation to create enriched representations:

$$\mathbf{H}_{\text{enriched}}(t) = [\text{GCN}(X_p(t)); X_p(t)] \tag{1}$$

To capture temporal dependencies, we employ a multi-layer Transformer encoder with multiple attention heads. Before processing, the game state features $X_g(t)$ are expanded along the player dimension to enable element-wise operations with the player-specific features in $\mathbf{H}_{\text{enriched}}(t)$. The Transformer processes these combined features through self-attention mechanisms, enabling the model to identify relevant temporal patterns and long-range dependencies in player behavior. The multi-head attention architecture allows the model to capture different aspects of temporal relationships simultaneously:

$$\mathbf{H}_{\text{temp}}(t) = \text{Transformer}(\mathbf{H}_{\text{enriched}}(t), X_g(t)) \tag{2}$$

$$\mathbf{H}_{\text{pool}}(t) = \text{Pool}(\mathbf{H}_{\text{temp}}(t)) \tag{3}$$

The categorical features are processed through embedding layers that map each discrete feature to a lower-dimensional dense representation $\mathbf{H}_{\text{cat}}(t) = \text{Embed}(X_c(t))$. All processed features are then combined through a multi-layer perceptron (MLP) with specialized output heads for skill and challenge measurement:

$$\mathbf{Z}(t) = \text{MLP}([\mathbf{H}_{\text{pool}}(t); \mathbf{H}_{\text{cat}}(t)]) \tag{4}$$

This hierarchical processing enables our framework to capture complex player interactions at multiple scales while maintaining the temporal context necessary for engagement measurement.

D. Engagement Classification

The final component produces engagement measures through a two-step process that explicitly models the relationship between player skill and game challenge. First, from the processed features $\mathbf{Z}(t)$, we measure skill and challenge levels through separate prediction heads:

$$\hat{S}(t) = f_{\text{skill}}(\mathbf{Z}(t)) \tag{5}$$

$$\hat{C}(t) = f_{\text{challenge}}(\mathbf{Z}(t)) \tag{6}$$

The activation functions for these measures are chosen based on the nature of the underlying skill and challenge proxies. For example, when using normalized ranking as a skill proxy, we employ a sigmoid activation for f_{skill} to bound the output between 0 and 1. In contrast, when using metrics like damage received as challenge proxies, we utilize Rectified Linear Unit (ReLU) activation for $f_{\text{challenge}}$ to handle unbounded positive values. This choice of activation functions can and should be adapted based on the specific proxies used in different game contexts.

The measured skill and challenge levels are then mapped to binary engagement states through a Random Forest classifier:

$$\hat{E}(t) = f_{\text{classify}}(\hat{S}(t), \hat{C}(t)) \tag{7}$$

where engagement is defined relative to a baseline $\bar{E}$:

$$E_{\text{binary}}(t) = \begin{cases} 1 & \text{if } E(t) > \bar{E} \text{ (High Engagement)} \\ 0 & \text{if } E(t) \leq \bar{E} \text{ (Low Engagement)} \end{cases} \tag{8}$$

The baseline $\bar{E}$ is established through a one-time calibration process using self-reported engagement levels from player surveys. This calibration is crucial for inherently engaging game genres, such as competitive multiplayer games, where most players maintain some baseline level of engagement. In such contexts, the baseline helps distinguish subtle variations in engagement levels rather than merely detecting obvious disengagement. While our implementation uses survey responses for baseline calibration, alternative approaches such as expert annotations or behavioral indicators could be used depending on the available data and specific game context.

The choice of Random Forest for the final classification aligns with the non-linear nature of the skill-challenge relationship in Flow Theory. It can capture complex decision boundaries between engagement states while providing interpretable feature importance scores that help validate the relative impact of skill and challenge on engagement predictions.

V. EVALUATION

We evaluate our engagement estimation framework through a systematic validation process, focusing both on component-level performance and end-to-end effectiveness. Our evaluation employs PlayerUnknown's Battlegrounds (PUBG) as a case study, leveraging its rich telemetry data and diverse gameplay mechanics to thoroughly assess our framework's capabilities in a real-world setting. Throughout this evaluation section, all reported uncertainties (±) represent one standard deviation of the corresponding metric.

A. Experimental Setup

1) Framework Implementation

We implement our framework for PUBG, a battle royale game where approximately 100 players compete in teams across large maps, starting with no equipment and scavenging for resources while avoiding elimination. The game naturally segments into distinct phases as the playable area progressively shrinks, with each phase typically lasting 2-3 minutes. For skill and challenge quantification, we define:

$$S_p(t) = \frac{\text{number of players eliminated before p}}{\text{total number of players - 1}} \tag{9}$$

$$C_p(t) = \text{total damage taken by player p in phase t} \tag{10}$$

Our implementation samples telemetry data at 10-second intervals. The feature dimensions are:

- $X_p(t) \in \mathbb{R}^{T \times N \times 35}$ for player features
- $X_g(t) \in \mathbb{R}^{T \times 4}$ for game state features
- $X_r(t)$ includes map identifier, team size, and phase index

The player interaction graph $G(t)$ is constructed with edge weights:

$$w_{ij} = \max(0, 1 - \frac{d_{ij}}{d_{max}}) \cdot [(1-\sigma(\theta)) \cdot \mathbb{1}_{enemy} + \sigma(\theta) \cdot \mathbb{1}_{teammate}] \tag{11}$$

where d_{ij} is the Euclidean distance between players (capped at $d_{max} = 100$ meters), θ is a learnable team weight parameter, and $\mathbb{1}_{enemy}$, $\mathbb{1}_{teammate}$ are binary indicators for enemy/teammate relationships.

While we demonstrate our framework using PUBG as our primary case study, the approach is designed to be genre-agnostic. We specifically selected PUBG because it represents a highly complex gaming environment that spans multiple genres (shooting, combat, scavenging, multiplayer, survival), providing an exceptionally challenging test scenario that is also commercially popular and realistic. If our framework can effectively measure engagement in PUBG's complex environment with its varied gameplay elements, it should be adaptable to less complex gaming scenarios

across different genres including sports games, racing games, strategy games, and more. For instance, in sports games, skill could be measured through performance metrics like scoring efficiency or ball possession, while challenge might be quantified through opponent defensive pressure. In racing games, skill could be measured via lap times or overtaking maneuvers, while challenge might be represented by track difficulty or competitor performance. This flexibility allows our engagement measurement approach to extend beyond combat-focused games to virtually any interactive gaming experience that generates telemetry data.

