Interdisciplinary Description of Complex Systems 24(4), 430-445, 2026

DATA-DRIVEN TACTICAL ARCHETYPES IN EUROPEAN PROFESSIONAL FOOTBALL: A K-MEANS CLUSTERING APPROACH

Tomislav Medić*, Antonio Pavlečić and Mirjana Pejić Bach

University of Zagreb, Faculty of Economics & Business
Zagreb, Croatia

DOI: 10.7906/indecs.24.4.5
Regular article

Received: 4 March 2026.
Accepted: 4 May 2026.

ABSTRACT

Positional labels in football no longer adequately reflect the fluid and role-based nature of modern play. This study examines whether player roles can be identified directly from match performance data, without relying on predefined positional categories. Using a dataset of 1988 outfield players from the top five European leagues (2024/2025 season), K-means clustering was applied to eight per-appearance performance metrics. The optimal number of clusters was determined using the Elbow method and Silhouette scores. The analysis produced five statistically distinct player archetypes: Conservative/Disciplined players, Defensive Enforcers, Primary Attackers/Finishers, Active Wingers, and Creative Playmakers. The results demonstrate that data-driven clustering captures functional differences in player behaviour that are not reflected in traditional positional classifications. In particular, the model distinguishes clearly between defensive disruption roles, high-output attacking profiles, and intermediary creative functions. These findings support a shift from position-based to behaviour-based player evaluation. The study contributes to football analytics by providing an interpretable framework for tactical archetyping based on unsupervised learning. Practical implications relate to scouting and recruitment, where player identification can be aligned with functional performance profiles. Limitations arise from the absence of defensive tracking data, and future research should incorporate richer datasets and alternative clustering approaches.

KEY WORDS

football, K-means clustering, tactical archetypes, performance metrics, unsupervised machine learning

CLASSIFICATION

JEL: C38, C55, L83

*Corresponding author, η: tmedic@net.efzg.hr; +385 91 2635 066;
Faculty of Economics and Business, Trg John F. Kennedy 6, HR – 10 000, Zagreb, Croatia

INTRODUCTION

The traditional vocabulary used to describe football players has not kept pace with the game's evolution. Positional labels such as “defender,” “midfielder,” or “forward” originate from relatively rigid tactical systems and assume a stable spatial organisation of play. However, contemporary football is characterised by fluid, role-based dynamics in which players frequently operate across multiple zones and contribute to both defensive and offensive phases [1, 2]. As a result, conventional positional classifications often fail to accurately capture a player’s actual tactical function and on-field behaviour.

This discrepancy between nominal position and functional role presents a practical challenge for performance analysis, scouting, and recruitment. Modern tactical systems increasingly rely on hybrid roles, such as inverted full-backs or false nines, in which player contributions are defined by behavioural patterns rather than fixed positions [3, 4]. Consequently, evaluating players solely within predefined positional categories risks overlooking the multidimensional and context-dependent nature of their performance.

At the same time, the rapid expansion of data collection in professional football has enabled more advanced analytical approaches. The integration of sports analytics and big data techniques has significantly enhanced the ability to quantify and interpret player performance [5]. Despite these developments, a substantial portion of existing analyses continues to rely on positional frameworks as the primary organising principle. This reliance introduces potential bias, as it depends on human-assigned labels that may reinforce traditional assumptions and obscure non-standard yet effective playing styles [6].

Machine learning offers an alternative framework for addressing these limitations. In particular, unsupervised learning techniques enable the identification of latent structures in data without relying on predefined labels. Clustering algorithms group observations based on similarity, allowing player roles to emerge directly from performance patterns [7]. Among these methods, K-means clustering is widely used due to its interpretability and computational efficiency, making it suitable for exploratory analysis in applied contexts [8, 9]. Unlike supervised approaches, which rely on labelled training data and may inherit existing biases, unsupervised methods offer a data-driven means of uncovering behavioural regularities [10].

Previous applications of clustering in sports analytics have demonstrated that data-driven groupings can reveal meaningful distinctions between athletes that are not captured by traditional classifications [11]. In football, this creates the opportunity to define “tactical fingerprints” based on actual match behaviour rather than nominal roles. Such an approach has important implications for player evaluation and recruitment, as it enables the identification of functionally similar players across teams, leagues, and positional labels 0.

Despite these advances, existing research has not sufficiently addressed the problem of deriving tactical roles exclusively from match performance data without incorporating positional information. Most approaches either rely on predefined categories or focus on specific performance dimensions, limiting their ability to capture the full behavioural complexity of modern football. This gap underscores the need for a fully data-driven framework that identifies coherent, interpretable player archetypes solely from observed performance patterns.

This study addresses this gap by applying K-means clustering to a dataset of 1988 outfield players from the top five European leagues during the 2024/2025 season. Using per-appearance performance metrics, the study examines whether meaningful and statistically distinct player groups can be identified independently of positional labels. The results are evaluated using standard clustering validation techniques and interpreted in terms of tactical function.

The contribution of this research is twofold. First, it provides an empirical demonstration of how unsupervised machine learning can be used to derive behaviour-based player archetypes in professional football. Second, it offers a practical, interpretable framework for analysing player roles based on functional output, thereby supporting the ongoing shift from position-based classification to data-driven performance modelling in football analytics.

