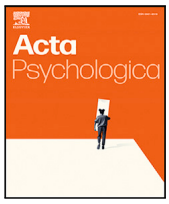
Acta Psychologica 267 (2026) 106947

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



# Interpretable behavioral clusters of gamblers through unsupervised learning

Mana Azizsoltani [a,*], Ismael Gomez-Talal [a,b,**], José Luis Rojo Alvarez [b], Kasra Ghaharian [a]

[a] International Gaming Institute at the University of Nevada, Las Vegas., 4505 S. Maryland Pkwy, Las Vegas, 89154, NV, USA
[b] Department of Signal Theory and Communications and Telematic Systems and Computation at Rey Juan Carlos University, Camino del Molino, s/n, Fuenlabrada, 28943, Madrid, Spain

ARTICLE INFO

*Keywords:*
Gambling disorder
Behavioral tracking data
Artificial intelligence
Machine learning
Unsupervised learning

ABSTRACT

Understanding the heterogeneity among highly involved gamblers is critical for the development of effective harm reduction strategies. This study employs unsupervised machine learning to segment a population of high-intensity Electronic Gambling Machine (EGM) users based on behavioral indicators derived from transactional data. Using a combination of Uniform Manifold Approximation and Projection (UMAP) for nonlinear dimensionality reduction and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for cluster identification, we performed a systematic grid search to optimize internal validity metrics (Silhouette = 0.5827, Davies–Bouldin Index = 0.4442, Calinski–Harabasz Index = 8907.46). The analysis yielded four well-separated behavioral clusters: one marked by impulsive withdrawals and night-time play; another showing consistent and high-frequency gambling; a third characterized by structured, high-stakes sessions; and a fourth exhibiting rapid, binge-like activity within short time windows. To facilitate interpretation, we trained cluster-wise random forest classifiers, identifying key discriminative features such as balance trajectory, inter-session timing, and variability in transaction intervals. Our findings demonstrate that high involvement is not a uniform construct, but rather encompasses diverse behavioral subtypes, each potentially associated with different levels of gambling-related harm. This segmentation framework offers practical implications for personalized responsible gambling initiatives and contributes to ongoing research advocating for data-driven player protection strategies.

## 1. Introduction

Over the past 20 years, innovations in information technology in the gambling industry have increased the availability and capabilities of data, particularly in the context of player behavior tracking (Chagas & Gomes, 2017; Deng et al., 2019). Simultaneously, the industry-wide adoption and implementation of artificial intelligence (AI) based solutions has allowed stakeholders to extract novel insights about players and leverage them for responsible gambling (Auer & Griffiths, 2023; Marionneau et al., 2025). The intersection of bet-by-bet transaction data and AI has been said to have the potential to solve the "casino AI puzzle" (Peister et al., 2024).

Technological innovations have been particularly transformative in the online gambling sector, largely due to the ease of collecting and tracking user behavior data, which has allowed operators to gain deeper insights into behavioral markers of gambling harm (Delfabbro et al., 2024; Deng et al., 2019). Although there is a substantial body of responsible gambling (RG) research in land-based settings (e.g., survey-, clinical-, and venue-based studies), land-based work remains largely underrepresented in the literature surrounding behavioral tracking data (BTD), particularly when compared to the online setting (Ghaharian, Abarbanel, Phung, et al., 2023). This is likely due to a variety of factors, including: the lack of mandatory carded play, the inability to track cash withdrawals, the challenge of tracking players geographically, and the inaccuracy of table games data (due to lack of technological infrastructure). Yet, new technological advances in data collection and storage have opened the door for seeking out data-driven insights for land-based gambling modalities (Peister et al., 2024). Before the advent of account-based play (Gainsbury, 2011), Electronic Gambling Machine (EGM) data was only available at the machine-level. But with the ability to obtain granular data for individual players, researchers can now obtain a deeper understanding of the behaviors associated with gambling harm for EGM players. These innovations have created novel, under-explored areas of research surrounding the harms of EGM play and opportunities for prevention.

* Corresponding author.
** Corresponding author at: Department of Signal Theory and Communications and Telematic Systems and Computation at Rey Juan Carlos University, Camino del Molino, s/n, Fuenlabrada, 28943, Madrid, Spain.
*E-mail addresses:* mana@diffgaming.com (M. Azizsoltani), ismael.gomez.talal@urjc.es (I. Gomez-Talal).

https://doi.org/10.1016/j.actpsy.2026.106947
Received 14 August 2025; Received in revised form 1 April 2026; Accepted 23 April 2026
Available online 16 May 2026

Furthermore, the combination of AI and granular BTD has shaped RG practice, as operators and regulators can rely on AI models to detect patron profiles that have a high risk of problem gambling. In parallel, these developments have also shaped regulatory and compliance discussions around the auditability of AI models, privacy of data, and transparency of analytics practices (Marionneau et al., 2025).

We seek to address the gap in the research by administering an explainable unsupervised machine learning (ML) methodology on a large dataset containing land-based EGM BTD from Poland to identify different behavioral profiles of highly engaged gamblers.

### 1.1. Electronic gambling machines and gambling harm

Slot machines, or more broadly EGMs, are one of the dominant segments of the gambling industry (American Gaming Association, 2024). In fact, slot machines accounted for over 50% of the total commercial gross gaming revenue (GGR) in the United States in 2023 (Tolan, 2024). But while this segment attracts a plethora of players and represents a significant proportion of casino revenue, it has also raised concerns about its association with impulse control disorders, such as gambling disorder (Mosquera & Keselj, 2017). Once referred to as "the crack cocaine of gambling", EGMs are considered to be one of the gambling modalities with the highest addictive potential and capacity for harm (Dowling et al., 2005). Prior research frequently identifies that the structural characteristics, rapid, continuous play cycles, and jackpot mechanisms of EGMS are significant contributors to the experience of gambling-related harms (Dowling et al., 2005; Rockloff & Hing, 2013; Schüll, 2012).

Behavioral research has found that multi-line betting, skill-based games, free spins, and sensory stimuli contribute to the risk of gambling harm that may arise from EGM play. In a study on EGM players in Australia, Woolley and Livingstone (2009) concluded that a combination of free spins and multi-line betting arrangements induced potentially high-risk gambling behaviors such as increased betting and time spent on the machine. Other studies have also found that machines with multi-line betting tend to be more immersive for players, which can lead to higher risk of gambling-related harms (Dixon et al., 2014; Murch & Clark, 2019). Pickering et al. (2020) found that skill-based games in the form of an EGM had high risks for harm as they caused higher levels of persistence within a gambling session, loss chasing, and initiating new gambling sessions. Furthermore, the sensory characteristics of the EGM (e.g.: graphics, sounds, and haptics) have been associated with risk of addiction (Delfabbro et al., 2005; Loba et al., 2001). In fact, biological research has shown that continuous exposure to the kind of stimulation obtained from EGMs can lead to problem gambling (Zack et al., 2020).

EGM play is also characterized by rapid and continuous play cycles, where the outcomes are given on an almost immediate basis (Harris & Griffiths, 2018; Landon et al., 2018). In fact, gamblers that seek treatment or intervention for gambling disorder disproportionately report the cause of their problem gambling to be rapid forms of gambling (such as EGMs) (Australian Productivity Commission et al., 2010; Griffiths & Meredith, 2009; Meyer et al., 2009). Empirical evidence shows that faster-paced games are more enjoyable and stimulating for gamblers (Delfabbro et al., 2005; Loba et al., 2001), but may increase the risk of losing control. Similarly, high-risk gambling behaviors such as increased time spent gambling, reduced ability to stop gambling, and increased wagering activity are associated with faster games both for problem gamblers and non-problem gamblers (Harris & Griffiths, 2018).

Additionally, some of the mechanisms of certain EGM games such as near misses and losses disguised as wins (LDWs) have been proven to be associated with the increased risk of harm. Near misses are events that are characterized by the appearance of being close to winning, which enhances the perceived reward and triggers similar brain regions associated with reinforcement and reward processing (Horton et al., 2006). They have been shown to significantly increase arousal, motivation, and frustration compared to regular losses, leading players to continue gambling despite losing outcomes (Barton et al., 2017). LDWs, on the other hand, occur when players win less than they bet but are still celebrated with winning sounds and animations, creating a cognitive distortion that can lead to an overestimation of actual winnings (Dixon et al., 2010). LDWs have been associated with increased session length, higher spending, and greater difficulty stopping play, thus exacerbating financial and psychological harms for gamblers (Barton et al., 2017). Both near misses and LDWs exploit the structural design of EGMs, increasing gambling intensity and contributing to problem gambling behaviors.

Lastly, the jackpot mechanisms may increase the risk factors of EGM play. The risks of harm surrounding EGM jackpots are derived from (1) that people are influenced by the potential of a jackpot rather than actually hitting the jackpot, and (2) the experience of winning a jackpot can lead to subsequent high-risk gambling behavior (Rockloff & Hing, 2013). In the context of the former, it has been found that jackpot-oriented machines are associated with greater spend (Browne et al., 2015) and that hidden or mystery jackpots contribute to higher gambling intensity (Donaldson et al., 2016). With respect to the latter, Young et al. (2008) discovered that high-risk EGM players who hit a significant win were more inclined to continue gambling compared to other individuals. Similarly, Wilkes et al. (2010) provided evidence that large wins often triggered physiological arousal, which has been linked to faster gambling, longer gambling sessions, and larger bets (Rockloff & Greer, 2010; Rockloff & Hing, 2013).

### 1.2. Identifying gambling harm using behavioral tracking data and clustering

In recent years, a stream of research has emerged surrounding the use of behavioral tracking data (BTD) to explore risk for gambling-related harm (Ghaharian, Abarbanel, Phung, et al., 2023). BTD allows gambling operators to monitor the interactions that players have with a gambling medium over time (Deng et al., 2019). Typically, such data contains account-level data containing user information (e.g., birthday, gender, address) as well as game data, such as amount wagered/won, spend limits, and length of sessions (Chagas & Gomes, 2017). Within the context of slot machine play, BTD refers to the detailed recording of every interaction a player has with a slot machine, from the moment and amount of initial deposit into the machine to the exact value, time, and outcome of each wager. The granularity of this data may allow operators to identify specific behaviors and patterns that can identify individuals at-risk of gambling-related harms and inform RG strategies and interventions.

There has been considerable research done to identify specific gambling behavioral profiles using BTD, where a number of studies have applied unsupervised machine learning clustering algorithms to a sample of gambler BTD to extract potential behavioral markers of harm (Ghaharian, Abarbanel, Phung, et al., 2023). These algorithms are particularly useful for RG research due to the lack of labeled data (i.e., problem gambler/non-problem gambler) or specific proxies for problem gambling arising from the financial and logistic challenges in obtaining problem gambling scores.

Nonetheless, the application of clustering algorithms to BTD has yielded various insights into the different profiles of gamblers and has managed to tease out different high-risk behaviors among gambler groups. The vast majority of the existing research is conducted on online gamblers, where gamblers are clustered on a variety of variables that are engineered as markers of gambling problem harm. These markers of harm generally revolve around gambling frequency, variability, intensity, amount, and trajectory, and other behavior-related metrics.

A large share of BTD studies has relied on k-means due to its simplicity and scalability, typically using engineered markers related to gambling frequency, intensity, variability, and trajectory. Across

these studies, a recurring pattern is the presence of a dominant low-involvement segment (often the majority of the sample), alongside one or more smaller segments characterized by elevated intensity and/or variability; however, the number of clusters and the interpretation of high-risk groups varies with the feature set and the gambling context.

For example, Braverman and Shaffer (2012) and Dragicevic et al. (2011) both applied k-means to online gamblers using related markers (frequency, variability, intensity, trajectory) and obtained four-cluster solutions, yet they differed in how many clusters were characterized as high-risk. Subsequent work illustrates further heterogeneity driven by feature engineering and domain context: Adami et al. (2013) incorporated a "sawtooth" marker and reported five clusters, while Suriadi et al. (2016) identified a larger number of clusters in a remote betting sample, with only a subset exhibiting problematic patterns. Similar k-means pipelines have also been applied to adjacent online contexts (e.g., daily fantasy sports) with clusters reflecting highly involved players with different outcome profiles (Wiley et al., 2020). Time-series representations can yield yet another typology: Peres et al. (2021) clustered sequences of wager outcomes and identified groups spanning regular to pathological patterns across products.

These studies serve as evidence that k-means can uncover meaningful segmentation structure in BTD, but the specific cluster labels and counts are study- and feature-dependent. In line with this, selected k-means after benchmarking against alternative algorithms, emphasizing that conclusions depend on both the clustering model and the underlying behavioral representation.

A smaller set of studies has employed alternative models either to increase interpretability (e.g., decision-tree segmentation) or to capture different distributional assumptions (e.g., probabilistic latent class approaches). For instance, CHAID-based analyses have been used to identify high-intensity/high-variability segments or expenditure-driven profiles when demographic covariates are included (Braverman & Shaffer, 2012; Chagas et al., 2021). Latent class analysis has been applied to lottery contexts combining traditional RG variables with behavioral dynamics (e.g., changes over time, loyalty bonuses), yielding multi-class typologies with only a fraction of players classified as moderate/high risk (Perrot et al., 2018). These approaches further reinforce that "clusters" in the literature are not homogeneous objects but rather modeling constructs shaped by data scope and methodological choices.