When calculating skill based on player ranking, we maintained the natural composition of PUBG matches, including both human players and AI-controlled bots. This approach preserves the authentic gameplay experience, as players typically don't distinguish between human and AI opponents during combat. While bots are present in the environment, our skill and challenge validation metrics were evaluated specifically on human player data, ensuring the framework's effectiveness for measuring human engagement.

2) Dataset

Our evaluation utilizes two complementary datasets: a skill-challenge estimation dataset for model development and a survey dataset for engagement validation. Both datasets share the same underlying structure, capturing game telemetry at 10-second intervals.

For the skill-challenge estimation dataset, we implemented a systematic sampling strategy starting with five seed players (professionals, streamers, community members), expanding to 1,684 unique players through their recent match histories. Players were categorized into five tiers based on lifetime match count using IQR-based outlier removal and quantile-based discretization, ranging from Rookies (< 374 matches, avg. 146±95) to Masters ($> 6,369$ matches, avg. 10,345±4,034), with Amateur (374-1,161 matches), Veteran (1,161-2,780 matches), and Elite (2,780-6,369 matches) tiers in between.

For efficient data collection, we selected four players from each tier, resulting in a balanced sample of 20 players. We monitored their battle royale matches (solo, duo, and squad modes) over a two-week period, collecting complete telemetry data for 2,673 matches. After preprocessing and phase-based segmentation, this yielded 20,030 data points, which we split into training (16,267), validation (1,866), and testing (1,897) sets.

For engagement validation, we conducted a data collection experiment involving 31 players. The experiment was approved by University of Ottawa's Office of Research Ethics and Integrity, file Number H-07-23-9439. Participants registered with their PUBG username and demographic information in our web application, then submitted post-match experiences through a structured questionnaire derived from the Game Experience Questionnaire (GEQ) [14]. For engagement measurement, participants rated their level from "Disengaged" (feeling bored, unfocused) to "Highly Engaged" (losing track of time, fully immersed), with intermediate levels of "Slightly," "Moderately," and "Fairly" engaged. Each level included descriptive examples to ensure consistent interpretation.

Our data collection workflow prioritized ecological validity - participants played PUBG matches normally, then immediately reported their skill level, perceived challenge, and engagement on 5-point Likert scales to minimize recall bias [15]. We aligned survey responses with telemetry data by matching submission timestamps with corresponding match data retrieved via the PUBG API using participants' usernames. We processed the telemetry data using the same phase-based approach, resulting in 120 labeled data points. To address measurement uncertainty in self-reported engagement, we transformed the Likert responses into binary classifications using the mean reported engagement (3.58) as the threshold and employed group-stratified cross-validation to account for person-specific reporting tendencies. These self-reported scores provided ground truth labels for framework validation. Table 1 summarizes the overall dataset statistics.

TABLE 1: Dataset Statistics Summary

Dataset	Matches	Datapoints	Labels
Skill-Challenge	2,673	20,030	-
Train	2,173	16,267	-
Validation	250	1,866	-
Test	250	1,897	-
Survey	31	120	31

3) Training Configuration

For PUBG implementation, we configured the framework with 6 temporal snapshots per sequence (1 minute of gameplay) and 35 player features. The categorical embeddings were dimensioned specifically for PUBG's feature cardinality: team size (3→2), map ID (12→6), and phase index (11→4).

The GCN implementation uses two convolutional layers (64 and 16 output units) to process the player interaction graph. The Transformer encoder was configured with two layers, two attention heads, and a hidden dimension of 128. We maintained pre-layer normalization for stable training with the game's variable player counts.

Training proceeded with the AdamW optimizer (learning rate: $3e^{-4}$, weight decay: 0.01) using cosine annealing schedule. The model trained for maximum 50 epochs with early stopping (patience: 5), using batches of 128 sequences. To handle PUBG's variable player counts per match, we implemented masked loss computation for active players only.

The engagement classifier was calibrated using our PUBG survey dataset (n=31). We addressed the granularity mismatch between phase-level predictions and match-level surveys by aggregating phase estimates using final normalized ranking for skill and mean damage across phases for challenge. High engagement thresholds were determined using mean reported engagement scores (3.58 ± 0.56). This notably high baseline engagement aligns with PUBG's status as a competitive battle royale game, where the inherent match stakes and elimination mechanics naturally foster high player investment.

B. Results & Analysis

1) Framework Validation

We evaluate our engagement estimation framework through a hierarchical validation approach that progresses from theoretical foundations to practical implementation. Beginning with survey-based engagement prediction as an upper bound, we systematically validate our telemetry-based proxies before assessing the complete end-to-end framework, as illustrated in Figure 5.

The first stage, shown in the top of the figure, employs Clf_1, an AdaBoost classifier (empirically selected for optimal performance on survey data) that maps self-reported skill and challenge to self-reported engagement. This establishes our empirical baseline for engagement prediction from explicit player feedback. We configured AdaBoost with SAMME boosting algorithm specifically because of its effectiveness with categorical features and resilience to overfitting on small datasets like our survey responses. This stage establishes the theoretical ceiling for engagement prediction accuracy using explicit skill-challenge relationships.

The second stage, shown in the middle of the figure, evaluates our telemetry-based proxies using Clf_2, a Random Forest classifier (chosen for its superior performance on game statistics) operating on match completion statistics—final ranking for skill and average damage taken per phase for challenge. While this stage requires complete match data, it serves to validate our proxy selection methodology. This intermediate validation stage is crucial for demonstrating that our selected telemetry proxies can approach the performance of explicit player feedback, proving the viability of non-intrusive measurement. It validates that game-derived metrics can effectively replace traditional survey methods while maintaining accuracy.

Our proposed end-to-end framework, represented by the third stage shown at the bottom of the figure, combines skill-challenge measurements from the GCN-Transformer model with Clf_2. We implement a forward-sliding cross-validation scheme that holds out five survey responses in each fold, maintaining complete separation between training and testing data. This real-time validation approach represents the framework's primary innovation: providing engagement estimates during gameplay rather than retroactively. By successfully approximating match-level predictions using only partial gameplay data, our framework enables adaptive game mechanics that can respond to fluctuating engagement levels within a single match. During real-time operation, the framework processes each phase independently: measuring current skill from performance up to the current phase, measuring the challenge level for the current phase, and applying the appropriate fold-specific classifier to these phase-level measures.