LITERATURE REVIEW

DATA-DRIVEN TRANSFORMATION OF SPORTS ANALYTICS

The integration of big data and artificial intelligence has enabled more advanced modelling and decision-making processes [13]. The rapid expansion of data availability in professional sports has significantly transformed performance analysis, particularly in football. Within this context, sports analytics has evolved from descriptive statistics to data-driven frameworks capable of capturing complex, multidimensional performance patterns. Bibliometric and review-based studies confirm a substantial growth in research combining data mining, machine learning, and performance analytics in sport [14]. These developments have improved the ability to quantify performance and identify patterns across players and teams. However, despite methodological progress, much of the existing work continues to rely on traditional positional categorisations as the primary analytical structure. As a result, the interpretation of performance often remains constrained by static labels that fail to reflect the fluid, hybrid nature of modern football roles.

MACHINE LEARNING APPLICATIONS IN FOOTBALL

Machine learning has been widely applied in football across several domains, including match outcome prediction, player valuation, and decision support. Predictive models have demonstrated strong performance in forecasting results and identifying key determinants of team success [15]. Similarly, machine learning techniques have been used to estimate player market value and assess individual performance based on structured datasets [16]. In addition to predictive applications, machine learning has been increasingly used in operational decision-making contexts. For example, data-driven approaches have been applied to support team management decisions and optimise training processes [17, 18]. Comparative modelling studies further highlight the importance of methodological choices in performance analysis, emphasising the trade-offs between interpretability and predictive accuracy [19]. Despite these advances, the majority of machine learning applications in football remain focused on prediction or evaluation tasks within predefined frameworks. Player roles are typically treated as given inputs rather than as constructs that can be derived from data. Consequently, these approaches do not fully address the challenge of redefining player roles to reflect observed behaviour rather than nominal classification.

CLUSTERING AND UNSUPERVISED LEARNING IN PLAYER PROFILING

Unsupervised learning methods, particularly clustering techniques, offer an alternative approach by enabling the identification of latent structures in performance data without predefined labels. In contrast to supervised models, clustering algorithms group players based on similarity in observed metrics, allowing patterns of behaviour to emerge directly from the data. K-means clustering is one of the most widely used methods in this context due to its simplicity, interpretability, and scalability. Previous studies have applied K-means and related techniques to group athletes based on performance characteristics, training data, or economic indicators [20-22]. More advanced approaches, such as deep embedded clustering, have also been introduced to capture complex patterns in high-dimensional datasets [23]. Recent football-specific studies have also applied clustering techniques to identify player profiles

based on technical and physical performance data, further confirming the applicability of unsupervised learning in football analytics [11, 21]. While these studies demonstrate the methodological potential of clustering in sports analytics, they often rely on predefined variables, contextual constraints, or domain-specific assumptions. In many cases, clustering is used to refine existing categories rather than to challenge them. As a result, the ability of clustering methods to generate fully data-driven representations of player roles, independent of positional labels, remains underexplored.

METHODOLOGICAL FOUNDATIONS OF CLUSTERING AND DATA PREPARATION

The effectiveness of clustering methods is strongly influenced by data pre-processing, feature selection, and normalisation procedures. K-means clustering is sensitive to variable scaling and distance metrics, which can significantly affect the resulting cluster structure [24, 25]. Determining the optimal number of clusters is another critical step, typically addressed using diagnostic techniques such as the Elbow method and Silhouette scores [26, 27]. Additional methodological considerations include cluster stability, statistical power, and the representation of high-dimensional data [28]. The choice of distance metrics and feature construction is especially important when dealing with heterogeneous performance indicators [29]. Evidence from related domains, such as marketing analytics, further shows that addressing data imbalance can improve model robustness and interpretability [30]. These methodological insights are directly relevant to sports analytics, where performance data are often unevenly distributed and context-dependent. They highlight the importance of careful model design when using clustering techniques to yield meaningful, interpretable results.

BEHAVIOURAL AND CONTEXTUAL PERSPECTIVES IN SPORTS DATA

Recent research has emphasised the importance of behavioural and contextual dimensions in performance analysis. Studies examining user behaviour and engagement in sports-related contexts demonstrate how data-driven approaches can capture complex interaction patterns [31, 32]. Although these studies are not directly focused on on-field performance, they provide a conceptual foundation for analysing behaviour as a multidimensional construct. Similarly, research on simulation-based environments highlights the potential of data-driven models to represent complex systems and interactions [33]. These perspectives support the view that performance should be understood as an emergent property of multiple interacting factors, rather than as a function of static categorical assignments.

RESEARCH GAP

Although existing literature demonstrates extensive use of machine learning and clustering techniques in football analytics, several limitations remain. First, most studies operate within predefined positional or contextual frameworks, thereby reintroducing classification bias into otherwise data-driven approaches. Second, clustering applications often focus on specific performance dimensions – such as physical output, economic value, or training load – rather than providing a comprehensive representation of in-game behaviour. Most importantly, there is a lack of research applying unsupervised learning to derive tactical player roles solely from match performance data, without incorporating positional information. This limits the ability of existing models to capture the behavioural complexity and functional diversity of modern football. This study addresses this gap by applying K-means clustering to performance metrics alone, with the objective of identifying coherent and interpretable tactical archetypes based solely on observed player behaviour. In doing so, it contributes to the shift from position-based classification toward behaviour-based modelling in football analytics.