While previous works have applied an array of clustering algorithms over a wide range of contexts, clusters cannot be directly compared for a variety of reasons. This is because of (1) gambling product and channel (e.g., online casino, sports betting, lottery, payments), (2) unit of analysis (account-level vs. session-level vs. time-series), (3) feature engineering choices (frequency, intensity, variability, trajectory, and bespoke markers), and (4) model selection decisions (algorithm family, hyperparameters, and criteria used to choose the final solution). This is consistent with (Ghaharian, Abarbanel, Kraus, et al., 2023), who emphasize that conclusions depend both on the clustering model and the underlying behavioral representation.

Despite the breadth of online BTD work, EGM-focused studies remain limited, particularly for brick-and-mortar contexts. Excell et al. (2014) used land-based EGM tracking data with supervised prediction of problem gambling, leveraging PGSI-based labels and behavioral markers such as frequency, amount, and loss chasing. In contrast, Mosquera and Keselj (2017) applied clustering to session-level EGM data to characterize high-risk sessions rather than player-level profiles, which constrains inference about individual harm risk. Neither of these studies seek to understand the behavioral profiles of high risk EGM gamblers at the individual level.

## 2. The present study

Despite the growing body of work on behavioral markers of gambling-related harm, several gaps remain. First, while there is substantial RG research in land-based settings, player-level BTD for land-based EGMs remains underrepresented, largely due to structural constraints in data capture and linkage. Second, prior BTD-driven clustering studies typically segment the full player population, which can obscure heterogeneity within the most engaged subgroup that is most relevant for harm prevention. Third, the methodological toolkit used in the literature has been dominated by relatively simple partitioning approaches (e.g., k-means), with limited adoption of modern clustering strategies that are well-suited to the distributional properties of gambling BTD.

To address these gaps, we adopt a two-stage, unsupervised framework that combines targeted identification of a highly-involved subgroup and profiling within that subgroup using a UMAP+DBSCAN pipeline. This approach is advantageous for gambling BTD because it (1) does not require specifying the number of clusters a priori, (2) can recover non-convex cluster structure, and (3) is more robust to outliers and heterogeneous-density patterns that are common in heavy-tailed behavioral markers. Rather than focusing on limitations of any single baseline method, our emphasis is on selecting a clustering strategy aligned with the geometry of the data and the practical constraints of label scarcity in RG research.

Accordingly, this study performs a two-fold exploratory study using BTD from a large sample of EGM gamblers to describe gambling behavioral activity using statistical and machine learning techniques. First, this study discerns a subset of "highly-involved" gamblers and compares their behavior to the rest of the sample (i.e., non highly-involved gamblers), analyzing differences in both demographic and gambling behavior variables. Highly involved players are defined as those that are disproportionately more engaged (in terms of frequency or intensity) than other players. Second, using the highly-involved gambler group as a proxy for gamblers with a high risk of GD, this study identifies various distinct behavioral subgroups of high-risk gamblers within the group of highly-involved gamblers using advanced ML clustering algorithms. Specifically, we made the following exploratory propositions in relation to EGM gambler behavior.

### *2.1. RQ 1: Does a small subgroup of EGM gamblers account for a disproportionate share of gambling involvement (e.g., sessions, spend, and losses)?*

Because clinically validated harm labels (e.g., PGSI/DSM) are rarely available at scale in operator data, many studies adopt pragmatic splits that isolate the extreme tail of the distribution to create a high-engagement analytic subgroup for profiling or downstream modeling. Previous research that used gambling BTD data from online poker (Tom et al., 2022), online sports betting (Nelson et al., 2021), online casinos (Edson et al., 2022), and daily fantasy sports (Nelson et al., 2019) have found that a small minority of players accounts for a large share of overall activity (e.g., sessions, spend, or time played). Other studies using financial data have also found that these groups of "highly involved" gamblers (Ghaharian, Abarbanel, Kraus, et al., 2023; Zendle & Newall, 2024). These studies typically report between a 95/5 to 99/1 percentage split in terms of the percentage of gamblers that exhibited the disproportionately high gambling behaviors. These thresholds are heuristic and context-dependent; they provide a relative, sample-specific definition of extreme engagement and should not be interpreted as a clinically validated cut-off for gambling-related harm. While there has not been a previous study that analyzes this phenomenon for strictly EGM gamblers in a land-based environment, we hypothesize that our sample will contain a similar skew when it comes to the proportion of "highly-involved" gamblers.

### *2.2. RQ 2: How does the group of "highly-involved" EGM gamblers differ from the rest of the population in terms of demographics and gambling behaviors?*

Extensive research on problem gambling risk and prevalence has demonstrated that risk factors for problem gambling vary across gender

and age, specifically finding that younger individuals and males are at higher risk of problem gambling (Moreira et al., 2023). Accordingly, we hypothesize that the highly-involved gambler group will be composed of a significantly higher composition of males and be younger on average. Similarly, based on results from existing studies that found that the highly-involved gambling group exhibited discontinuously high gambling behaviors (LaBrie et al., 2007; Nelson et al., 2021), we expect to find that highly-involved EGM players also display significantly greater values across gambling behavioral variables.

### 2.3. RQ 3: Is it possible to distinguish subgroups of highly involved EGM gamblers? what do these subgroups look like in terms of demographics and gambling behaviors?

Previous studies that apply clustering algorithms to gambler BTD have almost exclusively conducted the clustering algorithm on the entire available sample using different markers of gambling harm to gain insights on sub-groups of gamblers, including those at a potential risk of gambling-related harm (Delfabbro et al., 2024). While these sort of analyses have yielded insights into different profiles of gamblers, they do not specifically explore the profiles of the at-risk population. Essentially, the clustering algorithm picks up on the noise of the entire dataset and leaves us ignorant to the more granular behavioral profiles of more engaged gamblers that may be at higher risk of gambling-related harm.

One study conducted by Pricewaterhouse Coopers (PWC) (2017) uses a two-phase clustering technique, where they cluster the general population, and then perform a sub-clustering of daily triggers for the problem gambler subgroup, which is determined using the PGSI. In their analysis, they were able to identify certain daily behavioral triggers that were distinctive to problem gamblers. Accordingly, while there has been no research done to sub-cluster problem gamblers at the player level, we hypothesize that there will be different underlying behavior profiles specific to the highly-involved gambler group.

## 3. Methodology

This section outlines the methodological framework employed to identify and characterize subgroups of highly-involved gamblers within a large-scale transactional dataset of Polish slot machine players. The approach integrates a comprehensive data processing pipeline, beginning with the engineering of behavioral features from raw transaction logs and demographic data. We define session boundaries, aggregate behavioral indicators, and conduct descriptive and inferential analyses to distinguish highly-involved gamblers from the general population. To uncover latent behavioral profiles within this group, we implement unsupervised machine learning techniques — specifically, UMAP for dimensionality reduction and DBSCAN for density-based clustering — followed by a supervised feature importance analysis using Random Forest classifiers. This multi-step methodology enables the extraction of interpretable patterns in gambling behavior and offers methodological innovations that go beyond traditional clustering approaches limited by dimensional constraints.

### 3.1. Data

This study used data from 168,536 Polish slot machine players over a 15-month period spanning between March 12, 2023 to June 3, 2024. The structure of the gambling market in Poland includes a government monopoly on certain segments (e.g., online casinos and slot machines outside casinos), while private operators may obtain licenses for land-based casinos and online betting (ICLG, 2025). Land-based casinos operate under a licensing system that restricts the number of available licenses. No universal player card system is mandated, though standard age and licensing requirements apply. Online gambling is partially liberalized, with online betting permitted via licensed operators, while online casinos remain under the state-run monopoly operated by Totalizator Sportowy (Stelmachowski et al., 2023).

Two large account-level and transaction-level datasets are used in this study and were provided by a casino supply company that offers cloud-based IoT services for data tracking. The transaction-level dataset includes information on player withdrawals and deposits at the machine level, such as the amount deposited or withdrawn, the use of cash versus promotional balance, and the balance of the player card. The account-level dataset was obtained from the player card and contains demographic data (gender and age) and location information (play site and machine number).

The data was donated by a casino supply company for research purposes under a data-sharing agreement after being fully anonymized. The authors were granted access to a de-identified, read-only snapshot delivered via secure AWS S3 transfer with time-limited credentials. No direct personal identifiers (e.g., names, addresses, government IDs) were included. Demographic attributes were limited to age and gender as recorded by the operator, and internal player references were replaced by random pseudonymous IDs generated by the provider prior to transfer. All files were stored on encrypted drives, and access was restricted to the research team in compliance with applicable data-protection standards.

To obtain our analytical sample, we created an inclusion criterion of players who made at least five gambling deposit transactions during the sample period, yielding $N = 61{,}631$. This is to ensure reliable feature engineering and filter out all of the sparse or one-off play.

### 3.2. Measures and descriptive statistics

In this study, we used the raw data to derive a set of candidate variables that are based on previous research in RG (Ghaharian, Abarbanel, Phung, et al., 2023). Specifically, we identified behavioral variables such as frequency, intensity, amount wagered, variability, trajectory of wager changes, deposit/withdrawal frequency, and demographic data (age and gender). In order to engineer session-level data, we used the 95th percentile of the time elapsed between transactions to set the threshold for the creation of a new session, or approximately 13 h. For a full list of the engineered variables and their respective definitions, see Table 1.

The data came at the transaction level. In order to obtain player-level data, the data was aggregated two times. First, the data was aggregated at the session-level. To conduct this aggregation, we set the cutoff threshold between sessions using the 95th percentile of the distribution of elapsed time between transactions. Once the data is aggregated at the session level, we then summarized the sessions at the player-level.

To understand the general gambling behaviors and demographic makeup of our sample, we conducted various descriptive analyses. First, we looked at the distributions across the demographic variables and calculated seven-point summary statistics for all of the behavioral variables. Next, we analyzed pairwise Spearman correlations to understand the relationships between the different behavioral variables.

### 3.3. Data analysis

To judge the skewness in the data and discern the existence of a highly-involved gambler group, we followed the methodology of previous literature by generating centile plots for debits, total debited amounts, and net balances across all gambling transactions (Nelson et al., 2021, 2019; Zendle & Newall, 2024). Similar to how scree plots help determine cutoffs in factor or cluster analysis, we identified points where the distribution showed disproportionate shifts as done in previous gambling studies (Ghaharian et al., 2025; Zendle & Newall, 2024). Based on these points, we categorized a group of "highly-involved gamblers".

**Table 1**
Engineered features: definitions and variable codes.

| Variable | Definition | Variable code |
|---|---|---|
| Average monthly number of sessions | Mean sessions per month over 15 months. | avg_n_sesh_mth |
| Average monthly win/loss | Mean net spend per month. | avg_WL_mth |
| Max number of monthly sessions | Highest sessions in a single month. | max_n_sesh_mth |
| Average weekly number of sessions | Mean sessions per week. | avg_n_sesh_wk |
| Average weekly W/L | Mean weekly net spend. | avg_WL_wk |
| Max number of weekly sessions | Max sessions in a week. | max_n_sesh_wk |
| Average daily W/L | Mean daily net spend. | avg_WL_day |
| Average number of txns per session | Mean deposit/withdrawals per session. | avg_txns_per_sesh |
| Average elapsed hours between sessions | Mean hours between sessions. | avg_hrs_btwn_sesh |
| Average time between transactions | Mean seconds between transactions. | avg_secs_btwn_txns |
| Average time between deposits | Mean seconds between deposits. | avg_secs_btwn_deps |
| Average time between withdrawals | Mean seconds between withdrawals. | avg_secs_btwn_wth |
| Average session duration | Mean session length (hours). | avg_session_duration |
| Average number of deposits per session | Mean deposits per session. | avg_n_dep_sesh |
| Average deposit amount per session | Mean deposited amount per session. | avg_dep_amnt |
| Total number of deposits | Total deposits made. | LTD_deps |
| Total deposit amount | Total amount deposited. | LTD_dep_amnt |
| Average withdrawal amount per session | Mean withdrawn amount per session. | avg_wth_amnt |
| Average number of withdrawals per session | Mean withdrawals per session. | avg_n_wth_sesh |
| Total number of withdrawals | Total withdrawals made. | LTD_n_wth |
| Total amount withdrawn | Total amount withdrawn. | LTD_wth_amnt |
| Std dev of withdrawal amounts | Variability in withdrawal amounts. | sd_wth |
| Std dev of deposit amounts | Variability in deposit amounts. | sd_dep |
| Std dev of session durations | Variability in session lengths. | sd_sesh_duration |
| Std dev of time between sessions | Variability in time between sessions. | sd_time_btwn_sesh |
| Std dev of time between deposits | Variability in deposit timing. | sd_time_btwn_deps |
| Total number of sessions | Total sessions over 15 months. | n_seshs |
| Average net win/loss per session | Mean net spend per session. | avg_WL_sesh |
| Total net win/loss | Total net spend. | avg_WL_mth |
| Average session deposit trajectory | Mean slope of deposit sequence per session. | avg_dep_traj |
| Average balance trajectory | Mean slope of account balance during session. | avg_bal_traj |
| Night time session percent | of sessions between 12:00 AM and 5:00 AM. | night_time_pcnt |

Then, to test whether the demographic and behavioral variables differed between the highly-involved group and the rest of the sample (i.e., "not highly-involved gamblers"), we conducted chi-squared tests and Kruskal–Wallis tests, respectively. To test whether the distributions of the demographic variables (salary bands, age bands, and gender) were the same across the two groups, we used a chi-square test of independence. With respect to the behavioral variables, we tested the differences of each variable between groups by running Kruskal–Wallis (KW) tests. Given the expected skewness in the data, KW tests were the best fit for non-parametric median comparisons.