Given the granularity mismatch between phase-level measures and match-level ground truths, we average phase-level skill $\hat{S}(t)$ and challenge $\hat{C}(t)$ measures before engagement classification. This approach demonstrates our framework's potential for real-world deployment across varying time scales, from moment-to-moment gameplay adaptation to session-level analytics. Game designers can leverage these multi-resolution engagement signals to optimize both immediate mechanics and broader game progression systems. We specifically avoid averaging phase-level engagement measures $\hat{E}(t)$ since Clf_2 was trained on match-level skill and challenge scores, so averaging engagement would incorrectly assume linear composition across phases, contradicting flow theory [17].

2) Training Dynamics

The learning curves (Figure 6) demonstrate stable convergence for both skill and challenge estimation. While the total validation loss shows some fluctuation early in training, it stabilizes around epoch 15, indicating robust model generalization. The skill component converges more quickly and shows minimal gap between training and validation performance, suggesting effective learning of ranking patterns. The challenge component exhibits a larger training-validation gap but maintains consistent improvement throughout training.

3) Phase-wise performance

As shown in Figure 7, our GCN-Transformer model's accuracy varies significantly across game phases. Skill estimation error (blue line) demonstrates a consistent improvement pattern, with RMSE decreasing steadily from 0.235 in phase 0 to 0.024 in phase 9. This trend reflects the model's increasing ability to more accurately measure player skill as more gameplay data becomes available.

Challenge estimation (red line) exhibits a different pattern, with high initial error (RMSE 0.703-0.747 in phases 0-1) followed by a sharp improvement in phase 2 (RMSE 0.294). The model achieves its best challenge predictions in phase 5 (RMSE 0.166), coinciding with mid-game player confrontations as the playable area constricts. However, both skill and challenge predictions show increased error in the final phases (9-10), likely due to reduced player count and heightened end-game volatility.

These results indicate that our model's accuracy is phase-dependent, performing optimally during mid-game phases

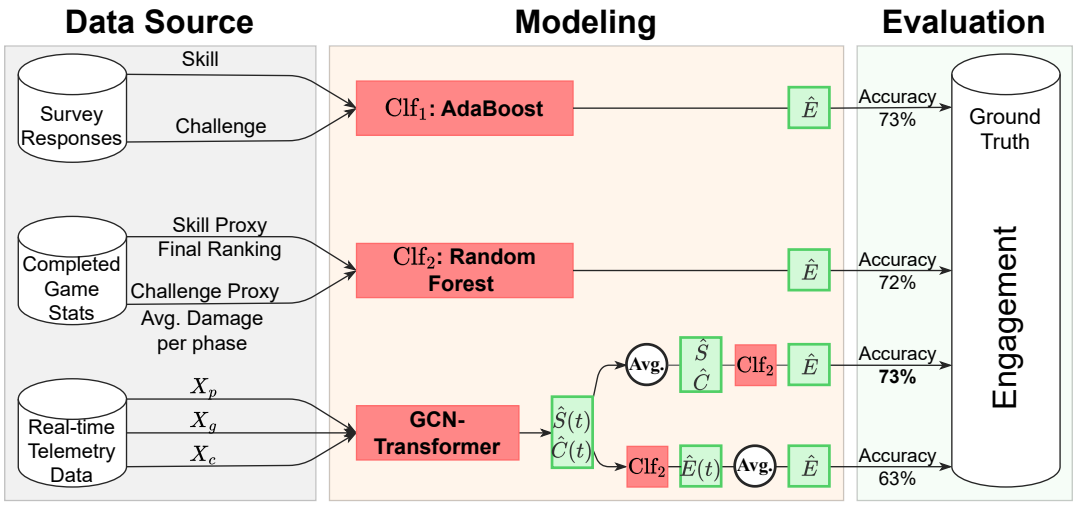


FIGURE 5: Validation framework for engagement estimation.

TABLE 2: Classification Performance Across Different Input Sources

Evaluation Stage	Model	Metric	Low	High	Macro Avg	ROC_AUC	accuracy
Clf_1 (Survey Responses)	AdaBoost	precision	0.64 ± 0.44	0.79 ± 0.22	0.71 ± 0.23		
		recall	0.55 ± 0.40	0.88 ± 0.16	0.72 ± 0.22	0.75 ± 0.32	0.73 ± 0.17
		f1-score	0.53 ± 0.35	0.80 ± 0.13	0.66 ± 0.22		
Clf_2 (Post-match Proxies)	Random Forest	precision	0.72 ± 0.30	0.81 ± 0.19	**0.76 ± 0.16**		
		recall	0.68 ± 0.30	0.77 ± 0.24	0.73 ± 0.12	0.76 ± 0.16	0.72 ± 0.12
		f1-score	0.63 ± 0.19	0.75 ± 0.13	0.69 ± 0.14		
Clf_2 (Avg. of Real-time Estimates)	Random Forest	precision	0.63 ± 0.19	0.87 ± 0.23	0.75 ± 0.16		
		recall	0.88 ± 0.21	0.66 ± 0.23	**0.77 ± 0.12**	**0.83 ± 0.17**	**0.73 ± 0.14**
		f1-score	0.71 ± 0.15	0.73 ± 0.19	**0.72 ± 0.15**		

where player interactions are most structured and predictable.

4) Sequence-Length Sensitivity

We systematically evaluated model architectures trained with different sequence lengths (1-10 timesteps at 10-second intervals) to determine the optimal temporal window for engagement measurement. As shown in Figure 8, all performance metrics peak at sequence length 6, with ROC-AUC reaching 0.83±0.17, achieving an accuracy of 0.73±0.14 and an F1-score of 0.71±0.15. This indicates that one minute of gameplay data is sufficient for reliable measures, which is particularly important given the short duration of game phases in PUBG.

Models trained with longer sequences not only show degraded performance but also exhibit increased variance, as evidenced by the widening standard deviation bands beyond length 7. To handle variable-length sequences in practice, our implementation employs padding when the available sequence is shorter than the target length (e.g., due to early player elimination), while longer sequences are trimmed. This approach ensures consistent input dimensionality while maintaining temporal relevance of the features.

The ability to make accurate predictions with just one minute of data enables responsive engagement measurement within the typical duration of game phases, making our framework practical for real-time applications.