METHODOLOGY

RESEARCH DESIGN

This study employs a quantitative research design based on unsupervised machine learning to identify data-driven tactical archetypes in professional football. The methodological approach is structured into several stages: data collection, pre-processing, feature selection, clustering, and model validation. The primary objective is to determine whether player roles can be derived directly from match performance data without relying on predefined positional labels. K-means clustering was selected as the core analytical method due to its interpretability, computational efficiency, and suitability for exploratory analysis in high-dimensional datasets [9]. The analysis was conducted using Python (version 3.12), with the scikit-learn library for clustering and pandas and matplotlib for data preprocessing and visualisation.

DATASET AND DATA PREPARATION

The dataset used in this study is the ESPN Soccer Dataset, publicly available via the Kaggle repository [34]. The dataset covers the 2024/2025 season and includes more than 30 000 matches across over 400 leagues worldwide. For this study, the sample was restricted to players from the five major European leagues: the English Premier League, Spanish La Liga, French Ligue 1, German Bundesliga, and Italian Serie A. After filtering, the final dataset consisted of 1988 unique outfield players. Goalkeepers were excluded because their performance metrics are fundamentally different and not directly comparable to those of outfield players. Additionally, players with fewer than three appearances were removed to avoid distortions caused by small sample sizes and unstable averages. Initial data preparation was conducted in Microsoft Excel, where raw performance indicators were transformed into per-appearance metrics. This approach was chosen over per-90-minute normalisation to better capture a player's typical contribution during a match, regardless of total playing time [35]. However, unlike per-90-minute metrics commonly used in football analytics, per-appearance measures do not fully control for differences in playing time within individual matches. As a result, players with shorter appearances may have a disproportionate influence on their average values. This limitation is acknowledged and should be considered when interpreting the results.

The dataset was subsequently imported into Python for final pre-processing and analysis. To ensure comparability across variables and to prevent scale-related bias in the clustering process, all features were standardised using z-score normalisation, resulting in variables with a mean of zero and unit variance [24]. This step is essential for distance-based algorithms such as K-means, where unscaled variables can disproportionately influence cluster formation.

FEATURE SELECTION

The analysis is based on eight performance metrics selected to capture key aspects of player behaviour during matches. These include total goals, total shots, shots on target, assists, offsides, fouls committed, fouls suffered, and yellow cards. All variables were calculated as per-appearance averages to account for differences in playing time across players. The selected variables were chosen to reflect a combination of offensive output, creative contribution, positional behaviour, and physical engagement. Goals, shots, and shots on target capture scoring activity and efficiency, while assists represent creative involvement. Offsides serve as a proxy for attacking positioning and forward movement, indicating how frequently a player operates near the defensive line. Fouls committed and fouls suffered reflect physical interaction and involvement in duels, while yellow cards provide an additional indicator of disciplinary behaviour. Rather than constructing composite indices or derived metrics, the study relies on directly observable match statistics to maintain transparency and interpretability. This approach ensures that the resulting clusters can be directly linked to specific on-field actions, facilitating practical interpretation of the identified player archetypes.

CLUSTERING PROCEDURE

K-means clustering was applied to the standardised dataset to identify groups of players with similar performance profiles. The algorithm partitions observations into K clusters by minimising within-cluster variance and assigns each player to the cluster with the nearest centroid. The optimal number of clusters was determined using two complementary diagnostic techniques: the Elbow method and the Silhouette score. The Elbow method evaluates the relationship between the number of clusters and the within-cluster sum of squares (WCSS), identifying a point at which additional clusters provide diminishing returns 0. The Silhouette score measures the degree of separation between clusters, indicating how well each observation fits within its assigned group compared to neighbouring clusters [27]. While lower values of K resulted in broader groupings that largely reflected general attacking and defensive distinctions, a five-cluster solution provided a more meaningful balance between interpretability and internal consistency. The Silhouette score reached a local maximum at $K = 5$, indicating improved cluster separation relative to neighbouring configurations. This solution enabled the identification of distinct and interpretable player profiles without introducing excessive fragmentation.

MODEL VALIDATION AND VISUALISATION

To assess the robustness and interpretability of the clustering results, both quantitative and qualitative validation approaches were employed. Quantitatively, the elbow method and silhouette scores were used to evaluate cluster structure and cohesion. These metrics provide complementary perspectives on cluster quality, ensuring that the selected solution is not only statistically valid but also analytically meaningful. Qualitatively, cluster centroids were analysed to interpret the behavioural characteristics of each group. Mean values of performance metrics within each cluster were examined to identify dominant patterns and define corresponding tactical archetypes. In addition, principal component analysis (PCA) was applied to reduce the dimensionality of the dataset and visualise the relative positioning of clusters in a two-dimensional space **Error! Reference source not found.**. This visualisation provides an intuitive representation of the separation and overlap between player groups, supporting the interpretation of cluster relationships and confirming the distinctiveness of the identified archetypes.

In addition to these diagnostics, it is important to note that clustering results may be sensitive to data preprocessing choices, feature construction, and random initialisation. While the present study focuses on interpretability and consistency of the identified profiles, future work should incorporate formal robustness checks, including stability analysis across multiple initialisations, alternative normalisation approaches, and comparison with different clustering algorithms.