### 3.4. Identifying subprofiles of highly involved gamblers

#### 3.4.1. Dimensionality reduction and clustering framework

After identifying the highly-involved gambling group, we used unsupervised ML models to identify patterns within the group of highly involved gamblers. Specifically, the analysis focused on identifying similar player behavior profiles across the highly-involved gamblers (determined to be $N_{HI} = 5373$) by reducing the dimensionality of the data to help visualize clusters of players in a 3D latent space, within which each point represents a player based on their behavior embeddings. To accomplish this, we employed Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for identifying different clusters.

UMAP is a non-linear dimensionality reduction technique that optimally conserves both local and global structures throughout the data (McInnes et al., 2018). Using UMAP for this task represents a methodological improvement over earlier clustering approaches, which often relied on eliminating correlated variables rather than embedding them in a lower-dimensional manifold. After dimensionality reduction, we applied DBSCAN to discover high-density subgroups in the latent space.

We adopt UMAP+DBSCAN because it addresses three recurring challenges in high-engagement gambling BTD: (1) high dimensionality and collinearity across engineered behavioral markers, (2) nonlinear similarity structure (players can be behaviorally close under complex combinations of intensity, variability, and timing), and (3) heterogeneous density with outliers driven by extreme-tail engagement (both in spend and in frequency). We note that other underutilized families offer complementary benefits: GMMs provide soft assignments and can model overlapping subtypes with different covariance structures; hierarchical clustering provides nested segmentations across resolutions; and spectral/graph-based clustering can separate groups defined by graph connectivity rather than spherical geometry. In this study, we adopt a density-based solution because it most directly matches the heavy-tailed, heterogeneous-density nature of high-engagement EGM behavior and supports robust discovery of small, non-spherical subprofiles in an unlabeled setting.

**Mathematical Formulation of UMAP and DBSCAN**

UMAP constructs a weighted graph of local fuzzy simplicial set relationships. The conditional probability of point $x_j$ being connected to $x_i$ is:

$$v_{j|i} = \exp\left(\frac{-d(x_i, x_j) - \rho_i}{\sigma_i}\right), \tag{1}$$

where $d(x_i, x_j)$ is the distance between points (e.g., cosine), $\rho_i$ is the distance to the nearest neighbor, and $\sigma_i$ is a scaling factor. These local similarities are symmetrized using a fuzzy union:

$$v_{ij} = v_{j|i} + v_{i|j} - v_{j|i}v_{i|j}. \tag{2}$$

In the low-dimensional space, the similarity between embeddings $y_i$ and $y_j$ is modeled as:

$$w_{ij} = \left(1 + a\|y_i - y_j\|^{2b}\right)^{-1}, \tag{3}$$

where $a \approx 1.929$ and $b \approx 0.7915$. The loss function minimized by UMAP is the cross-entropy:

$$C_{\text{UMAP}} = \sum_{i \neq j} \left[ v_{ij} \log \left( \frac{v_{ij}}{w_{ij}} \right) + (1 - v_{ij}) \log \left( \frac{1 - v_{ij}}{1 - w_{ij}} \right) \right]. \quad (4)$$

This formulation makes UMAP scalable to large datasets and effective in preserving global structure. Once the embeddings are generated, we applied DBSCAN (Schubert et al., 2017), a density-based clustering algorithm that defines clusters as contiguous regions of high point density. For a given point $p$, the $\varepsilon$-neighborhood is:

$$N_\varepsilon(p) = \{ q \in D \mid \text{dist}(p, q) \leq \varepsilon \}. \quad (5)$$

A point is considered a core point if $|N_\varepsilon(p)| \geq \text{minPts}$. Clusters are formed by connecting density-reachable core points, while isolated points in low-density regions are labeled as noise. Grid-based approximations and indexing techniques (e.g., k-d trees, LSH, or cover-trees) may accelerate DBSCAN, but its flexibility to support arbitrary distance functions — such as cosine or geodesic distance — is a key advantage, especially for behavioral datasets like ours (Schubert et al., 2017).

Once the clusters were identified, we further investigated the characteristics of each group to derive insights into player behavior. This analysis included examining the engineered RG features such as frequency, average bet size, wager inclination, and potential risky betting behaviors (e.g., loss chasing or progressive betting). We validated the quality of the clusters using three well-established internal evaluation metrics suitable for unsupervised learning: the Silhouette Score, the Davies–Bouldin Index (DBI), the Calinski–Harabasz Index (CHI). These scores were used to capture different elements of cluster quality; the silhoutte score provides an interpretable summary of within-cluster cohesion relative to between-cluster separation, the DBI emphasizes average similarity of clusters and penalizes overlap, and the CHI measures the separation and compactness of the clusters from a variance-ratio.

UMAP is designed to preserve local neighborhood relationships and does not necessarily preserve global distances or densities. Therefore, to avoid evaluating the "clusterability" of the embedding rather than of the underlying behavioral space, we use the 3D UMAP representation strictly for visualization and qualitative interpretation of the resulting profiles. All internal cluster validity indices reported in this study (SSM, DBI, and CHI) are computed in the standardized original feature space (i.e., before UMAP), using the same feature set used for clustering.

#### 3.4.2. Feature importance by cluster

We utilize a supervised interpretation step as a proxy to determine which behaviors most distinguish each cluster. For each identified cluster, we define a binary classification task (cluster vs. rest) and train a Random Forest classifier using the full engineered feature set. We emphasize that this step is used to rank discriminative markers of cluster membership (i.e., which variables best separate one subgroup from the remainder of the highly involved population), and should not be interpreted as identifying causal determinants of gambling harm. Because cluster-vs-rest settings can be substantially imbalanced, we mitigate instability by using class-balanced training (e.g., class weights and stratified resampling) and we evaluate the classifier using out-of-bag (OOB) predictions (or a held-out split) to avoid overly optimistic separation. In addition, we acknowledge that impurity-based (Gini) feature importance can be biased in the presence of correlated predictors and may distribute importance unevenly across collinear features. Therefore, alongside the Gini ranking, we compute permutation importance on OOB/held-out predictions as a robustness check. The main conclusions are based on features that are consistently ranked highly across importance definitions, increasing confidence that the reported markers reflect stable, discriminative behavioral signatures rather than artifacts of impurity bias.

To further reduce redundancy induced by collinearity, we interpret importance rankings at the level of feature families (frequency, intensity, variability, and timing/regularity) and, where appropriate, we group highly correlated variables (e.g., $|\rho| > 0.8$) when summarizing the top discriminators for each cluster. This avoids over-emphasizing multiple near-duplicates of the same behavioral construct.

This feature importance methodology represents another key contribution to the literature. In previous studies, the importance of a particular set of variables could only be understood for whichever variables were included in the k-means clustering algorithm (typically 5-7), whereas our proposed approach allows feature importances to be extracted from all of the features. This gives a more holistic and unconstrained view of the features that characterize a particular behavioral profile.

### 3.5. Methodological framework diagram

Fig. 1 presents a comprehensive visual summary of the methodological framework implemented in this study. The process begins with the ingestion of raw transactional and demographic data from slot machine players, encompassing detailed records of deposits, withdrawals, session times, and basic user attributes such as age and gender. This data is then subjected to a structured preprocessing stage, where relevant features are engineered and aggregated at both the session and player levels using established criteria from prior literature.

Following this, a filtering step is introduced to isolate a subset of "highly involved gamblers", defined based on intensity, frequency, and volume metrics derived from centile plots. Once this group is identified, dimensionality reduction is performed using UMAP, which embeds each player into a latent behavioral space that preserves both local and global structure.

To uncover latent behavioral typologies, DBSCAN is applied in the embedded space. Unlike traditional clustering techniques, this method enables the identification of irregular cluster shapes and is robust to noise—both of which are important properties in behavioral datasets with skewed distributions. After clusters are detected, a supervised learning approach is employed to compute feature importance within each cluster using Random Forest classifiers. These models allow for the identification of the most discriminative behavioral characteristics associated with each group.

The final step involves interpreting the derived clusters to construct a set of behavioral subprofiles. These subprofiles reflect differences in gambling patterns such as session frequency, average amounts wagered, transaction variability, and indicators of risky behavior (e.g., loss chasing).

## 4. Results

This section presents the results of our multi-step analysis aimed at understanding gambling behavior patterns and identifying subgroups of highly involved gamblers. We begin by describing the demographic and behavioral characteristics of the overall sample, highlighting substantial variation and heavy skewness in financial metrics. Using centile plots and non-parametric statistical tests, we define and validate a subgroup of highly involved players based on intensity, frequency, and volume of gambling. We then compare this group to the rest of the population, revealing significant differences in demographics and behavioral patterns. Finally, we apply unsupervised clustering techniques to the highly involved group, identifying distinct behavioral subprofiles using UMAP and DBSCAN. These clusters are interpreted using feature importance rankings derived from supervised models, offering new insights into different forms of high-risk gambling behavior.

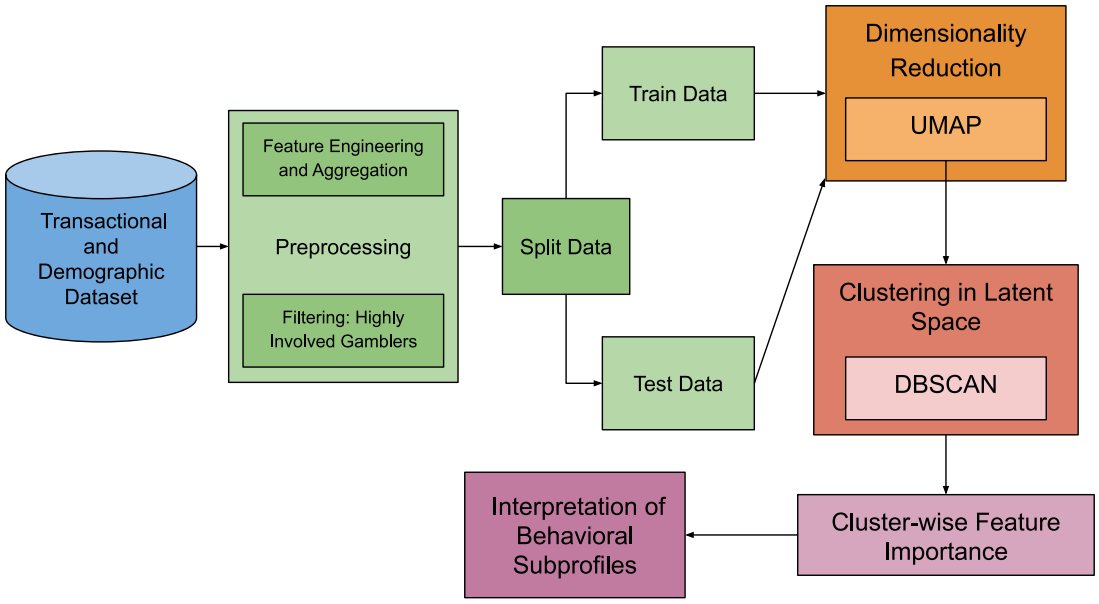



**Fig. 1.** Overview of the methodological pipeline used to identify and interpret behavioral clusters among highly involved gamblers.

### 4.1. Descriptive statistics

As displayed in Table 2, the sample was predominantly male (91%), with females comprising a mere 9% of the analytical sample. Age distribution was skewed toward younger players, with about 60% of players under the age of 40 and very few players over the age of 65.

Across the sample, there was substantial variation in gambling behavior, with many transactional variables exhibiting right-skewed distributions, where means exceeded medians. The typical player engaged in gambling infrequently, with a median of five monthly sessions (mean = 6), though some played as many as 34 times per month. The median player had a net spend averaging 5.5k PLN (≈USD $1.4k)per session, but extreme values ranged from ≈1.74M PLN (≈USD $ −440k) to 3.96M PLN (≈USD $1M). Additionally, deposit and withdrawal behaviors varied considerably. The median player made six deposits per session (mean = 10) with a median deposit amount of 4.2k PLN (≈USD $1k), but some players deposited amounts exceeding 1.58M PLN (≈USD $400k) in a single session. Similarly, the median player made two withdrawals per session (mean = 3), with a median withdrawal amount of 8k PLN (≈USD $2.1k), though the highest recorded withdrawal exceeded 2.5M PLN (≈USD $632k). These results are summarized in Table 3.

Session and transaction timing also differed widely across users. The median time elapsed between gambling sessions was 356 h (mean = 693), but some users had gaps as long as 8279 h, suggesting highly irregular engagement patterns. However, the standard deviation in session gaps was substantial (1.6k hours), reflecting diverse play patterns.