5) Engagement Classifier

We evaluated several classification algorithms including Random Forest, AdaBoost, and Support Vector Machines (SVM) for engagement classification. While both Random Forest and AdaBoost demonstrated competitive performance, their relative effectiveness varied based on the input features used. The Random Forest classifier emerged as the optimal choice for telemetry-based proxies, achieving an ROC-AUC of 0.83±0.17 and accuracy of 0.73±0.14.

Training on different input sources revealed distinct patterns in feature importance. With survey-based inputs, AdaBoost showed a clear bias toward challenge scores (0.622±0.062) over skill scores (0.378±0.062). In contrast, the Random Forest trained on telemetry-based proxies exhib-

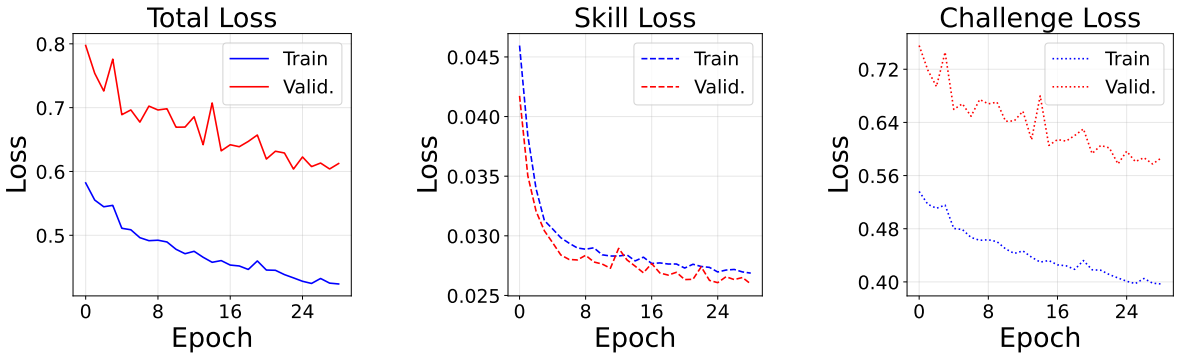


FIGURE 6: Training and validation loss curves.

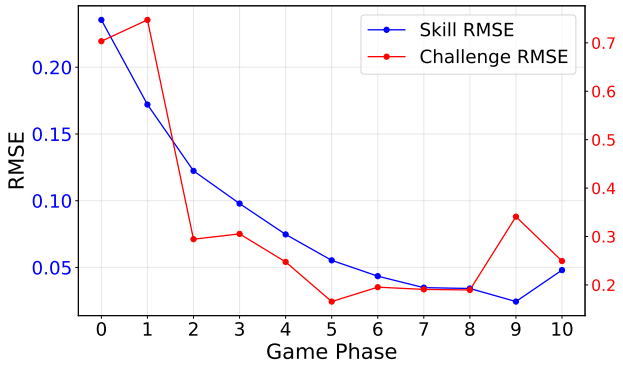


FIGURE 7: Phase-wise RMSE for skill and challenge measurement.

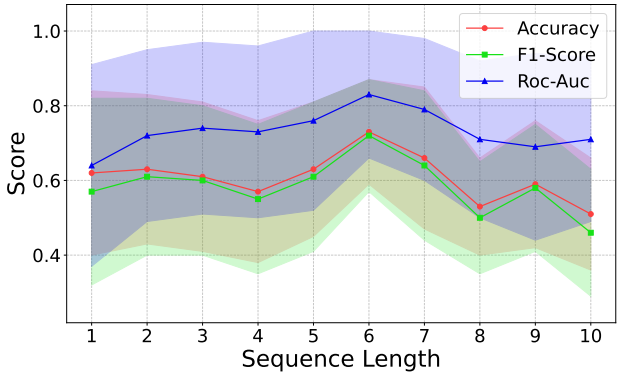


FIGURE 8: Impact of sequence length on model performance.

ited remarkably balanced feature importance between skill (0.502±0.029) and challenge (0.498±0.029). The minimal standard deviation (±0.029) across cross-validation folds indicates robust stability in this balanced relationship.

The contrast between survey-based and proxy-based feature importances reveals an interesting psychological aspect: while players may be more consciously aware of challenge levels during gameplay, our telemetry-based proxies capture a more balanced representation of the skill-challenge relationship. This finding suggests that objective gameplay metrics may better reflect the theoretical engagement model than subjective player assessments, possibly due to reporting biases or varying interpretations of skill and challenge across players.

6) Ablation Study

To assess the individual contribution of skill and challenge components, we conducted isolated evaluations using single-feature classifiers. The results are summarized in Table 3.

TABLE 3: Component-wise Classification Performance

Component	ROC-AUC	Accuracy
Skill-only	0.57 ± 0.20	0.63 ± 0.15
Challenge-only	0.69 ± 0.14	0.61 ± 0.16
Combined	**0.83 ± 0.17**	**0.73 ± 0.14**

The challenge-only classifier achieved higher ROC-AUC but lower accuracy compared to the skill-only variant, suggesting that challenge levels may be more discriminative but less reliable for binary engagement classification. However, the combined approach significantly outperformed both individual components, with improvements of 20.3% and 16.7% in ROC-AUC over skill-only and challenge-only classifiers, respectively. This substantial performance gain validates our framework's theoretical foundation in flow theory and demonstrates the synergistic relationship between skill and challenge in engagement measurement.

We also conducted extensive ablation experiments to evaluate the contribution of different feature categories, model architecture components, and assess the individual impact of skill and challenge components on the framework's performance.

To validate our hybrid architecture design, we compared the full GCN-Transformer model against a Transformer-only variant that excludes the graph convolutional component. The hybrid architecture (accuracy: 0.73±0.14, ROC-

AUC: 0.83±0.17) outperforms the Transformer-only model (accuracy: 0.67±0.19, macro F1: 0.65±0.21), suggesting that the GCN's ability to model player interactions provides valuable information for engagement measurement. The Transformer-only model shows stronger precision for high engagement states (0.80±0.28) but suffers from reduced recall (0.63±0.25) compared to the hybrid approach.

Table 4 presents the impact of removing different feature categories on both skill-challenge estimation and end-to-end engagement measurement. The baseline model, utilizing all features, achieves the best performance across most metrics. Removing player features significantly degrades skill estimation (RMSE increases from 0.163 to 0.319) and end-to-end accuracy (73% to 55%). Similarly, excluding game features impairs challenge estimation (RMSE increases from 0.541 to 0.604) and reduces end-to-end accuracy to 61%. Categorical features show the least impact, with marginal changes in skill-challenge estimation and moderate degradation in end-to-end performance.