RESULTS

MODEL VALIDATION AND SELECTION

The first stage of the analysis focused on determining the optimal number of clusters (K) for the K-means solution. To address this, two complementary diagnostic techniques were applied: the Elbow method and the Silhouette score. The Elbow method was used to examine the relationship between the number of clusters and model inertia (i.e., the within-cluster sum of squares). As shown in Figure 1, inertia declined substantially as K increased from 1 to 3, after which the rate of decrease became less pronounced. From a purely mathematical perspective, this pattern could suggest a lower-cluster solution, such as $K = 2$ or $K = 3$. However, although these smaller solutions improved parsimony, they did not provide sufficient analytical differentiation for this study. In substantive terms, lower values of K produced overly broad groupings that primarily reflected a general division between more defensive and more attacking player profiles. While

such a distinction is not without value, it does not capture the finer behavioural variation required for meaningful tactical archetyping in contemporary football.

For this reason, the Elbow method was complemented by an examination of Silhouette scores across cluster solutions ranging from K = 2 to K = 10. As shown in Figure 2, the Silhouette score improved notably at K = 5, indicating stronger cohesion within clusters and better separation between clusters compared with neighbouring solutions. This suggests that the five-cluster configuration achieved a more favourable balance between statistical consistency and interpretive usefulness. Accordingly, K = 5 was selected as the final clustering solution. This configuration enabled the analysis to move beyond broad positional distinctions and to identify more nuanced, behaviourally meaningful player archetypes. The selected solution was therefore not based solely on mathematical optimisation, but also on its capacity to generate analytically interpretable and tactically relevant groupings.

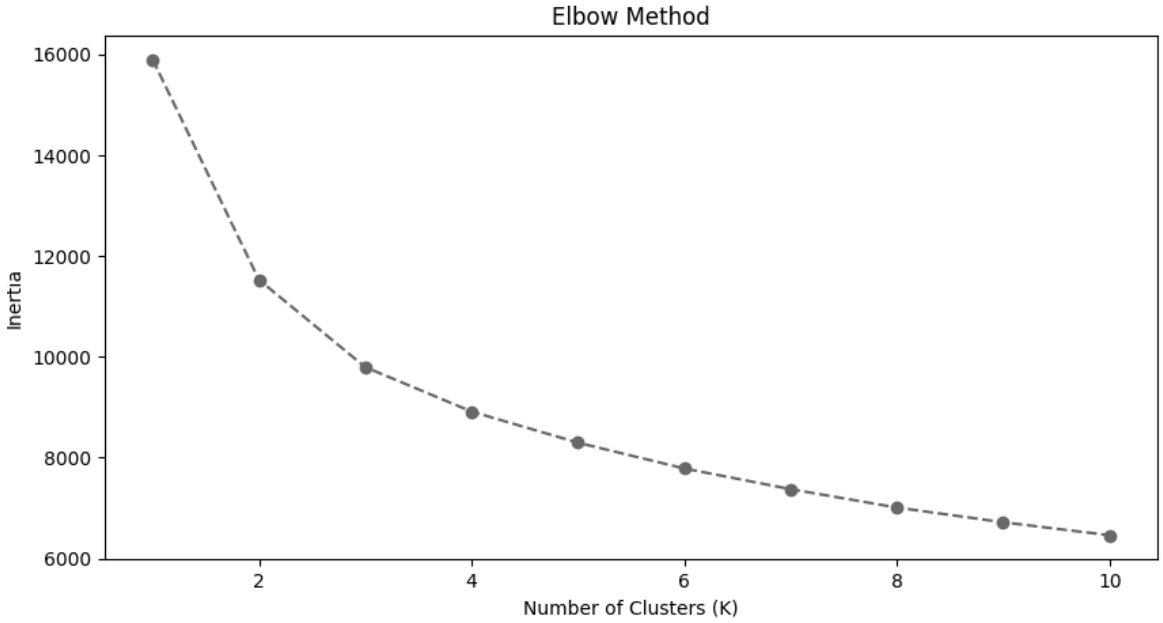

Figure 1. Elbow method plot.