Deposit trajectories had a median of −8, with a wide range from −135k to 127k, indicating substantial variability among players. Balance trajectories were generally positive, with a median of 448 and a maximum balance of 1.97M. High kurtosis values for both metrics reflect the presence of a small number of players with extreme trajectories. There were also high kurtosis values across financial and behavioral metrics, indicating that a small subset of players also accounts for a disproportionate share of gambling frequency, intensity, and spending.

### 4.2. Highly involved gamblers — research question 1

Fig. 2 displays the centile plots for each of the following three variables: number of debit transactions, total amount debited, and average net win/loss per session.

**Table 2**
Demographic distribution.

| Variable | Percentage |
|---|---|
| **Gender** | |
| Female | 9% |
| Male | 91% |
| **Age Band** | |
| 18–30 | 25% |
| 30–40 | 33% |
| 40–50 | 21% |
| 50–65 | 16% |
| 65+ | 5% |

Supporting H1, we observe a subgroup of users who exhibited disproportionately higher gambling activity across each of the three frequency and amount variables. According to the percentile plots, it appeared that the point of discontinuity for the frequency variable (total number of sessions) occurred at the top 1% mark, possibly suggesting that there were more casual or leisure gamblers. For the volume and intensity variables, the point of discontinuity did not come at the top 1% mark, but rather at the top 3% mark. In the case of the average net loss per session, we also found that the bottom 1% of players were found to be discontinuous with the rest of the population due to their big wins.

### 4.3. Demographic and behavioral differences — research question 2

Using the findings from the previous research question, we defined highly-involved players to be players that were in the top 3% of total debited amount (volume), total number of sessions (frequency), or net losses (intensity). This definition resulted in a highly-involved player group that consisted of 5373 players and a not highly-involved player group consisting of 57,258 players.

To understand the different demographic composition of the two groups, we conducted chi-square tests of independence to see if there was a significant difference between the two groups. A chi-square test revealed a significant difference in the age distribution between highly involved and not highly involved gamblers ($\chi^2 = 1936.46, p < .0001$), indicating that older individuals were more likely to be highly involved, whereas younger individuals were less likely to fall into the highly involved subgroup. The distribution of age across the two groups is shown

**Table 3**
Seven-point summary of gambling behavioral variables.

| Variable | Min | Q1 | Median | Mean | Q3 | Max | SD | Kurtosis |
|---|---|---|---|---|---|---|---|---|
| Average monthly number of sessions | 1 | 3 | 5 | 6 | 8 | 34 | 4 | 2 |
| Average monthly W/L | −17 416.7 | 1749 | 5582 | 13 742 | 14 075 | 3 965 000 | 44 894 | 1416 |
| Max number of monthly sessions | 1 | 4 | 7 | 9 | 12 | 44 | 7 | 1 |
| Average weekly number of sessions | 1 | 2 | 2 | 2 | 3 | 8 | 1 | 1 |
| Average weekly W/L | −17 416.7 | 1749 | 5582 | 13 742 | 14 075 | 3 965 000 | 44 894 | 1416 |
| Max number of weekly sessions | 1 | 2 | 4 | 4 | 5 | 13 | 2 | 0 |
| Average daily W/L | −17 416.7 | 1750 | 5597 | 13 766 | 14 111 | 3 965 000 | 44 862 | 1428 |
| Average number of txns per session | 1 | 5 | 8 | 12 | 14 | 473 | 15 | 116 |
| Average elapsed time between sessions (hrs) | 12 | 149 | 356 | 693 | 864 | 8279 | 890 | 9 |
| Average time between transactions (hrs) | 0.06 | 27.16 | 75.64 | 191.09 | 205.12 | 6668.96 | 336.77 | 38 |
| Average time between deposits (hrs) | 0.06 | 36.49 | 100.85 | 247.89 | 270.98 | 17 876.79 | 422.52 | 74 |
| Average time between withdrawals (hrs) | 0 | 0.15 | 0.25 | 3.26 | 0.44 | 5170.25 | 62.98 | 3150 |
| Average session duration (hrs) | 0 | 1 | 1 | 1 | 2 | 19 | 1 | 7 |
| Average number of deposits per session | 0 | 3 | 6 | 10 | 11 | 454 | 12 | 148 |
| Average deposit amount per session | 12 | 1650 | 4202 | 7760 | 8711 | 1 583 458 | 17 865 | 1490 |
| Total number of deposits | 0 | 51 | 138 | 667 | 457 | 110 069 | 2251 | 333 |
| Total deposit amount | 0 | 155 754 | 538 485 | 3 654 247 | 2 168 978 | 141 262.559 | 17 348 161 | 1760 |
| Average number of withdrawals per session | 0 | 1 | 2 | 3 | 3 | 214 | 3 | 348 |
| Average withdrawal amount per session | 1 | 3685 | 8279 | 18 954 | 18 728 | 2 527 040 | 45 659 | 529 |
| Total number of withdrawals | 0 | 14 | 40 | 210 | 137 | 80 635 | 893 | 2283 |
| Total amount withdrawn | 0 | 94 200 | 369 060 | 2 937 914 | 1 606 163 | 128 873.079 | 15 097 363 | 1825 |
| Std dev of withdrawals | 0 | 2385 | 5972 | 14 528 | 14 275 | 1 329 716 | 34 557 | 277 |
| Std dev of deposits | 0 | 591 | 1665 | 3416 | 3329 | 516 696 | 9431 | 747 |
| Std dev of session duration | 0 | 1 | 1 | 1 | 2 | 20 | 1 | 11 |
| Std dev of elapsed time between sessions | 1 | 250 | 672 | 1258 | 1598 | 16 983 | 1621 | 10 |
| Std dev of time between deposits | 1 | 193 036 | 475 527 | 1 048 208 | 1 204 984 | 52 224 164 | 1 657 320 | 53 |
| Total number of sessions | 6 | 10 | 20 | 54 | 56 | 1170 | 87 | 20 |
| Average W/L per session | −17 416.7 | 1749 | 5582 | 13 742 | 14 075 | 3 965 000 | 44 894 | 1416 |
| Net W/L | −104 500.0 | 28 988 | 128 987 | 716 333 | 523 437 | 315 310.90 | 2 903 981 | 2802 |
| Average deposit trajectory | −135629 | −219 | −8 | 33 | 123 | 127 227 | 2235 | 1001 |
| Average balance trajectory | −50000 | −49 | 448 | 2425 | 2195 | 1 977 479 | 13 404 | 9150 |

*Note:* Positive deposit trajectory means that the amount of the deposits trend up over time, negative deposit trajectory means that the deposits trend down over time. Positive balance trajectory means that the balance trends up over time, while negative balance trajectory menas that the balance trends down over time.

**Table 4**
Age band distribution of highly involved and not highly involved gamblers.

| Age band | 18-30 | 30-40 | 40-50 | 50-65 | 65+ |
|---|---|---|---|---|---|
| Not highly involved | 26.82% | 33.54% | 20.38% | 15.05% | 4.20% |
| Highly involved | 6.81% | 26.82% | 27.84% | 27.41% | 11.11% |

**Table 5**
Gender distribution of highly involved and not highly involved gamblers.

| Gender | Female | Male |
|---|---|---|
| Not highly involved | 9.29% | 90.71% |
| Highly involved | 10.81% | 89.19% |

in Table 4. In contrast, Table 5 shows the gender distribution across the two groups, which remained relatively stable, with no substantial shift in composition ($\chi^2 = 13.22, p = .0003$). This suggests that gambling involvement is not meaningfully associated with gender.

We also conducted Kruskal–Wallis (KW) tests between the highly involved and not highly involved groups across a variety of behavioral variables, which we display in Table 6. Significant differences were observed across all metrics. Highly involved gamblers engaged in more frequent sessions, with a median of 11.38 monthly sessions (vs. 4.33), and their session duration was nearly twice as long (2.26 h vs. 1.12 h). Median win/loss values were substantially higher: 105k PLN (≈USD $26.7k) vs. 26.2k PLN (≈USD $6.6k), a pattern that persisted across weekly and per-session metrics. Deposits and withdrawals were more frequent and larger in volume: total deposit amounts of 15.1M PLN (≈USD $3.8M) vs. 434.7k PLN (≈USD $110k), and total withdrawals of 11.3M PLN (≈USD $2.9M) vs. 297.3k PLN (≈USD $75k). The median deposit per session was higher (9.9k PLN ≈USD $2.5k) and the median withdrawal was 22,147.59 PLN (≈USD $5.6k) compared to 3.9k PLN (≈USD $1k) and 7.7k PLN (≈USD $1.9k), respectively.

Highly involved players also made more transactions per session and had shorter gaps gambling: elapsed time between sessions was 100.39 h vs. 388.91 h, and between deposits was 22.15 h vs. 114.23 h. Variability was greater among highly involved gamblers for withdrawals, deposits, and session durations, though session gap variability was lower (SD = 246.48 h vs. 727.45 h), indicating more consistent engagement. Finally, deposit trajectory shifted from −10.1 in the less involved group to + 11.39 in the highly involved group, and balance trajectory was nearly twice as high (822 vs. 420). This means that the highly involved players increased their amounts wagered over time, whereas the less involved players decreased the amounts of their wagers over time. These findings suggest highly involved gamblers engage more often and intensely, with larger, more variable financial transactions and indications of loss-chasing behaviors.

### 4.4. Clustering of the highly involved gambler group — research question 3

To evaluate the performance of the different clustering configurations, we conducted an extensive grid search over the hyperparameters of the UMAP dimensionality reduction and the DBSCAN clustering algorithm on the group of 5373 highly-involved players. Specifically, we varied the number of neighbors (`n_neighbors`) and the minimum distance parameter of UMAP (`min_dist`) as well as the distance metric (`metric`), the epsilon neighborhood radius (`eps`), and the minimum number of points required to form a dense region (`min_samples`) of the DBSCAN algorithm.

Table 8 presents the results of this grid search, showing the number of clusters and noise points detected, along with three internal

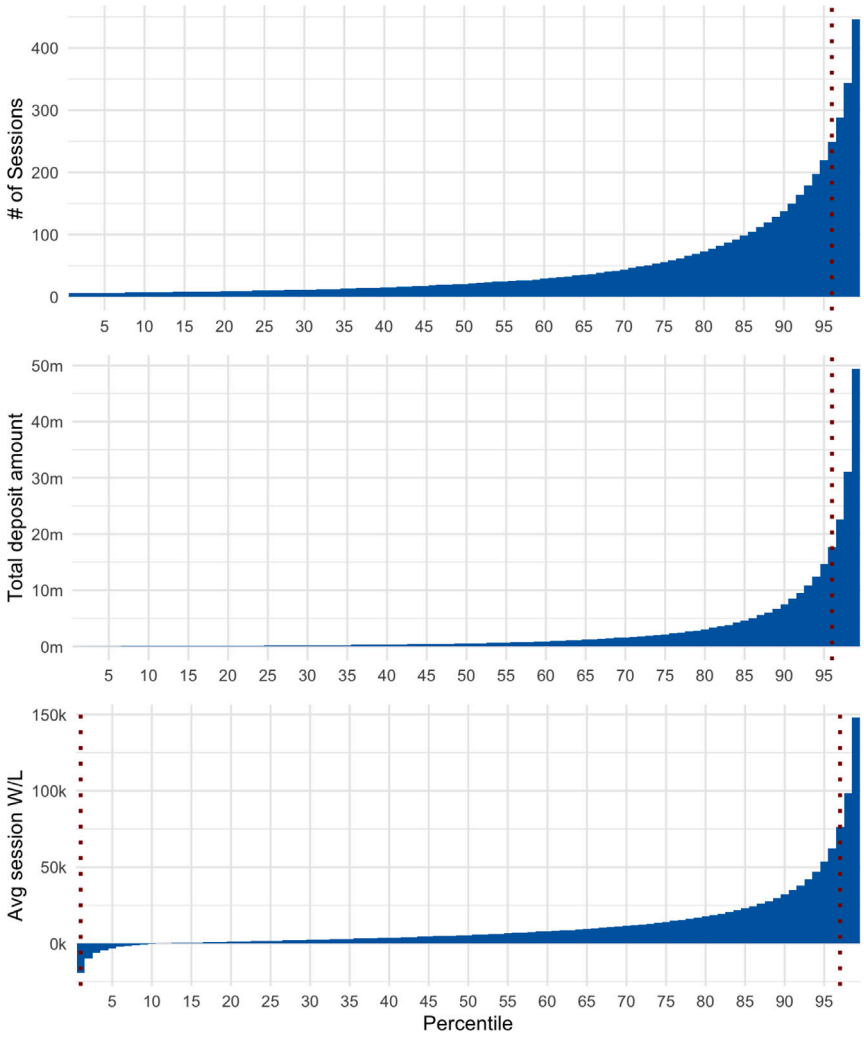



**Fig. 2.** Example of the cross validation procedure.

validation metrics and bootstrapped confidence intervals: Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. These metrics serve to quantify the quality of the clustering in terms of cohesion and separation.

The best configuration in terms of Silhouette Score is highlighted in bold and corresponds to the setting with `n_neighbors=5`, `min_dist=0`, `metric=cosine`, `eps=0.6`, and `min_samples=3`, achieving a silhouette of **0.5827**. For DBI, the lowest (i.e., best) value of **0.4442** was obtained under the same configuration, reinforcing the consistency of this result. Meanwhile, the highest CHI of **9097.37** was reached using `n_neighbors=10` and `min_dist=0`, also with a `cosine` metric.