TABLE 4: Impact of Feature Removal on Model Performance

Features Removed	Skill RMSE	Challenge RMSE	Acc. (%)	ROC-AUC
none	**0.163**	**0.541**	**73±14**	**83±17**
player	0.319	0.707	55±15	59±21
game	0.343	0.604	61±14	56±16
categorical	0.172	0.535	57±18	72±18

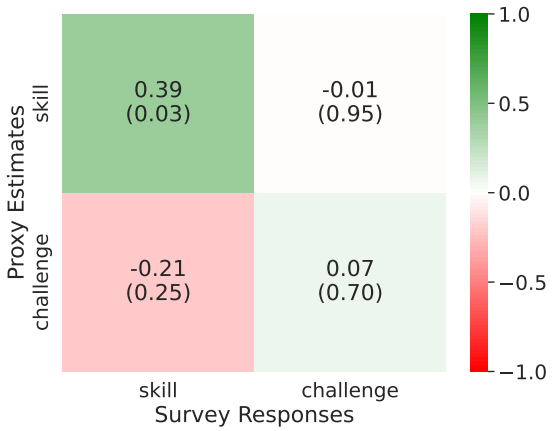


FIGURE 9: Correlation between model estimates and player survey responses

7) Robustness to Perceptual Variability

Figure 9 examines the relationship between our telemetry-based estimates and players' self-reported perceptions through Spearman's rank correlation analysis. Our skill estimates show a significant moderate correlation with self-reported skill ($\rho = 0.39$, $p = 0.03$), suggesting our ranking-based proxy effectively captures aspects of player-perceived skill. Interestingly, our challenge estimates show limited correlation with self-reported challenge ($\rho = 0.07$, $p = 0.70$). Given the strong predictive power of our challenge proxy demonstrated in the ablation study (ROC-AUC 0.69 for challenge-only classifier), this finding suggests that effective engagement measurement may not require direct alignment with players' subjective challenge perceptions. Instead, our telemetry-based challenge metric appears to capture gameplay patterns that, while distinct from players' self-reported experiences, provide valuable signals for engagement estimation.

8) Cross-Domain Validation

To assess our framework's generalizability beyond PUBG, we conducted extensive validations across two distinct game genres: FIFA'23, a sports game, and Street Fighter V, a 2.5D fighting game. This cross-domain study involved 39 participants across 900 gaming sessions. These games were specifically selected for their structured difficulty systems - 6 levels in FIFA'23, 8 in Street Fighter V - providing clear challenge metrics, and their distinct gameplay mechanics allowing robust assessment of skill-challenge dynamics across genres.

For this cross-domain validation, we focused on testing the core theoretical relationship between skill, challenge, and engagement rather than implementing full telemetry extraction systems. Natural gameplay segments defined session boundaries—between goals or halftimes in FIFA'23, and between rounds in Street Fighter V—which helped mitigate recall bias by allowing engagement measurement immediately after short, meaningful gameplay segments. We collected and processed three key measurements:

- **Skill Measurement**: We assessed initial skill levels through self-reported game familiarity using a 5-point Likert scale questionnaire with game-specific prompts: "How familiar are you with 2.5D fighting games (like Street Fighter)?" and "How familiar are you with FIFA games?". Each participant's reported familiarity with the corresponding game genre was used as their skill level for that specific gameplay session.
- **Challenge Measurement**: We utilized the preset difficulty settings available in each game as objective challenge metrics. For Street Fighter V, this ranged from levels 1 to 8, while FIFA'23 offered six difficulty tiers: "Beginner", "Amateur", "Semi-Pro", "Professional", "World Class", and "Legendary". To standardize these metrics across games, we normalized each difficulty setting by dividing by the maximum available level (e.g., difficulty 3 in FIFA "Semi-Pro" was normalized to 0.5 by dividing by 6).

- **Engagement Measurement**: After each gameplay session, participants reported their engagement on a 5-point Likert scale. Following our methodology from the PUBG study, we binarized these responses using the mean reported engagement score (2.87) as the threshold, classifying the lowest three levels as "Low Engagement" and the upper two levels as "High Engagement".

Although we did not extract actual telemetry data as part of this study, our framework could theoretically be implemented using game-specific metrics. For FIFA, these could include successful pass rate, possession time, yellow and red cards, fouls committed, and shots on goal as player features. Skill proxies in this context could be calculated from score differences in past matches, while challenge proxies could be derived from the opponent's possession statistics or shots on goal. Similarly, for Street Fighter V, while it does contain combat elements like PUBG (e.g., damage, health points), the gameplay mechanics, perspectives, and skill requirements differ substantially, demonstrating the framework's adaptability to varied combat paradigms.

TABLE 5: Classification Performance Across Game Genres

Metric	Low	High	Accuracy	Macro Avg
Precision	0.43 ± 0.16	0.77 ± 0.08	0.66 ± 0.06	0.60 ± 0.08
Recall	0.45 ± 0.14	0.75 ± 0.08	0.66 ± 0.06	0.60 ± 0.07
F1-score	0.43 ± 0.12	0.76 ± 0.05	0.66 ± 0.06	0.59 ± 0.08

As shown in Table 5, our framework as is and without applying transfer learning, maintains robust performance across these diverse game contexts. Using a Random Forest classifier with 7-fold stratified group cross-validation (ensuring participant-level separation), the model achieves 66% accuracy (±6%) with particularly strong performance in detecting high engagement states (precision: 0.77±0.08, recall: 0.75±0.08). These results demonstrate that our framework's fundamental premise — engagement as a function of skill-challenge balance — generalizes effectively across game genres when provided with appropriate skill and challenge metrics.

Notably, the framework's performance remains stable despite the significant differences in gameplay mechanics, session duration, and competitive dynamics between the test domains. This robustness suggests that our approach captures fundamental aspects of player engagement that transcend specific game mechanics, supporting its potential application across diverse gaming contexts. Performance is expected to further improve by applying transfer learning to our model for the specific game at hand, as demonstrated by the application of transfer learning to models in other domains [59].

For this validation, we focused on the core theoretical relationship between skill, challenge, and engagement rather than implementing full telemetry extraction systems. We collected three key measurements: initial skill level (self-reported genre familiarity), challenge level (normalized game difficulty), and post-session engagement (5-point Likert scale, binarized at mean reported engagement 2.87). While implementing comprehensive telemetry extraction for these games was beyond the scope of this study, our framework could theoretically be extended to utilize sports-specific metrics in FIFA (e.g., possession percentage, shots on goal) or fighting-game metrics in Street Fighter V (e.g., combo execution rates, defensive reactions).