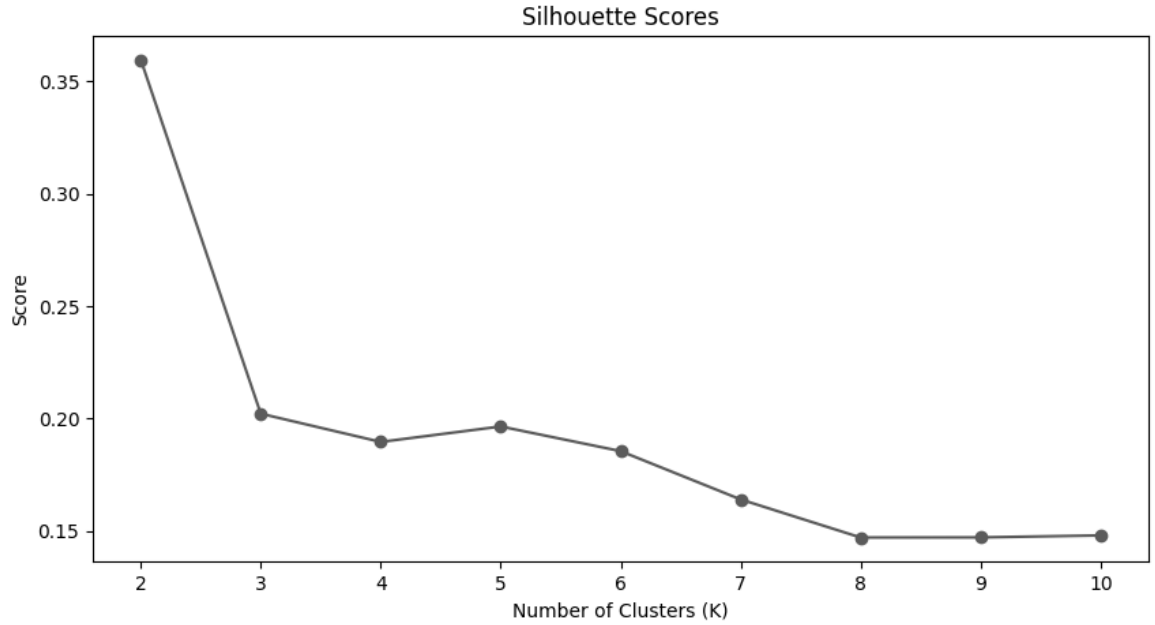

Figure 2. Silhouette scores plot.

CLUSTER CHARACTERISTICS

The mean values presented in Table 1 provide a clear basis for interpreting the five identified clusters in behavioural and tactical terms. The resulting groups extend beyond conventional positional labels such as midfielder or forward, as they reflect distinct patterns of match involvement rather than nominal role assignment.

Cluster 0 was labelled Conservative/Disciplined Players. This group is characterised by the lowest levels of physical and disciplinary involvement, with an average of 0,52 fouls committed and only 0,06 yellow cards per appearance. At the same time, these players show limited attacking output across goals, shots, and assists. Taken together, this profile suggests a low-risk and relatively restrained style of play, marked by positional stability and limited direct disruption or offensive initiative.

Cluster 1 was identified as Defensive Enforcers. In contrast to Cluster 0, this group records the highest average number of fouls committed (1,15) and yellow cards (0,24), indicating a substantially more aggressive and physically interventionist role. Although these players do not stand out in offensive production, their statistical profile suggests a tactical function centred on disruption, defensive resistance, and the interruption of opposition play.

The remaining three clusters reflect more clearly differentiated attacking roles. Cluster 2, labelled Primary Attackers/Finishers, displays the strongest scoring profile in the dataset, with the highest averages for goals (0,42), shots on target (1,03), total shots (2,35), and offsides (0,42). This combination indicates a role defined by direct attacking intent, frequent penetration of advanced spaces, and sustained involvement in finalising offensive actions.

Cluster 3 was labelled Active Wingers/Secondary Attackers. These players are characterised by a high level of attacking activity, particularly in terms of total shots (1,42) and fouls suffered (1,02). Their profile suggests players who are actively involved in progressive attacking movement and generate offensive pressure through repeated forward actions. Compared with Cluster 2, their contribution appears less focused on finishing efficiency and more on offensive dynamism and volume.

Table 1. Mean values of performance metrics for the five identified player clusters (based on ESPN Soccer Dataset [34]). All statistical values represent averages calculated on a per-appearance basis to account for varying player involvement across the season.

Cluster	cluster_role	Fouls Committed	Fouls Suffered	Yellow Cards	Goal Assists	Offsides	Shots On Target	Total Shots	Total Goals
0	Conservative/ Disciplined	0,518	0,445	0,064	0,023	0,054	0,138	0,462	0,037
1	Defensive Enforcers	1,147	0,821	0,238	0,029	0,050	0,143	0,531	0,035
2	Primary Attackers/ Finishers	0,965	1,118	0,110	0,153	0,417	1,031	2,351	0,418
3	Active Wingers/ Secondary Attackers	0,867	1,023	0,103	0,060	0,236	0,537	1,421	0,173
4	Creative Playmakers	0,815	0,824	0,117	0,197	0,087	0,273	0,867	0,074

Cluster 4 was interpreted as Creative Playmakers. This group records the highest average number of assists per appearance (0,20), while maintaining more moderate values for goals and shots relative to the more attack-oriented clusters. Such a pattern indicates a role focused

primarily on chance creation and connective play rather than direct finishing. From a tactical perspective, this cluster appears to represent players who function as creative intermediaries between ball progression and final attacking execution.

The cluster profiles show that the five-group solution captures substantive differences in player behaviour and tactical contribution. Rather than reproducing conventional positional categories, the model distinguishes between statistically grounded role profiles that reflect how players operate during matches. This makes the clustering solution more suitable for identifying functional similarities and differences among players than traditional position-based classification.

VALIDATION THROUGH OBSERVATION

To assess whether the identified clusters were meaningful in practical football terms, a small set of recognisable players was selected from each group for illustrative purposes (see Table 2). The clustering procedure was based solely on the eight performance variables used in the analysis, all expressed per appearance. During computation, each observation was identified solely by the athleteID variable, meaning the model grouped players solely by their statistical profiles rather than their names, reputations, or perceived roles.