These findings highlight the importance of tuning both dimensionality reduction and clustering parameters simultaneously. In particular, cosine distance and a low `min_dist` in UMAP, combined with moderate `eps` values in DBSCAN, yield better clustering structures in this dataset.

In addition to the DBSCAN-focused grid search, we benchmarked our final clustering choice against widely used *unsupervised* alternatives on the same UMAP embedding (e.g., k-means, Gaussian Mixture Models, and agglomerative clustering). Table 9 summarizes the comparative internal validity results using the same criteria (Silhouette, Davies–Bouldin, and Calinski–Harabasz), enabling a consistent assessment across methods.

We selected DBSCAN as the primary clustering algorithm because it aligns well with the geometry and distributional properties of highly-involved gambling markers, which often exhibit heavy tails, local concentration, and heterogeneous density patterns. Unlike centroid-based approaches (e.g., k-means) that implicitly favor spherical clusters and require specifying the number of clusters *a priori*, DBSCAN identifies clusters as dense regions separated by sparse areas, allowing (1) *non-convex* cluster shapes, (2) *data-driven* determination of the number of clusters, and (3) robustness to outliers and low-density behavioral profiles. These properties are particularly desirable in behavioral typology settings, where rare but meaningful player patterns may form small dense pockets rather than globally separable, convex groups.

Finally, we emphasize that internal indices are used as *guidance* rather than an absolute optimization target; therefore, the final choice also considers robustness and interpretability of the resulting subprofiles in downstream analyses (feature characterization and cluster-wise discrimination models).

Once the optimal clustering configuration was identified, we proceeded to interpret the resulting clusters by analyzing the features that most contributed to the formation of each group. Fig. 3 displays the cluster mapping over three UMAP dimensions and Table 10 summarizes the top 10 most important features for each cluster obtained through UMAP and DBSCAN, excluding variables that were either redundant or less informative for behavioral profiling (specifically, the standard deviation of withdrawal amounts, average daily win/loss, and the total number of withdrawals).

**Table 6**
Comparison of behavioral variables between highly involved and not highly involved gamblers.

| Comparison variable | Not highly involved | Highly involved | Significance |
|---|---|---|---|
| Average monthly number of sessions | 4.33 | 11.38 | * |
| Average monthly W/L | 5113.07 | 26 264.97 | * |
| Max number of monthly sessions | 6 | 19 | * |
| Average weekly number of sessions | 2 | 3.5 | * |
| Average weekly W/L | 5113.07 | 26 264.97 | * |
| Max number of weekly sessions | 3 | 6 | * |
| Average daily W/L | 5124.85 | 26 399.34 | * |
| Average number of transactions per session | 7.64 | 16 | * |
| Average elapsed hours between sessions | 388.91 | 100.39 | * |
| Average time between transactions | 308 663.01 | 58 279.08 | * |
| Average time between deposits | 411 209.84 | 79 752.03 | * |
| Average time between withdrawals | 888.98 | 1069.65 | * |
| Average session duration | 1.12 | 2.26 | * |
| Average number of deposits per session | 5.67 | 12.04 | * |
| Average deposit amount per session | 3912.17 | 9886.97 | * |
| Total number of deposits | 121 | 1750 | * |
| Total deposit amount | 434 717.5 | 15 104 260 | * |
| Average number of withdrawals per session | 1.75 | 3.62 | * |
| Average withdrawal amount per session | 7720.55 | 22 147.59 | * |
| Total number of withdrawals | 35 | 583 | * |
| Total amount withdrawn | 297 389.5 | 11 307 238 | * |
| Std dev of withdrawals | 5488.63 | 20 957.45 | * |
| Std dev of deposits | 1542.3 | 4896.73 | * |
| Std dev of session duration | 1.25 | 2.06 | * |
| Std dev of elapsed time between sessions | 727.45 | 246.48 | * |
| Std dev of time between deposits | 524 526.61 | 130 891.61 | * |
| Total number of sessions | 19 | 211 | * |
| Average W/L per session | 5113.07 | 26 264.97 | * |
| Net W/L | 105 695.5 | 2 692 870 | * |
| Average deposit trajectory | −10.1 | 11.39 | * |
| Average balance trajectory | 420.95 | 822.29 | * |

The feature importance was computed using a cluster-wise supervised approach, where for each cluster a binary classification model (Random Forest) was trained to distinguish members of that cluster from all others. We opted for a tree-based method due to its computational efficiency as well as its intuitive interpretability (James et al., 2023). The resulting Gini importance values are cluster-specific and indicate which behavioral variables are most characteristic within each group.

To further understand the behavioral composition of the highly involved gambler group, we examined the most relevant features driving the formation of each cluster. Table 7 shows a five-point summary of the most important variables across the different clusters. The resulting clusters highlight four distinct subgroups of highly-involved players with unique gambling patterns:

**Cluster 0** (n = 3317) consists of frequent players, characterized by the highest mean number of sessions (332) and substantial session engagement metrics. Players in this cluster average around 15 sessions per month and nearly 24 sessions per month at maximum, indicating a consistent play pattern over time. Average win/loss per session stands at approximately PLN 19.4k (USD $4.8k), reflecting moderate stakes compared to other clusters. The high deposit amounts (mean of PLN 5776 or USD $1.4k) further underscore their active engagement. Feature importance rankings highlight variables like `n_seshs`, `avg_n_sesh_mth`, and `avg_txns_per_sesh`, aligning with their sustained gambling behavior. This cluster represents steady, loyal patrons who engage frequently but may not exhibit the highest volatility in spend per session.

**Cluster 1** (n = 2031) includes high-stakes but moderate-frequency players. With an average of 41 sessions, they engage less frequently than Cluster 0 but exhibit the highest mean win/loss per session at over PLN 147.5k (USD $36.9k). Despite having fewer sessions (averaging just 5 per month), these players appear to stake larger amounts during each play session, reflected in moderate mean deposit amounts (PLN 687 or USD $175). Feature importance rankings emphasize session count, monthly sessions, and win/loss per session, supporting the profile of players who are less frequent but more financially engaged per session. This cluster could represent VIPs or high-rollers who bet larger amounts when they do participate.

**Cluster 2** (n = 18) represents a small group of low-frequency, high-stakes players. They have the fewest sessions on average (11 overall) but an average win/loss per session of over PLN 133k (USD $33k), indicating high-risk, high-reward gambling. These players likely appear for special events or occasional substantial betting activities. Mean deposit amounts are relatively low (PLN 94 or USD $24), suggesting that big wins or losses might not correspond with consistent deposits. Variability in play is high, both in session frequency and wager amounts, marking them as unpredictable but financially significant. Feature importance rankings focus on session counts and win/loss per session, aligning with infrequent but intense betting activity. This cluster likely includes special-event visitors or occasional high-rollers.

**Cluster -1** (n = 7) comprises of a small set of outlier players. They show moderate session counts (22 sessions), but the highest win/loss per session overall at nearly PLN 138k (USD $34.5k), coupled with lower deposit levels (PLN 569 or USD $142). These players exhibit highly variable playing patterns, with wide swings in their wager amounts and session frequencies. Although the low feature importance scores suggest that they may be noise or outliers, the players are particularly interesting because they appear more similar to Cluster 1 than to Cluster 2. This proximity hints that they might represent players who have branched off from more typical high-stakes play into more erratic or volatile patterns. The combination of low feature importance but behavioral similarities to Cluster 1 suggests they could reflect players transitioning between normal and outlier gambling behavior.

## 5. Discussion

This study analyzed a large dataset of EGM gamblers from Poland. We found that a small subset of players accounted for a disproportionate share of gambling frequency, deposits, and expenditures, reinforcing prior research across various gamling modalities (Edson et al., 2022; Nelson et al., 2021; Wiley et al., 2020; Zendle & Newall, 2024). While prior work often links younger adults and males to higher

**Table 7**
Five-Point Summary by Cluster. The table presents minimum, first quartile (Q1), median, third quartile (Q3), and maximum values for key gambling engagement metrics by cluster.

| Cluster | Variable | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|
| Cluster -1 | | | | | | |
| | n_seshs | 10 | 14.5 | 21 | 28 | 39 |
| | max_n_sesh_wk | 2 | 2 | 3 | 3 | 5 |
| | avg_n_sesh_mth | 1.6 | 2.3 | 3.15 | 3.33 | 5.63 |
| | max_n_sesh_mth | 3 | 3.5 | 4 | 5.5 | 8 |
| | LTD_deps | 149 | 311.5 | 479 | 768 | 1195 |
| | avg_txns_per_sesh | 14.43 | 16.8 | 47.05 | 54.33 | 64.15 |
| | avg_WL_sesh | 87 052.63 | 101 019.75 | 105 748.5 | 183 663.41 | 206 499.29 |
| | LTD_n_wth | 24 | 76 | 116 | 415.5 | 917 |
| | sd_sesh_duration | 0.83 | 1.53 | 1.76 | 2.47 | 2.69 |
| Cluster 0 | | | | | | |
| | n_seshs | 6 | 249 | 306 | 406 | 1170 |
| | max_n_sesh_wk | 1 | 6 | 7 | 7 | 13 |
| | avg_n_sesh_mth | 1.33 | 12.03 | 15.08 | 18.57 | 33.85 |
| | max_n_sesh_mth | 2 | 20 | 24 | 27 | 43 |
| | LTD_deps | 122 | 1730 | 3365 | 6586 | 110 069 |
| | avg_txns_per_sesh | 1.29 | 8.85 | 15.36 | 28.68 | 464.8 |
| | avg_WL_sesh | −237767.22 | 3425.71 | 9149.07 | 23 487.6 | 515 054.4 |
| | LTD_n_wth | 17 | 615 | 1117 | 2124 | 80 635 |
| | sd_sesh_duration | 0.18 | 1.67 | 2.26 | 2.84 | 18.97 |
| Cluster 1 | | | | | | |
| | n_seshs | 6 | 10 | 22 | 58 | 280 |
| | max_n_sesh_wk | 1 | 2 | 4 | 5 | 8 |
| | avg_n_sesh_mth | 1 | 2.78 | 4.33 | 6.79 | 17.78 |
| | max_n_sesh_mth | 1 | 4 | 7 | 11 | 27 |
| | LTD_deps | 10 | 126.5 | 331 | 853 | 9900 |
| | avg_txns_per_sesh | 1.48 | 10.33 | 16.91 | 27.8 | 183 |
| | avg_WL_sesh | −280600 | 81 850.24 | 106 069.52 | 160 519.03 | 3965000 |
| | LTD_n_wth | 0 | 27 | 75 | 215 | 2823 |
| | sd_sesh_duration | 0.1 | 1.15 | 1.75 | 2.36 | 14.22 |
| Cluster 2 | | | | | | |
| | n_seshs | 6 | 7 | 7 | 13.25 | 29 |
| | max_n_sesh_wk | 1 | 2 | 2 | 3 | 5 |
| | avg_n_sesh_mth | 1 | 1.29 | 2.12 | 2.86 | 5.48 |
| | max_n_sesh_mth | 1 | 2 | 3 | 4 | 9 |
| | LTD_deps | 22 | 41.25 | 69.5 | 110.75 | 308 |
| | avg_txns_per_sesh | 3.14 | 4.93 | 9.55 | 15.6 | 28 |
| | avg_WL_sesh | 76 571.43 | 91 631.8 | 97 208.45 | 122 191.57 | 508 993.17 |
| | LTD_n_wth | 0 | 6.25 | 17 | 25.75 | 145 |
| | sd_sesh_duration | 0.58 | 1.09 | 1.61 | 2.52 | 3.3 |

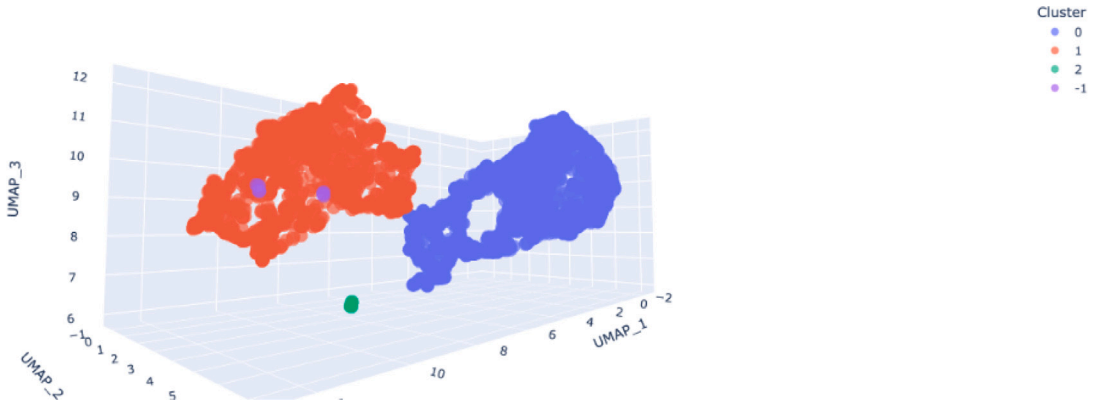



**Fig. 3.** 3D Visualization of Clusters Using UMAP and DBSCAN. The figure shows the three-dimensional projection of the high-dimensional dataset obtained using UMAP with cosine distance, followed by DBSCAN clustering.

gambling-related harm (Gainsbury et al., 2015; Humphreys & Perez, 2012), our findings indicated that among Poland's EGM players, neither gender nor age were associated with an increased rate of males were not more involved and that highly involved gamblers skewed older. This could reflect the heavy male bias in the Polish player base, but it also highlights how the relationship between demographics and gambling engagement may vary by gambling type.