VI. CONCLUSION

This paper introduces a framework for real-time player engagement estimation that advances the state-of-the-art through its predictive capabilities and non-intrusive nature. By combining GCNs for player interaction modeling with Transformer networks for temporal processing, our framework successfully predicts skill and challenge levels before their manifestation in gameplay. Our analysis revealed several interesting findings: phase-specific predictions can be effectively made using one minute of gameplay data, player features proved most critical for accurate estimation, and objective gameplay metrics may better reflect engagement than subjective player assessments, possibly due to reporting biases.

Our results empirically demonstrate the effectiveness of Flow Theory's skill-challenge relationship in quantifying engagement within modern multiplayer games, particularly during structured mid-game interactions where our framework showed peak performance. This work provides game developers with a powerful tool for understanding and optimizing player engagement without disrupting the gaming experience, marking a significant step toward more engaging and adaptive multiplayer online games.

Importantly, our cross-domain validation demonstrates that the framework extends well beyond combat games, with robust performance observed in sports games (FIFA'23) and fighting games (Street Fighter V), suggesting broad applicability across the gaming industry regardless of genre.

This work provides game developers and researchers with a practical, non-intrusive instrument for analyzing and monitoring player engagement using existing telemetry data. The framework's predictive capabilities enable proactive game adjustments and more efficient resource allocation in cloud gaming [28], [60]. Its scalability makes it suitable for large-scale deployment across various gaming platforms and genres, with experimental validation demonstrating robust measurement performance across different game types including sports games and fighting games. By detecting subtle variations in engagement levels relative to an established baseline, our measurement approach offers more nuanced insights than traditional binary engagement classifications while maintaining the natural flow of gameplay [61].

REFERENCES

[1] K. Arora, "The gaming industry: A behemoth with unprecedented global reach," 2023.

[2] L. A. Gil-Aciron, "The gamer psychology: a psychological perspective on game design and gamification," *Interactive Learning Environments*, vol. 32, no. 1, pp. 183–207, 2024.

[3] A. Z. Abbasi, D. H. Ting, and H. Hlavacs, "Engagement in games: Developing an instrument to measure consumer videogame engagement and its validation," *International Journal of Computer Games Technology*, vol. 2017, no. 1, p. 7363925, 2017.

[4] S. Poeller, S. Seel, N. Baumann, and R. L. Mandryk, "Seek what you need: Affiliation and power motives drive need satisfaction, intrinsic motivation, and flow in league of legends," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CHI PLAY, pp. 1–23, 2021.

[5] D.-I. D. Han, F. Melissen, and M. Haggis-Burridge, "Immersive experience framework: a delphi approach," *Behaviour & information technology*, pp. 1–17, 2023.

[6] Z. Yu, M. Gao, and L. Wang, "The effect of educational games on learning outcomes, student motivation, engagement and satisfaction," *Journal of Educational Computing Research*, vol. 59, no. 3, pp. 522–546, 2021.

[7] F. Hadiji, R. Sifa, A. Drachen, C. Thurau, K. Kersting, and C. Bauckhage, "Predicting player churn in the wild," in *2014 IEEE Conference on Computational Intelligence and Games*. Dortmund, Germany: IEEE, 2014, pp. 1–8.

[8] V. Bonometti, C. Ringer, M. Hall, A. Wade, and A. Drachen, "Modelling early user-game interactions for joint estimation of survival time and churn probability," in *2019 IEEE Conference on Games (CoG)*. London, UK: IEEE, 2019, pp. 1–8.

[9] B. Bontchev and D. Vassileva, "Assessing engagement in an emotionally-adaptive applied game," in *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, ser. TEEM '16. New York, NY: Association for Computing Machinery, 2016, p. 747–754.

[10] T. H. Laine and R. S. N. Lindberg, "Designing engaging games for education: A systematic literature review on game motivators and design principles," *IEEE Transactions on Learning Technologies*, vol. 13, no. 4, pp. 804–821, 2020.

[11] X. Zhong and J. Xu, "Game updates enhance players' engagement: A case of dota2," in *Proceedings of the 4th International Conference on Information Management and Management Science*, 2021, pp. 117–123.

[12] D. Gábana Arellano, L. Tokarchuk, and H. Gunes, "Measuring affective, physiological and behavioural differences in solo, competitive and collaborative games," in *Intelligent Technologies for Interactive Entertainment*, R. Poppe, J.-J. Meyer, R. Veltkamp, and M. Dastani, Eds. Cham: Springer International Publishing, 2017, pp. 184–193.

[13] W. Yang, M. Rifqi, C. Marsala, and A. Pinna, "Towards better understanding of player's game experience," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '18. New York, NY: Association for Computing Machinery, 2018, p. 442–449.

[14] K. L. Norman, "Geq (game engagement/experience questionnaire): a review of two papers," *Interacting with computers*, vol. 25, no. 4, pp. 278–283, 2013.

[15] E. Hassan, "Recall bias can be a threat to retrospective and prospective research designs," *The Internet Journal of Epidemiology*, vol. 3, no. 2, pp. 339–412, 2006.

[16] G. Hookham and K. Nesbitt, "A systematic review of the definition and measurement of engagement in serious games," in *Proceedings of the Australasian Computer Science Week Multiconference*, ser. ACSW 2019. New York, NY: Association for Computing Machinery, 2019.

[17] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*. New York, NY: Harper Perennial, 1990.

[18] A. Apicella, P. Arpaia, G. Mastrati, N. Moccaldi, and R. Prevete, "Preliminary validation of a measurement system for emotion recognition," in *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2020, pp. 1–6.

[19] A. Eddin Alchalabi, M. Elsharnouby, S. Shirmohammadi, and A. Nour Eddin, "Feasibility of detecting adhd patients' attention levels by classifying their eeg signals," in *2017 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2017, pp. 314–319.

[20] M. Catelani, L. Ciani, and C. Risaliti, "Risk assessment in the use of medical devices: A proposal to evaluate the impact of the human factor," in *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2014, pp. 1–6.

[21] A. Miguel-Cruz, A. M. R. Rincon, C. Daum, D. A. Q. Torres, R. De Jesus, L. Liu, and E. Stroulia, "Predicting engagement in older adults with and without dementia while playing mobile games," *IEEE Instrumentation Measurement Magazine*, vol. 24, no. 6, pp. 29–36, 2021.