Player names were linked to cluster assignments only after the clustering procedure was completed. This two-step approach helped preserve the objectivity of the analysis and allowed the resulting groups to be interpreted independently of prior expectations. The examples shown in Table 2 indicate that the clusters correspond to recognisable patterns of football behaviour. For instance, the placement of Mohamed Salah and Bruno Fernandes in the Primary Attackers/Finishers cluster is consistent with their strong attacking output, while the inclusion of Joe Gomez in the Creative Playmakers cluster shows that data-driven role assignment does not necessarily align with official position labels.

This is particularly important for interpreting the results. The clustering solution does not classify players by their position on a team sheet, but rather by how they contribute during matches. In this sense, the observed examples provide practical support for the study's central argument: match statistics can reveal tactical utility in ways that conventional positional categories often cannot.

Table 2. Representative players by data-derived tactical cluster (based on ESPN Soccer Dataset [34])

Full Name	Athlete Id	Cluster	cluster_role
Michael Keane	4 946	0	Conservative/Disciplined
James Milner	26 843	0	Conservative/Disciplined
Ilia Gruev	30 116	0	Conservative/Disciplined
Ryan Christie	93 435	0	Conservative/Disciplined
Matt Doherty	43 575	1	Defensive Enforcers
David Brooks	72 487	1	Defensive Enforcers
Cristian Romero	96 970	1	Defensive Enforcers
Kyle Walker	129 358	1	Defensive Enforcers
Mohamed Salah	173 896	2	Primary Attackers/Finishers
Bruno Fernandes	124 091	2	Primary Attackers/Finishers
Harry Wilson	194 801	2	Primary Attackers/Finishers
Leandro Trossard	174 980	2	Primary Attackers/Finishers
Evanilson	102 368	3	Active Wingers/Secondary Attackers
Chris Wood	134 190	3	Active Wingers/Secondary Attackers
Callum Wilson	138 924	3	Active Wingers/Secondary Attackers
Danny Welbeck	115 271	3	Active Wingers/Secondary Attackers
Jordan Henderson	127 262	4	Creative Playmakers
Granit Xhaka	149 981	4	Creative Playmakers
Wataru Endo	152 479	4	Creative Playmakers
Joe Gomez	102 053	4	Creative Playmakers

While Table 2 highlights specific examples for clarity, the full classification of all 1988 players and their corresponding cluster assignments is available in Digital Appendix A.

VISUALISATION OF TACTICAL SPACES

To examine the spatial relationships among the identified clusters, PCA was used to reduce dimensionality. PCA transformed the original eight standardised performance variables into two principal components, enabling the multidimensional clustering solution to be represented in a two-dimensional space [35]. The resulting visualisation, presented in Figure 3, provides a simplified two-dimensional visual representation of the clustering structure of player profiles and helps illustrate the relative proximity, separation, and overlap between clusters.

The PCA projection reveals a meaningful visible separation in the projection between player groups with predominantly defensive and predominantly attacking profiles. On one side of the plot, the Conservative/Disciplined Players and Defensive Enforcers occupy more compact and dense regions. This concentration suggests that these clusters are relatively homogeneous, meaning that players assigned to them tend to exhibit more similar patterns of behaviour. In practical terms, this indicates that lower attacking involvement combined with either restrained or aggressive defensive engagement forms relatively stable and recognisable behavioural profiles.

By contrast, the more attack-oriented clusters are distributed more broadly across the tactical space. The Primary Attackers/Finishers cluster extends toward the part of the plot associated with stronger offensive output, reflecting elevated values in goals, total shots, shots on target, and offsides. This wider spread indicates greater internal variation among attacking players, consistent with the reality that offensive contributions can take multiple forms even within a broadly similar tactical function. Some players in this cluster may operate as direct central finishers, while others may combine scoring with more mobile or hybrid attacking movement.

A similar but distinct pattern is visible for the Active Wingers/Secondary Attackers cluster. These players also occupy an attacking part of the visual projection, but their distribution suggests a profile that is less concentrated around pure finishing and more associated with activity, movement, and repeated offensive involvement. Their relative position in the PCA space supports the interpretation that this cluster captures players who generate attacking pressure through volume and dynamism rather than through highly concentrated scoring efficiency alone.

The location of the Creative Playmakers cluster is particularly informative. These players occupy a more central region in the projection of the visual projection, positioned between the more defensive clusters and the more direct attacking groups. This centrality is analytically meaningful because it reflects the intermediary tactical role suggested by their statistical profile. Rather than being defined by either defensive disruption or goal-oriented finishing, these players appear to function as connectors between phases of play, contributing to chance creation and attacking progression while maintaining a balanced presence across several dimensions of performance. Their placement in the PCA plot visually reinforces the interpretation of this cluster as a intermediate statistical profile within the broader tactical structure.

At the same time, the visualisation shows that the cluster boundaries are not completely isolated. A limited degree of overlap is visible at the margins of several groups, particularly between the more advanced attacking clusters and the centrally positioned playmakers. This is not a methodological weakness, but rather a realistic reflection of the fluidity of football roles. Modern players rarely conform to entirely discrete functional categories, and partial overlap between behavioural profiles is therefore expected. What matters is that the central mass of each cluster occupies a distinguishable region of the tactical space, indicating that the clustering solution captures meaningful structural differences despite the natural continuity of player roles.