Transactional behaviors further differentiated highly involved players. High variability in deposit and withdrawal amounts along with extreme net win/loss values suggest significant financial fluctuations, echoing the finding that gambling engagement is often concentrated among a minority of high-intensity players (Deng et al., 2021; Louderback et al., 2024). Deposit and balance trajectory metrics revealed both stable and escalating deposit patterns, with some players showing substantial balance increases, possibly reflecting large wins or continued deposits (Adami et al., 2013; Gainsbury et al., 2014; Philander, 2014). These insights hint at heterogeneity in the behaviors of highly-involved gamblers, suggesting that targeted harm reduction strategies tailored

**Table 8**
Grid Search Results for UMAP + DBSCAN Clustering. The table presents clustering evaluation metrics (Silhouette, Davies–Bouldin, and Calinski–Harabasz) for different combinations of hyperparameters. Higher Silhouette and Calinski–Harabasz values indicate better clustering performance, while lower Davies–Bouldin values are preferred.

| n_neighbors | min_dist | metric | eps | min_samples | n_clusters | n_noise | Silhouette | Davies–Bouldin | Calinski–Harabasz |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 0.0 | cosine | 0.4 | 3 | 7 | 2 | −0.0947 | 1.5565 | 3084.8400 |
| 5 | 0.0 | cosine | 0.4 | 5 | 8 | 15 | −0.0147 | 0.9682 | 2674.7300 |
| 5 | 0.0 | cosine | 0.5 | 3 | 5 | 0 | 0.2575 | 1.6675 | 4453.9600 |
| 5 | 0.0 | cosine | 0.5 | 5 | 3 | 7 | 0.5826 | 0.4445 | 8879.4900 |
| 5 | 0.0 | cosine | 0.6 | 3 | 3 | 0 | **0.5827** | **0.4442** | 8907.4600 |
| 5 | 0.0 | cosine | 0.6 | 5 | 3 | 0 | 0.5827 | 0.4442 | 8907.4600 |
| 5 | 0.0 | euclidean | 0.4 | 3 | 10 | 3 | −0.6143 | 2.7056 | 10.0500 |
| 5 | 0.0 | euclidean | 0.4 | 5 | 9 | 12 | −0.3218 | 1.1006 | 1784.2800 |
| 5 | 0.0 | euclidean | 0.5 | 3 | 4 | 1 | −0.3737 | 4.7999 | 10.6500 |
| 5 | 0.0 | euclidean | 0.5 | 5 | 4 | 1 | −0.3737 | 4.7999 | 10.6500 |
| 5 | 0.0 | euclidean | 0.6 | 3 | 2 | 0 | −0.0277 | 1.2830 | 7.4700 |
| 5 | 0.0 | euclidean | 0.6 | 5 | 2 | 0 | −0.0277 | 1.2830 | 7.4700 |
| 5 | 0.1 | cosine | 0.4 | 3 | 8 | 12 | −0.1413 | 1.4054 | 1808.9500 |
| 5 | 0.1 | cosine | 0.4 | 5 | 8 | 30 | −0.2037 | 1.0251 | 1811.6100 |
| 5 | 0.1 | cosine | 0.5 | 3 | 3 | 5 | 0.5514 | 0.5005 | 6259.9900 |
| 5 | 0.1 | cosine | 0.5 | 5 | 3 | 7 | 0.5514 | 0.5005 | 6255.8900 |
| 5 | 0.1 | cosine | 0.6 | 3 | 3 | 2 | 0.5514 | 0.5004 | 6265.6400 |
| 5 | 0.1 | cosine | 0.6 | 5 | 3 | 2 | 0.5514 | 0.5004 | 6265.6400 |
| 5 | 0.1 | euclidean | 0.4 | 3 | 8 | 8 | −0.5508 | 2.0143 | 4.8700 |
| 5 | 0.1 | euclidean | 0.4 | 5 | 10 | 19 | −0.5994 | 1.8357 | 4.8400 |
| 5 | 0.1 | euclidean | 0.5 | 3 | 2 | 0 | −0.0607 | 1.4414 | 6.4100 |
| 5 | 0.1 | euclidean | 0.5 | 5 | 2 | 0 | −0.0607 | 1.4414 | 6.4100 |
| 5 | 0.1 | euclidean | 0.6 | 3 | 2 | 0 | −0.0607 | 1.4414 | 6.4100 |
| 5 | 0.1 | euclidean | 0.6 | 5 | 2 | 0 | −0.0607 | 1.4414 | 6.4100 |
| 10 | 0.0 | cosine | 0.4 | 3 | 4 | 2 | 0.1320 | 1.0926 | 6062.5400 |
| 10 | 0.0 | cosine | 0.4 | 5 | 3 | 7 | 0.2815 | 0.9401 | 9078.5300 |
| 10 | 0.0 | cosine | 0.5 | 3 | 4 | 1 | 0.1321 | 1.0926 | 6065.0400 |
| 10 | 0.0 | cosine | 0.5 | 5 | 3 | 5 | 0.2816 | 0.9400 | 9083.6400 |
| 10 | 0.0 | cosine | 0.6 | 3 | 3 | 0 | 0.2817 | 0.9390 | **9097.3700** |
| 10 | 0.0 | cosine | 0.6 | 5 | 3 | 0 | 0.2817 | 0.9390 | 9097.3700 |

**Table 9**
Comparative unsupervised clustering baselines on the UMAP embedding. We report internal validation metrics (Silhouette, Davies–Bouldin, and Calinski–Harabasz). Higher Silhouette/Calinski–Harabasz indicate better structure, while lower Davies–Bouldin is preferred.

| Method | Key hyperparameters | Silhouette ↑ | DBI ↓ | CHI ↑ |
|---|---|---|---|---|
| **UMAP+DBSCAN** | $n\_n = 5$, $d = 0.0$, metric=cosine, $\epsilon = 0.6$, minPts= 3 | **0.5827** | **0.4442** | **8907.46** |
| UMAP+KMeans | $k = TBD$ | *TBD* | *TBD* | *TBD* |
| UMAP+GMM | $k = TBD$, cov=*TBD* | *TBD* | *TBD* | *TBD* |
| UMAP+Agglomerative | $k = TBD$, link=*TBD* | *TBD* | *TBD* | *TBD* |

**Table 10**
Top 10 Important Features per Cluster (excluding `sd_wth`, `avg_WL_day`, `LTD_n_wth`).

| Cluster -1 | | Cluster 0 | | Cluster 1 | | Cluster 2 | |
|---|---|---|---|---|---|---|---|
| Feature | Imp. | Feature | Imp. | Feature | Imp. | Feature | Imp. |
| avg_n_wth_sesh | 0.0643 | max_n_sesh_mth | 0.1242 | max_n_sesh_mth | 0.1177 | avg_secs_btwn_txns | 0.2553 |
| night_time_pcnt | 0.0580 | avg_n_sesh_mth | 0.1072 | avg_n_sesh_mth | 0.1081 | avg_secs_btwn_deps | 0.2149 |
| avg_bal_traj | 0.0576 | n_seshs | 0.0888 | avg_WL_sesh | 0.0908 | sd_time_btwn_deps | 0.0440 |
| avg_secs_btwn_deps | 0.0542 | avg_WL_sesh | 0.0886 | n_seshs | 0.0877 | avg_hrs_btwn_sesh | 0.0407 |
| sd_time_btwn_deps | 0.0540 | avg_WL_mth | 0.0740 | avg_WL_mth | 0.0736 | sd_time_btwn_sesh | 0.0350 |
| avg_hrs_btwn_sesh | 0.0476 | avg_n_wth_sesh | 0.0822 | avg_n_wth_sesh | 0.0809 | avg_n_sesh_mth | 0.0336 |
| max_WL_sesh | 0.0472 | avg_bal_traj | 0.0807 | avg_bal_traj | 0.0715 | n_seshs | 0.0315 |
| avg_WL_sesh | 0.0469 | avg_secs_btwn_deps | 0.0736 | max_WL_sesh | 0.0701 | avg_WL_sesh | 0.0304 |
| avg_n_sesh_mth | 0.0443 | avg_hrs_btwn_sesh | 0.0719 | avg_hrs_btwn_sesh | 0.0639 | avg_WL_mth | 0.0289 |
| avg_n_sesh_wk | 0.0431 | night_time_pcnt | 0.0651 | avg_secs_btwn_deps | 0.0630 | max_WL_sesh | 0.0271 |

to different behavioral profiles may be more effective than universal approaches (Gainsbury et al., 2018)

This study also introduces a new means of understanding which variables should be clustered upon. By utilizing the UMAP dimensionality reduction technique and feature importances for each cluster classification, we were able to incorporate all of the information (variables) available to us to make the clusters. This is fundamentally different from previous studies, which perform variable selection first, leaving out all but 5-7 variables in order to meet the requirements of the computationally and dimensionally inefficient k-means algorithm (Ghaharian, Abarbanel, Phung, et al., 2023). By using UMAP and feature importance metrics, we cannot only look at feature importances after-the-fact for the entire feature set, but also scale to bigger and higher-dimensional datasets much more efficiently.

In this study, we demonstrated heterogeneity within the group of highly involved gamblers through the identification of four distinct behavioral subtypes. These findings indicate that high involvement is not a uniform construct, reinforcing the value of segmentation when considering RG strategies. Two of the identified clusters are characterized primarily by either sustained high frequency play or structured high volume play, with relatively stable behavioral patterns and limited evidence of escalation or volatility. For these groups, light touch and

informational responsible gambling approaches are likely most appropriate. This could be in the form of passive feedback on session duration or cumulative spend, optional self monitoring tools, and non intrusive reminders aimed at supporting awareness rather than restricting play. In contrast, the remaining clusters represent outlier behavioral profiles marked by greater irregularity or intensity. Notably, one outlier cluster is positioned closer to a core cluster in the latent space, suggesting a transitional profile that may reflect movement between high value engagement and emerging risk. For these players, more proactive and adaptive RG strategies may be warranted, including enhanced monitoring, sending an RG representative, and/or graduated interventions responsive to behavioral escalation.

These findings emphasize that high gambling involvement is multifaceted, encompassing different modes of risky behavior, where some are driven by intensity and volatility, others by persistence and regularity. This segmentation aligns with prior theoretical distinctions between gamblers who play frequently versus those who wager heavily or chase losses (Gainsbury et al., 2018). Recognizing these subtypes is crucial for designing targeted interventions. For instance, strategies focusing on financial limits or self-exclusion may be more effective for clusters with high intensity or volume, whereas behavioral nudges and time reminders may better serve those with sustained, high-frequency engagement. By incorporating behavioral segmentation into responsible gambling frameworks, operators and regulators can better tailor harm reduction efforts to the diverse patterns observed in real-world gambling data.

In practice it is often difficult or costly to obtain labels for high-risk individuals, making it very difficult to conduct inference on the population. While there have been various attempts to develop markers of gambling harm based on gambler behavior (Dragicevic et al., 2011; McAuliffe et al., 2022), there is no consensus in terms of which behaviors (or combination of behaviors) should be used to determine thresholds to label a particular player as a potential high risk gambler. Thus, this study provides a contribution in that it provides an alternate way of creating a proxy for high-risk gamblers, paving the way for supervised machine learning applications.

From an operator and regulator perspective, this research could be utilized to understand and track different customer profiles. In fact, some jurisdictions are rolling out obligatory RG checks using pre-trained AI models (Guimarães, 2025). In practice, an operator could (1) compute the same behavioral features on rolling windows, (2) embed and cluster players using the trained UMAP+DBSCAN workflow (with scheduled re-fitting or incremental refresh), and (3) assign each player to a behavioral subtype that triggers tiered, pre-defined interventions. The UMAP+DBSCAN approach warrants particularly well for brick and mortar (or even hybrid) operators as their patron databases are often large and require model scalability. Feature importance summaries for cluster membership can provide an explanation of why a player is flagged, aligning with the strategic direction of recent AI regulation which is concerned with auditability of AI models and their use cases (Hartmann et al., 2025).

Furthermore, much of the previous research around clustering gamblers based on their behavior is done in static settings and on relatively small datasets. This makes it possible to use computationally inefficient algorithms such as k-means clustering. That said, operator player databases could contain thousands of entries for millions of players, particularly with new forms of data infrastructure coming to market. Our approach to clustering utilizing UMAP and DBSCAN eliminates the computational barrier for industrial-scale applications.

## 6. Limitations and future research

This study presents several limitations that should be acknowledged when interpreting the results. First, the clustering analysis focused exclusively on behavioral tracking data, without incorporating psychological, contextual, or demographic variables beyond basic attributes such as age and gender. While behavioral features are powerful indicators of gambling patterns, future work could integrate survey-based measures (e.g., impulsivity, gambling motives) or contextual information (e.g., socioeconomic status, life events) to enhance the interpretability and evaluate the external validity of the identified clusters.