[22] A. Lochbihler, B. Wallace, K. V. Benthem, C. Herdman, W. Sloan, K. Brightman, J. Goheen, F. Knoefel, and S. Marshall, "Assessing driver engagement through machine learning classification of physiological measures," in *2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2023, pp. 1–6.

[23] A. Baldwin, D. Johnson, and P. A. Wyeth, "The effect of multiplayer dynamic difficulty adjustment on the player experience of video games," in *CHI'14 extended abstracts on human factors in computing systems*, 2014, pp. 1489–1494.

[24] P. D. Paraschos and D. Koulouriotis, "Game difficulty adaptation and experience personalization: A literature review," *International Journal of Human–Computer Interaction*, vol. 39, pp. 1 – 22, 2022.

[25] M. F. Maleki and R. Zhao, "Procedural content generation in games: A survey with insights on emerging llm integration," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 20, no. 1, 2024, pp. 167–178.

[26] M. Chen, A. N. Elmachtoub, and X. Lei, "Matchmaking strategies for maximizing player engagement in video games," *Available at SSRN 3928966*, 2021.

[27] K. Wang, H. Liu, Z. Hu, X. Feng, M. Zhao, S. Zhao, R. Wu, X. Shen, T. Lv, and C. Fan, "Enmatch: matchmaking for better player engagement via neural combinatorial optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 9098–9106.

[28] A. A. Laghari, H. He, K. A. Memon, R. A. Laghari, I. A. Halepoto, and A. Khan, "Quality of experience (qoe) in cloud gaming models: A review," *multiagent and grid systems*, vol. 15, no. 3, pp. 289–304, 2019.

[29] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," *ACM SIGCOMM computer communication review*, vol. 41, no. 4, pp. 362–373, 2011.

[30] K. Doherty and G. Doherty, "Engagement in hci: Conception, theory and measurement," *ACM Comput. Surv.*, vol. 51, no. 5, nov 2018.

[31] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary educational psychology*, vol. 25, no. 1, pp. 54–67, 2000.

[32] E. N. Wiebe, A. Lamb, M. Hardy, and D. Sharek, "Measuring engagement in video game-based environments: Investigation of the user engagement scale," *Computers in Human Behavior*, vol. 32, pp. 123–132, 2014.

[33] D. Weibel and B. Wissmath, "Immersion in computer games: The role of spatial presence and flow," *International Journal of Computer Games Technology*, vol. 2011, jan 2011.

[34] K. Procci, "The subjective gameplay experience: An examination of the revised game engagement model," Ph.D. dissertation, University of Central Florida, 2015.

[35] T. Terkildsen and G. Makransky, "Measuring presence in video games: An investigation of the potential use of physiological measures as indicators of presence," *International Journal of Human-Computer Studies*, vol. 126, pp. 64–80, 2019.

[36] L. Ermi and F. Mäyrä, "Fundamental components of the gameplay experience: Analyzing immersion," *Worlds in play: International perspectives on digital games research*, vol. 21, p. 37, 2007.

[37] B. Cowley, D. Charles, M. Black, and R. Hickey, "Toward an understanding of flow in video games," *Computers in Entertainment (CIE)*, vol. 6, no. 2, pp. 1–27, 2008.

[38] J. C. Read, S. MacFarlane, and C. Casey, "Endurability, engagement and expectations: Measuring children's fun," in *Interaction design and children*, vol. 2. Eindhoven, Netherlands: Shaker Publishing, 2002, pp. 1–23.

[39] H. L. O'Brien and E. G. Toms, "What is user engagement? a conceptual framework for defining user engagement with technology,"

Journal of the American society for Information Science and Technology, vol. 59, no. 6, pp. 938–955, 2008.

[40] P. Sweetser and P. Wyeth, "Gameflow: A model for evaluating player enjoyment in games," *Comput. Entertain.*, vol. 3, no. 3, p. 3, jul 2005.

[41] G. B. Moneta, "On the measurement and conceptualization of flow," *Advances in flow research*, pp. 23–50, 2012.

[42] M. Bassi and A. Delle Fave, *Flow in the Context of Daily Experience Fluctuation*. Cham: Springer International Publishing, 2016, pp. 181–196.

[43] E. A. Boyle, T. M. Connolly, T. Hainey, and J. M. Boyle, "Engagement in digital entertainment games: A systematic review," *Computers in human behavior*, vol. 28, no. 3, pp. 771–780, 2012.

[44] A. Voida and S. Greenberg, "Wii all play: The console game as a computational meeting place," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2009, pp. 1559–1568.

[45] N. Ducheneaut, N. Yee, E. Nickell, and R. J. Moore, "" alone together?" exploring the social dynamics of massively multiplayer online games," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 407–416.

[46] M.-V. Aponte, G. Levieux, and S. Natkin, "Scaling the level of difficulty in single player video games," in *Entertainment Computing – ICEC 2009*, S. Natkin and J. Dupire, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 24–35.

[47] D. Wheat, M. Masek, C. P. Lam, and P. Hingston, "Modeling perceived difficulty in game levels," in *Proceedings of the Australasian Computer Science Week Multiconference*, ser. ACSW '16. New York, NY: Association for Computing Machinery, 2016.

[48] N. M. Diah, A. P. Sutiono, L. Zuo, N. Nossal, H. Iida, N. Azan, and M. Zin, "Quantifying engagement of video games: Pac-man and dota (defense of the ancients)," in *17th International Conference on Mathematical and Computational Methods in Science and Engineering (MACMESE15)*. Kuala Lumpur: WSEAS, 2015, pp. 49–55.

[49] X. Chen, L. Niu, A. Veeraraghavan, and A. Sabharwal, "Faceengage: Robust estimation of gameplay engagement from user-contributed (youtube) videos," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 651–665, 2019.

[50] A. Fortin-Côté, C. Chamberland, M. Parent, S. Tremblay, P. Jackson, N. Beaudoin-Gagnon, A. Campeau-Lecours, J. Bergeron-Boucher, and L. Lefebvre, "Predicting video game players' fun from physiological and behavioural data," in *Advances in Information and Communication Networks*. Cham: Springer International Publishing, 2019, pp. 479–495.

[51] D. Melhart, A. Azadvar, A. Canossa, A. Liapis, and G. N. Yannakakis, "Your gameplay says it all: Modelling motivation in tom clancy's the division," in *2019 IEEE Conference on Games (CoG)*. London, UK: IEEE, 2019, pp. 1–8.