It should be noted that the PCA visualisation is used for illustrative purposes only and does not represent an independent validation of the clustering structure.

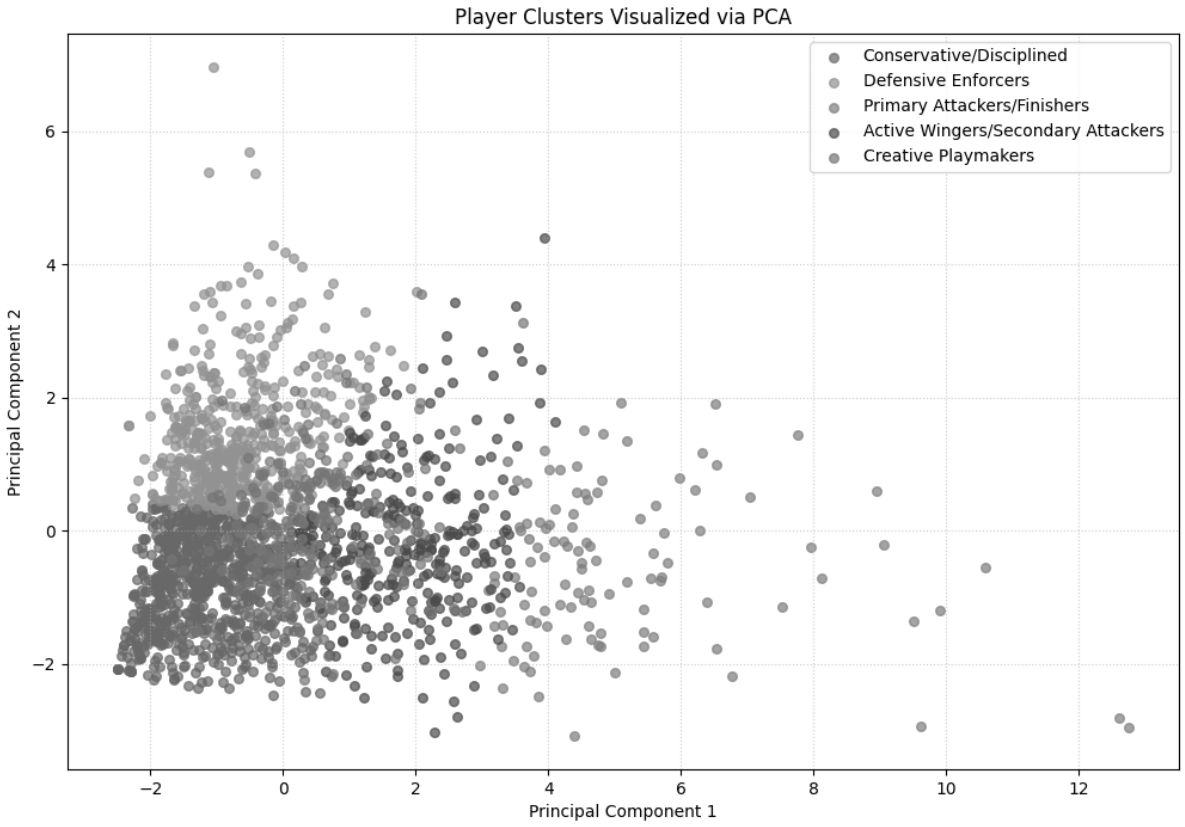

Figure 3. Player clusters visualised via PCA.

DISCUSSION

The findings of this study support the growing body of literature showing that data-driven approaches can provide a more precise understanding of football performance than traditional descriptive categories. Prior research has already demonstrated the broader expansion of big data, artificial intelligence, and machine learning in sport analytics, including football-specific applications in prediction, valuation, and decision support [13, 15]. However, much of this literature still operates within inherited football taxonomies, where players are evaluated as defenders, midfielders, or forwards. The present results suggest that such labels are increasingly insufficient for capturing actual tactical behaviour. Instead, the clustering solution indicates that players can be more meaningfully understood through behavioural profiles derived directly from match statistics. In this sense, the study directly shows that role-based groupings can emerge from performance data even when positional labels are not included in the model.

This finding is consistent with research emphasising the increasing complexity and fluidity of football roles, as well as the broader evolution of sports analytics toward multidimensional performance modelling [1, 4]. In this study, the five identified clusters do not simply reproduce conventional positional divisions. Rather, they reveal functional archetypes such as Creative Playmakers, Defensive Enforcers, and Primary Attackers/Finishers, each characterised by a distinct statistical signature. This is an important contribution because it shows that unsupervised learning can move analysis beyond nominal squad labels and toward role definitions grounded in actual on-pitch activity. In that sense, the results align with clustering-based studies in sport that have shown the value of unsupervised grouping for identifying meaningful performance structures [20, 21], while also extending that line of work by focusing specifically on tactical archetyping in elite European football.