Second, the dataset was derived from a single gambling operator in Poland, and may not generalize to other types of gambling activities (e.g., sports betting, online poker), regions, or cultural contexts. Gambling motivations and patterns can vary significantly across jurisdictions due to legal, social, and economic differences. Future research should validate these findings using multi-operator datasets and across diverse geographic settings to examine the stability and replicability of the cluster typologies.

Third, the clustering process was unsupervised, and while we used internal validation metrics (Silhouette, Davies–Bouldin, and Calinski–Harabasz scores) to select optimal configurations, these do not guarantee external validity. The functional significance and long-term predictive utility of the clusters — such as their association with future gambling problems, self-exclusion, or financial distress — remain untested. Longitudinal studies linking cluster membership to future outcomes would be instrumental in assessing the practical relevance of these behavioral segments.

Moreover, the dimensionality reduction step with UMAP introduces stochasticity, which could lead to variability in the clustering output. Although we fixed the random seed to ensure reproducibility, further work could explore ensemble or consensus clustering methods to improve the robustness of cluster assignments.

Lastly, feature importance was computed using a cluster-vs-rest binary classification approach. While this approach helps to interpret the unique characteristics of each cluster, it may oversimplify the complex interdependencies among clusters. Future studies may explore multivariate or network-based feature importance models such as graphical LASSO or Normal-Block to capture more nuanced relationships (Tan et al., 2015; Tous & Chiquet, 2025).

Future research could also examine real-time clustering systems and dynamic updates of player profiles, especially in high-risk gambling environments. This would allow for adaptive harm-reduction interventions that evolve in response to behavioral change over time.

## 7. Conclusion

The clustering analysis offered a more nuanced understanding of the heterogeneity within the highly involved gambler group. Using UMAP for non-linear dimensionality reduction and DBSCAN for density-based clustering, we identified four distinct subgroups based on behavioral markers. These clusters differed not only in gambling frequency and intensity but also in their temporal dynamics, transaction patterns, and volatility profiles. For instance, **Cluster -1** exhibited erratic withdrawal behavior and frequent night-time sessions—characteristics previously linked to impulsivity and emotional gambling. In contrast, **Cluster 0** showed a pattern of regular, high-frequency play with stable financial behavior, suggestive of habitual engagement. **Cluster 1** was similar in volume but demonstrated more deliberate and spaced-out gambling sessions, while **Cluster 2** was distinguished by compressed time intervals between transactions, indicating rapid and potentially compulsive play.

These differentiated behavioral profiles confirm that high involvement in gambling is not a monolithic construct. Instead, it comprises subtypes that may reflect distinct psychological and motivational mechanisms. This segmentation supports earlier findings that frequent play and high spending may stem from different underlying processes (Gainsbury et al., 2018; Philander, 2014). The identification of such subgroups has practical implications for responsible gambling policies: rather than applying uniform interventions, tailored strategies could be designed to address the specific behavioral risks associated with each cluster. For example, real-time alerts and time-out suggestions might benefit players in Cluster 2, while personalized feedback on

financial losses or withdrawal patterns could support those in Cluster -1. Overall, the results underscore the value of unsupervised learning methods in identifying latent structures within gambling behavior data and highlight the potential of these insights to inform harm reduction frameworks.

Over the past three years, Mana Azizsoltani has either worked on projects funded by or has personally received funding, honoraria, travel reimbursement, or consulting fees from AXES.ai, Walker Digital Table Systems, Melco Resorts, Crown Melbourne, Inspire Resort, Bally's Resorts, the Nevada Council on Problem Gambling, and the Nevada Department of Health and Human Services. None of these entities had any role in the design, analysis, or interpretation of the present study and imposed no constraints on its publication.

Over the past five years, Kasra Ghaharian has either worked on projects funded by, or has personally received funding, honoraria, travel reimbursement, or consulting fees from the International Center for Responsible Gambling, the Nevada Department of Health and Human Services, the Nevada Governor's Office of Economic Development, the Massachusetts Gaming Commission, AXES.ai, Playtech, Evoke, Gaming Analytics, Walker Digital Table Systems, the Responsible Online Gaming Association, Videopoker.com, Kindbridge Behavioral Health, IGT, Differential Labs, Yaspa, Focal Research Consultants, Bet Blocker, Sports Betting Alliance, ESPN, Sightline, Global Payments, Telus, GP Consulting, the Responsible Gambling Council, the Illinois Council on Problem Gambling, the Alberta Gambling Research Institute, and Kindred Group. None of these entities had any role in the design, analysis, or interpretation of the present study and imposed no constraints on its publication.

During the past five years, the International Gaming Institute (IGI) at University of Nevada, Las Vegas, has received research and program funding from DraftKings, Inc., the American Gaming Association, ESPN, MGM Resorts International, Wynn Resorts Ltd, Las Vegas Sands Corporation, Entain Foundation, Aristocrat Gaming, San Manuel Band of Mission Indians, Axes.ai, Sports Betting Alliance, Playtech, Sightline Payments, Global Payments, the State of Nevada Knowledge Fund, and the State of Nevada Department of Health and Human Services. IGI runs the triennial research-focused International Conference on Gambling and Risk Taking, whose sponsors include industry, academic, and legal/regulatory stakeholders in gambling. A full list of sponsors for the most recent conference can be found at https://www.unlv.edu/igi/conference/18th/sponsors. IGI maintains a strict research policy (https://www.unlv.edu/igi/research-policy), as well as partnership and transparency framework (https://www.unlv.edu/igi/policies/partnership) to ensure appropriate firewalls exist between funding entities — no matter the entity's classification — and IGI's research and programs.

## CRediT authorship contribution statement

**Mana Azizsoltani:** Writing – original draft, Software, Methodology, Conceptualization. **Ismael Gomez-Talal:** Writing – original draft, Software, Data curation, Conceptualization. **José Luis Rojo Alvarez:** Writing – review & editing, Supervision, Methodology. **Kasra Ghaharian:** Writing – review & editing, Supervision, Methodology.

## Declaration of the Use of Generative AI Tools

The authors acknowledge that generative AI tools, such as ChatGPT (OpenAI), were used to assist in improving the language and readability of the manuscript. All content was reviewed and edited by the authors to ensure accuracy and clarity. The use of such tools was limited strictly to the writing process, and not to the analysis or generation of research data or results. The authors take full responsibility for the integrity and originality of the content presented in this article, in accordance with Elsevier's policies on the use of AI and AI-assisted technologies.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

Funding for this study was provided by the Nevada Department of Health and Human Services (NVDHHS) and the Nevada Council on Problem Gambling (NVCPG). This work was supported by the CyberFold project, funded by the European Union through the NextGenerationEU instrument (Recovery, Transformation, and Resilience Plan), and managed by Instituto Nacional de Ciberseguridad de España (INCIBE), Spain, under reference number ETD202300129. Partially supported by the Autonomous Community of Madrid (ELLIS Madrid Node), by project PID2022-140786NB-C32 (LATENTIA) and project AIA2025-163540-C31 (EmbedWorld) from the Spanish Ministry of Science and Innovation (AEI/10.13039/501100011033). Finally, the authors would like to thank those at AXES.ai for their support and guidance throughout the research project.

## Data availability

The authors do not have permission to share data.

## References

Adami, N., Benini, S., Boschetti, A., Canini, L., Maione, F., & Temporin, M. (2013). Markers of unsustainable gambling for early detection of at-risk online gamblers. *International Gambling Studies*, *13*(2), 188–204. http://dx.doi.org/10.1080/14459795.2012.754919, Publisher: Taylor & Francis.

American Gaming Association (2024). *Commercial gaming revenue tracker*: *Technical report*, (pp. 1–7). American Gaming Association, URL: https://www.americangaming.org/wp-content/uploads/2024/08/Q2-2024_CGRT-1.pdf.

Auer, M., & Griffiths, M. D. (2023). Using artificial intelligence algorithms to predict self-reported problem gambling with account-based player data in an online casino setting. *Journal of Gambling Studies*, *39*(3), 1273–1294. http://dx.doi.org/10.1007/s10899-022-10139-1.

Australian Productivity Commission, et al. (2010). Gambling productivity commission inquiry report. *Australian Government*.

Barton, K., Yazdani, Y., Ayer, N., Kalvapalle, S., Brown, S., Stapleton, J., Brown, D., & Harrigan, K. (2017). The effect of losses disguised as wins and near misses in electronic gaming machines: A systematic review. *Journal of Gambling Studies*, *33*, 1241–1260. http://dx.doi.org/10.1007/s10899-017-9688-0.

Braverman, J., & Shaffer, H. J. (2012). How do gamblers start gambling: Identifying behavioural markers for high-risk internet gambling. *The European Journal of Public Health*, *22*(2), 273–278. http://dx.doi.org/10.1093/eurpub/ckp232, Publisher: Oxford University Press.

Browne, M., Langham, E., Rockloff, M. J., Li, E., Donaldson, P., & Goodwin, B. (2015). EGM jackpots and player behaviour: An in-venue shadowing study. *Journal of Gambling Studies*, *31*, 1695–1714. http://dx.doi.org/10.1007/s10899-014-9485-y.

Chagas, B. T., & Gomes, J. (2017). Internet gambling: A critical review of behavioural tracking research. *Journal of Gambling Issues*, *36*(1), 1–27. http://dx.doi.org/10.4309/jgi.2017.36.1, Publisher: Centre for Addiction and Mental Health (CAMH).

Chagas, B. T., Gomes, J., & Griffiths, M. D. (2021). Consumer profile segmentation in online lottery gambling utilizing behavioral tracking data from the portuguese national lottery. *Journal of Gambling Studies*, 1–23. http://dx.doi.org/10.1007/s10899-021-10072-9.

Delfabbro, P., Falzon, K., & Ingram, T. (2005). The effects of parameter variations in electronic gambling simulations: Results of a laboratory-based pilot investigation. *Gambling Research: Journal of the National Association for Gambling Studies (Australia)*, *17*(1), 7–25, https://search.informit.org/doi/10.3316/informit.844557461996677.

Delfabbro, P., Parke, J., & Catania, M. (2024). Behavioural tracking and profiling studies involving objective data derived from online operators: A review of the evidence. *Journal of Gambling Studies*, *40*(2), 639–671. http://dx.doi.org/10.1007/s10899-023-10247-6.

Deng, X., Lesch, T., & Clark, L. (2019). Applying Data Science to Behavioral Analysis of Online Gambling. *Current Addiction Reports*, *6*(3), 159–164. http://dx.doi.org/10.1007/s40429-019-00269-9, URL: http://link.springer.com/10.1007/s40429-019-00269-9.

Deng, X., Lesch, T., & Clark, L. (2021). Pareto distributions in online casino gambling: Sensitivity to timeframe and associations with self exclusion. *Addictive Behaviors*, *120*, Article 106968. http://dx.doi.org/10.1016/j.addbeh.2021.106968.

Dixon, M. J., Graydon, C., Harrigan, K. A., Wojtowicz, L., Siu, V., & Fugelsang, J. A. (2014). The allure of multi-line games in modern slot machines. *Addiction*, *109*(11), 1920–1928. http://dx.doi.org/10.1111/add.12675, Publisher: Wiley Online Library.

Dixon, M. J., Harrigan, K. A., Sandhu, R., Collins, K., & Fugelsang, J. A. (2010). Losses disguised as wins in modern multi-line video slot machines. *Addiction*, *105*(10), 1819–1824. http://dx.doi.org/10.1111/j.1360-0443.2010.03050.x.

Donaldson, P., Langham, E., Rockloff, M. J., & Browne, M. (2016). Veiled EGM jackpots: The effects of hidden and mystery jackpots on gambling intensity. *Journal of Gambling Studies*, *32*, 487–498. http://dx.doi.org/10.1007/s10899-015-9566-6.

Dowling, N., Smith, D., & Thomas, T. (2005). Electronic gaming machines: are they the 'crack-cocaine'of gambling? *Addiction*, *100*(1), 33–45. http://dx.doi.org/10.1111/j.1360-0443.2005.00962.x, Publisher: Wiley Online Library.

Dragicevic, S., Tsogas, G., & Kudic, A. (2011). Analysis of casino online gambling data in relation to behavioural risk markers for high-risk gambling and player protection. *International Gambling Studies*, *11*(3), 377–391. http://dx.doi.org/10.1080/14459795.2011.629204.

Edson, T. C., Tom, M. A., Louderback, E. R., Nelson, S. E., & LaPlante, D. A. (2022). Returning to the virtual casino: a contemporary study of actual online casino gambling. *International Gambling Studies*, *22*(1), 114–141. http://dx.doi.org/10.1080/14459795.2021.1985581, Publisher: Routledge.

Excell, D., Bobashev, G., Gonzalez-Ordonez, D., Wardle, H., Whitehead, T., Morris, R. J., & Ruddle, P. (2014). Report 3: Predicting problem gamblers: analysis of industry data. Gambling Machines Research Program. Responsible Gambling Trust.

Gainsbury, S. (2011). Player account-based gambling: Potentials for behaviour-based research methodologies. *International Gambling Studies*, *11*(2), 153–171. http://dx.doi.org/10.1080/14459795.2011.571217.