[52] D. Reguera, P. Colomer-de Simón, I. Encinas, M. Sort, J. Wedekind, and M. Boguñá, "Quantifying human engagement into playful activities," *Scientific Reports*, vol. 10, no. 1, p. 4145, 2020.

[53] D. Melhart, D. Gravina, and G. N. Yannakakis, "Moment-to-moment engagement prediction through the eyes of the observer: Pubg streaming on twitch," in *Proceedings of the 15th International Conference on the Foundations of Digital Games*, ser. FDG '20. New York, NY, USA: Association for Computing Machinery, 2020.

[54] A. Denisova, A. I. Nordin, and P. Cairns, "The convergence of player experience questionnaires," in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. NY, USA: ACM, 2016, p. 33–37.

[55] K. Pinitas, D. Renaudie, M. Thomsen, M. Barthet, K. Makantasis, A. Liapis, and G. N. Yannakakis, "Predicting player engagement in tom clancy's the division 2: A multimodal approach via pixels and gamepad actions," in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23. NY, USA: ACM, 2023, p. 488–497.

[56] P. Mavromoustakos-Blom, D. Melhárt, A. Liapis, G. N. Yannakakis, S. Bakkes, and P. Spronck, "Multiplayer tension in the wild: A hearthstone case," in *Proceedings of the 18th International Conference on the Foundations of Digital Games*, ser. FDG '23. NY, USA: ACM, 2023.

[57] D. Rae Selvig and H. Schoenau-Fog, "Non-intrusive measurement of player engagement and emotions - real-time deep neural network analysis of facial expressions during game play," in *HCI in Games*, X. Fang, Ed. Cham: Springer International Publishing, 2020, pp. 330–349.

[58] H. Schoenau-Fog, "The player engagement process–an exploration of continuation desire in digital games," in *Proceedings of DiGRA 2011 Conference: Think Design Play*. Hilversum, The NL: Digital Games Research Association (DiGRA), 2011, p. 18.

[59] S. A. Mohammed, S. Shirmohammadi, and A. E. Alchalabi, "Network delay measurement with machine learning: From lab to real-world deployment," *IEEE Instrumentation Measurement Magazine*, vol. 25, no. 6, pp. 25–30, 2022.

[60] I. Slivar, L. Skorin-Kapov, and M. Suznjevic, "Qoe-aware resource allocation for multiple cloud gaming users sharing a bottleneck link," in *2019 22nd conference on innovation in clouds, internet and networks and workshops (ICIN)*. IEEE, 2019, pp. 118–123.

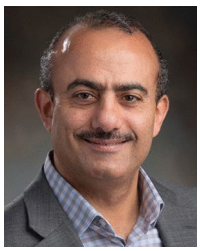
[61] K. Xiaohan, M. N. A. Khalid, and H. Iida, "Player satisfaction model and its implication to cultural change," *IEEE Access*, vol. 8, pp. 184 375–184 382, 2020.

Ammar Rashed received his Bachelor of Science degree in Computer Science and Engineering with the highest standing in his graduating class in 2018 from İstanbul Şehir University, Türkiye, and his Master of Science in Computer Science degree in 2020 from Özyeğin University, Türkiye. He is currently a scholarship-holding Computer Science PhD student at the University of Ottawa, Canada, doing research in machine learning-assisted player engagement measurement.

Shervin Shirmohammadi(M '04, SM '04, F '17) received his Ph.D. in Electrical Engineering in 2000 from the University of Ottawa, Canada, and after spending 3 years in the industry as a senior architect and project manager, joined as Assistant Professor the same University, where since 2012 he has been a Full Professor with the School of Electrical Engineering and Computer Science. He is Director of the Discover Laboratory, doing research in machine learning-assisted measurements, especially vision-based measurement, IoT measurements, and multimedia and network measurements. The results of his research, funded by more than $28 million from public and private sectors, have led to over 400 publications, over 80 researchers trained at the postdoctoral, PhD, and Master's levels, 30+ patents and technology transfers to the private sector, and four Best Paper awards. He is the Associate Editor-in-Chief of the IEEE Open Journal of Instrumentation and Measurement, for which he was the Founding Editor-in-Chief from 2021 to 2023, and was the Editor-in-Chief of the IEEE Transactions on Instrumentation and Measurement from 2017 to 2021, the Associate Editor-in-Chief of IEEE Instrumentation and Measurement Magazine in 2014 and 2015, and is currently on the latter's editorial board.

He has been on the Administrative Committee (AdCom) of the IEEE Instrumentation and Measurement Society (IMS) since 2014, currently serves as IMS's President, and was a member of the IEEE I^2MTC Board of Directors from 2014 to 2016.

Dr. Shirmohammadi is an IEEE Fellow "for contributions to multimedia systems and network measurements", and recipient of the 2019 George S. Glinski Award for Excellence in Research, the 2021 IEEE IMS Distinguished Service Award, and the 2023 IEEE IMS Technical Award "for contributions to the advancement of machine learning-assisted measurements".



Mohamed Hefeeda(Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Mansoura University,Mansoura, Egypt, in 1994 and 1997, respectively, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2004. He is a Professor in the School of Computing Science at Simon Fraser University (SFU), Burnaby, BC, Canada, where he served as the Director of the School between 2018 and 2023. He founded and leads the Network and Multimedia Systems Laboratory (http://nmsl.cs.sfu.ca) at SFU. His research interests include multimedia systems, mobile and wireless video streaming, immersive video processing and delivery, and network systems and protocols. He has authored or co-authored more than 150 papers and multiple granted patents. Dr. Hefeeda was the recipient of the prestigious NSERC Discovery Accelerator Supplements (DAS) awards in 2011, which is granted to a select group of distinguished researchers from all Science and Engineering disciplines in Canada. His research on mobile multimedia systems has resulted in multiple patents and conference awards (e.g., ACM MMSys Best Paper, ACM Multimedia Best Demo, and IEEE Innovation Best Paper), and has been featured in several news venues, including ACM Tech News, World Journal News, and CTV British Columbia. He has served on the editorial boards of premier journals, such as the ACM Transactions on Multimedia Computing, Communications and Applications (TOMM), where he was named the Best Associate Editor in 2014, and on the organization committees and/or Co/Chaired several international conferences, such as ACM MMSys, ACM MM, IEEE ICME, and ACM NOSSDAV.