Gainsbury, S. M., Abarbanel, B. L., Philander, K. S., & Butler, J. V. (2018). Strategies to customize responsible gambling messages: a review and focus group study. *BMC Public Health*, *18*, 1–11. http://dx.doi.org/10.1186/s12889-018-6281-0, Publisher: Springer.

Gainsbury, S. M., Russell, A., Hing, N., Wood, R., Lubman, D., & Blaszczynski, A. (2015). How the Internet is changing gambling: Findings from an Australian prevalence survey. *Journal of Gambling Studies*, *31*, 1–15. http://dx.doi.org/10.1007/s10899-013-9404-7, Publisher: Springer.

Gainsbury, S. M., Suhonen, N., & Saastamoinen, J. (2014). Chasing losses in online poker and casino games: Characteristics and game play of Internet gamblers at risk of disordered gambling. *Psychiatry Research*, *217*(3), 220–225. http://dx.doi.org/10.1016/j.psychres.2014.03.033, Publisher: Elsevier.

Ghaharian, K., Abarbanel, B., Kraus, S. W., Singh, A., & Bernhard, B. (2023). Players Gonna Pay: Characterizing gamblers and gambling-related harm with payments transaction data. *Computers in Human Behavior*, *143*, Article 107717. http://dx.doi.org/10.1016/j.chb.2023.107717, URL: https://www.sciencedirect.com/science/article/pii/S0747563223000687.

Ghaharian, K., Abarbanel, B., Phung, D., Puranik, P., Kraus, S., Feldman, A., & Bernhard, B. (2023). Applications of data science for responsible gambling: a scoping review. *International Gambling Studies*, *23*(2), 289–312. http://dx.doi.org/10.1080/14459795.2022.2135753.

Ghaharian, K., Peterson, J., Azizsoltani, M., Young, R. J., & Louderback, E. R. (2025). Across the bettor-verse: an open banking perspective on gambling in the united kingdom. *Journal of Gambling Studies*, *41*(4), 1419–1436. http://dx.doi.org/10.1007/s10899-025-10419-6.

Griffiths, M. D., & Meredith, A. (2009). Videogame addiction and its treatment. *Journal of Contemporary Psychotherapy*, *39*, 247–253. http://dx.doi.org/10.1007/s10879-009-9118-4.

Guimarães, S. (2025). Spain develops AI system to monitor 60 indicators of gambling risk. *Esports.Gg*, URL: https://esports.gg/news/betting/spain-ai-system-monitor-gambling-risk/.

Harris, A., & Griffiths, M. D. (2018). The impact of speed of play in gambling on psychological and behavioural factors: A critical review. *Journal of Gambling Studies*, *34*, 393–412. http://dx.doi.org/10.1007/s10899-017-9701-7, Publisher: Springer.

Hartmann, D., De Pereira, J. R. L., Streitbörger, C., & Berendt, B. (2025). Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act by including civil society. *AI and Ethics*, *5*(4), 3617–3638. http://dx.doi.org/10.1007/s43681-024-00595-3.

Horton, K., Turner, N., & Horbay, R. (2006). Do weighted reels on a slot machine distort a gambler's judgment of probability? The effect of near misses. Report Submitted To Ontario Problem Gambling Research Centre.

Humphreys, B. R., & Perez, L. (2012). Participation in internet gambling markets: An international comparison of online gamblers' profiles. *Journal of Internet Commerce*, *11*(1), 24–40. http://dx.doi.org/10.1080/15332861.2012.650987.

ICLG (2025). Gambling laws and regulations Poland 2025. https://iclg.com/practice-areas/gambling-laws-and-regulations/poland. (Accessed: 09 June 2025).

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Tree-based methods. In *An introduction to statistical learning: with applications in python* (pp. 331–366). Springer, http://dx.doi.org/10.1007/978-3-031-38747-0_8.

LaBrie, R. A., LaPlante, D. A., Nelson, S. E., Schumann, A., & Shaffer, H. J. (2007). Assessing the playing field: A prospective longitudinal study of internet sports gambling behavior. *Journal of Gambling Studies*, *23*(3), 347–362. http://dx.doi.org/10.1007/s10899-007-9067-3, Publisher: Springer.

Landon, J., Palmer Du Preez, K., Page, A., Bellringer, M., Roberts, A., & Abbott, M. (2018). Electronic gaming machine characteristics: It's the little things that count. *International Journal of Mental Health and Addiction*, *16*(2), 251–265. http://dx.doi.org/10.1007/s11469-016-9666-2, URL: http://link.springer.com/10.1007/s11469-016-9666-2.

Loba, P., Stewart, S. H., Klein, R. M., & Blackburn, J. R. (2001). Manipulations of the features of standard video lottery terminal (VLT) games: Effects in pathological and non-pathological gamblers. *Journal of Gambling Studies*, *17*, 297–320. http://dx.doi.org/10.1023/A:1013639729908.

Louderback, E. R., Tom, M. A., Edson, T. C., & LaPlante, D. A. (2024). The stability of gambling expenditure distributions over time and associations with the use of gambling self-regulatory tools. *International Journal of Mental Health and Addiction*, 1–23. http://dx.doi.org/10.1007/s11469-024-01399-6.

Marionneau, V., Ristolainen, K., & Roukka, T. (2025). Duty of care, data science, and gambling harm: A scoping review of risk assessment models. *Computers in Human Behavior Reports*, Article 100644. http://dx.doi.org/10.1016/j.chbr.2025.100644.

McAuliffe, W. H., Louderback, E. R., Edson, T. C., LaPlante, D. A., & Nelson, S. E. (2022). Using "markers of harm" to track risky gambling in two cohorts of online sports bettors. *Journal of Gambling Studies*, *38*(4), 1337–1369. http://dx.doi.org/10.1007/s10899-021-10097-0.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. http://dx.doi.org/10.48550/arXiv.1802.03426, arXiv preprint arXiv:1802.03426.

Meyer, G., Hayer, T., & Griffiths, M. (2009). *Problem gambling in Europe: Challenges, prevention, and interventions*. Springer, http://dx.doi.org/10.1007/978-0-387-09486-1.

Moreira, D., Azeredo, A., & Dias, P. (2023). Risk factors for gambling disorder: A systematic review. *Journal of Gambling Studies*, *39*(2), 483–511. http://dx.doi.org/10.1007/s10899-023-10195-1.

Mosquera, M. G., & Keselj, V. (2017). Identifying electronic gaming machine gambling personae through unsupervised session classification. *Big Data & Information Analytics*, *2*(2), 141–175. http://dx.doi.org/10.3934/bdia.2017015, URL: https://www.aimsciences.org/en/article/doi/10.3934/bdia.2017015. Publisher: Big Data & Information Analytics.

Murch, W. S., & Clark, L. (2019). Effects of bet size and multi-line play on immersion and respiratory sinus arrhythmia during electronic gaming machine use. *Addictive Behaviors*, *88*, 67–72. http://dx.doi.org/10.1016/j.addbeh.2018.08.014, Publisher: Elsevier.

Nelson, S. E., Edson, T. C., Louderback, E. R., Tom, M. A., Grossman, A., & LaPlante, D. A. (2021). Changes to the playing field: A contemporary study of actual European online sports betting. *Journal of Behavioral Addictions*, *10*(3), 396–411. http://dx.doi.org/10.1556/2006.2021.00029, URL: https://akjournals.com/view/journals/2006/10/3/article-p396.xml.

Nelson, S. E., Edson, T. C., Singh, P., Tom, M., Martin, R. J., LaPlante, D. A., Gray, H. M., & Shaffer, H. J. (2019). Patterns of Daily Fantasy Sport Play: Tackling the Issues. *Journal of Gambling Studies*, *35*(1), 181–204. http://dx.doi.org/10.1007/s10899-018-09817-w, URL: http://link.springer.com/10.1007/s10899-018-09817-w.

Peister, C., Acres, N., & Moore, S. (2024). Winning the AI arms race with data. *IAG*, URL: https://www.asgam.com/index.php/2024/07/31/winning-the-ai-arms-race-with-data/. Section: Columnists.

Peres, F., Fallacara, E., Manzoni, L., Castelli, M., Popovič, A., Rodrigues, M., & Estevens, P. (2021). Time series clustering of online gambling activities for addicted users' detection. *Applied Sciences*, *11*(5), 2397. http://dx.doi.org/10.3390/app11052397.

Perrot, B., Hardouin, J.-B., Grall-Bronnec, M., & Challet-Bouju, G. (2018). Typology of online lotteries and scratch games gamblers' behaviours: A multilevel latent class cluster analysis applied to player account-based gambling data. *International Journal of Methods in Psychiatric Research*, *27*(4), Article e1746. http://dx.doi.org/10.1002/mpr.1746.

Philander, K. S. (2014). Identifying high-risk online gamblers: A comparison of data mining procedures. *International Gambling Studies*, *14*(1), 53–63. http://dx.doi.org/10.1080/14459795.2013.841721, Publisher: Taylor & Francis.

Pickering, D., Philander, K. S., & Gainsbury, S. M. (2020). Skill-based electronic gaming machines: A review of product structures, risks of harm, and policy issues. *Current Addiction Reports*, *7*, 229–236. http://dx.doi.org/10.1007/s40429-020-00309-9.

Pricewaterhouse Coopers (PWC) (2017). *Remote gambling research interim report on phase II: Technical report*, PricewaterhouseCoopers LLP, URL: https://www.begambleaware.org/media/1549/gamble-aware_remote-gambling-research_phase-2_pwc-report_august-2017-final.pdf.

Rockloff, M. J., & Greer, N. (2010). Never smile at a crocodile: Betting on electronic gaming machines is intensified by reptile-induced arousal. *Journal of Gambling Studies*, *26*(4), 571–581. http://dx.doi.org/10.1007/s10899-009-9174-4.

Rockloff, M. J., & Hing, N. (2013). The impact of jackpots on EGM gambling behavior: A review. *Journal of Gambling Studies*, *29*, 775–790. http://dx.doi.org/10.1007/s10899-012-9336-7, Publisher: Springer.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, *42*(3), 1–21. http://dx.doi.org/10.1145/3068335.

Schüll, N. D. (2012). Addiction by design: Machine gambling in Las Vegas. In *Addiction by design*. Princeton University Press.

Stelmachowski, A., Dynowski, P., & Zawadzki, P. (2023). A general introduction to gambling law in Poland. https://www.lexology.com/library/detail.aspx?g=e913608f-2909-44f8-b018-b865ccb3b734. (Accessed: 09 June 2025).

Suriadi, S., Susnjak, T., Ponder-Sutton, A., Watters, P., & Schumacher, C. (2016). Using data-driven and process mining techniques for identifying and characterizing problem gamblers in New Zealand. *Complex Systems Informatics and Modeling Quarterly*, *9*, 44–66. http://dx.doi.org/10.7250/csimq.2016-9.03.

Tan, K. M., Witten, D., & Shojaie, A. (2015). The cluster graphical lasso for improved estimation of Gaussian graphical models. *Computational Statistics & Data Analysis*, *85*, 23–36. http://dx.doi.org/10.1016/j.csda.2014.11.015.

Tom, M. A., Edson, T. C., Louderback, E. R., Nelson, S. E., Amichia, K. A., & LaPlante, D. A. (2022). Second Session at the Virtual Poker Table: A Contemporary Study of Actual Online Poker Activity. *Journal of Gambling Studies*, *39*(3), 1295–1317. http://dx.doi.org/10.1007/s10899-022-10147-1, URL: https://link.springer.com/10.1007/s10899-022-10147-1.

Tous, J., & Chiquet, J. (2025). An integrated method for clustering and association network inference. http://dx.doi.org/10.48550/arXiv.2503.22467, arXiv preprint arXiv:2503.22467.

Wiley, R. C., Tom, M. A., Edson, T. C., & LaPlante, D. A. (2020). Behavioral markers of risky daily fantasy sports play. *Journal of Sport and Social Issues*, *44*(4), 356–371. http://dx.doi.org/10.1177/0193723520919819.

Wilkes, B. L., Gonsalvez, C. J., & Blaszczynski, A. (2010). Capturing SCL and HR changes to win and loss events during gambling on electronic machines. *International Journal of Psychophysiology*, *78*(3), 265–272. http://dx.doi.org/10.1016/j.ijpsycho.2010.08.008.

Woolley, R., & Livingstone, C. (2009). Into the zone: innovating in the Australian poker machine industry. In *Global gambling - cultural perspectives on gambling organisations* (pp. 38–63). United Kingdom: Routledge.

Young, M. M., Wohl, M. J., Matheson, K., Baumann, S., & Anisman, H. (2008). The desire to gamble: The influence of outcomes on the priming effects of a gambling episode. *Journal of Gambling Studies*, *24*, 275–293. http://dx.doi.org/10.1007/s10899-008-9093-9.

Zack, M., George, R. S., & Clark, L. (2020). Dopaminergic signaling of uncertainty and the aetiology of gambling addiction. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *99*, Article 109853. http://dx.doi.org/10.1016/j.pnpbp.2019.109853, Publisher: Elsevier.

Zendle, D., & Newall, P. (2024). The relationship between gambling behaviour and gambling-related harm: A data fusion approach using open banking data. *Addiction*, http://dx.doi.org/10.1111/add.16571, Publisher: Wiley Online Library.