A particularly relevant result is the clear separation between disruptive defensive profiles and high-output attacking profiles. Defensive Enforcers are statistically defined by a higher frequency of fouls committed and yellow cards, reflecting a role centred on physical interruption and defensive intensity. By contrast, Primary Attackers/Finishers are distinguished by elevated goals, shots, shots on target, and offsides, indicating an aggressive and vertically oriented attacking role. This distinction is analytically useful because it demonstrates that behavioural clusters can capture differences in tactical function that would be obscured if players were grouped only by nominal positions. From a practical perspective, this strengthens the argument that recruitment and squad planning should move toward function-based replacement logic. Replacing a player should not mean signing another athlete with the same listed position, but rather identifying a player with a similar behavioural and statistical profile. This interpretation is consistent with the increasing use of machine learning as a decision-support tool in football management and performance analysis [16, 17]. It also suggests that the practical value of clustering lies not only in classification, but in translating performance data into tactically usable categories.

The emergence of Creative Playmakers as a distinct cluster is also theoretically important. These players occupy an intermediary tactical space, with relatively strong assist output but more moderate finishing indicators. Their statistical profile suggests a intermediate statistical profile between possession development and final attacking execution. This aligns with the broader literature, which suggests that modern football performance should be understood in terms of functional contributions rather than static formation-based labels [2]. It also illustrates one of the key advantages of clustering methods: they allow latent role structures to emerge from data rather than forcing observations into predefined categories [7, 9]. In that respect, the study confirms the wider methodological value of unsupervised learning in football analytics, especially when the objective is not prediction but discovery of hidden structure.

At the same time, the results should be interpreted with the limitations of the underlying dataset in mind. The selected ESPN data provide a useful basis for identifying attacking and disciplinary tendencies, but they do not capture the full complexity of defensive or transitional performance. Variables such as interceptions, recoveries, pressures, duel success, spatial occupation, or progressive passing are absent. As a result, the two more defensive clusters are likely broader and less differentiated than the attacking clusters. This helps explain why highly specific defensive roles, such as ball-winning midfielders, screening pivots, or positional defenders, did not emerge as separate groups. Methodologically, this is consistent with the literature, which shows that clustering results are highly sensitive to feature selection, standardisation, and the representation of underlying behaviour [25, 29]. In other words, the cluster structure obtained here reflects not only the underlying tactical reality but also the boundaries imposed by the available variables.

This limitation does not diminish the model's practical value, but it does define its scope. The five-cluster solution should therefore be interpreted as a meaningful but partial map of tactical identity, one that is especially effective in distinguishing offensive contribution patterns and broad disciplinary-defensive behaviours. Future research could improve this framework by integrating richer event and tracking data and by testing alternative clustering techniques, including more advanced approaches that may capture more complex behavioural structures [23]. A longitudinal extension would also be valuable, since it would allow researchers to examine whether players transition between archetypes over time, for example, from high-intensity wide attackers to more centrally oriented creative roles.

The findings therefore suggest that tactical archetyping via K-means clustering offers a conceptually and practically useful alternative to conventional player categorisation. Even a relatively compact set of match statistics was sufficient to reveal coherent and interpretable

behavioural groups. More importantly, the study points toward a broader shift in football analytics: from asking what position a player occupies to asking what function the player performs.

CONCLUSION

This study demonstrates that unsupervised machine learning, specifically K-means clustering, can effectively identify data-driven tactical archetypes in professional football. By analysing performance metrics independently of positional labels, the results reveal five distinct player profiles that reflect actual on-pitch behaviour rather than nominal roles. This finding contributes to the growing literature on sports analytics and machine learning by showing that even relatively simple statistical inputs can uncover meaningful structural patterns in player performance.

The study extends existing research on machine learning applications in football, which has predominantly focused on prediction, valuation, and decision support, by shifting the analytical focus toward role discovery. In doing so, it aligns with and builds upon prior clustering-based approaches while addressing a key limitation in the literature: the continued reliance on predefined positional categories. The results suggest that a behavioural, data-driven classification framework offers a more precise and operationally relevant understanding of player roles, particularly in the context of modern, tactically fluid football systems.

From a practical perspective, the findings have implications for scouting, recruitment, and squad composition. Identifying players based on statistical similarity rather than positional labels enables more accurate replacement strategies and supports a function-based approach to team building. This is particularly relevant in high-performance environments where marginal gains and tactical balance are critical. The framework proposed in this study provides an interpretable basis for integrating behavioural profiling into decision-support processes within football organisations.

At the same time, the study is subject to several limitations. Reliance on the ESPN dataset limits the analysis to a subset of performance indicators, primarily related to attacking output and disciplinary actions. As a result, more nuanced defensive and off-ball contributions are not fully captured, which affects the granularity of defensive role differentiation. In addition, the use of K-means clustering, while methodologically transparent and interpretable, assumes relatively simple cluster structures and may not fully capture more complex relationships in high-dimensional performance data. Future research should therefore incorporate richer datasets, including tracking data and advanced event metrics, and explore alternative clustering techniques that may provide additional analytical depth. Longitudinal analyses would also be valuable in examining how player roles evolve over time and across different tactical systems. In addition, the use of per-appearance metrics instead of per-90 normalisation may introduce bias related to unequal playing time across appearances. Future research should therefore explicitly compare alternative normalisation strategies as part of robustness testing.

This study supports a broader shift in football analytics from position-based classification toward behaviour-based modelling. By showing that tactical roles can emerge directly from performance data, it contributes to a more flexible and tactically meaningful understanding of player performance.

SUPPLEMENTARY MATERIAL: FULL DATASET AND CLUSTER ASSIGNMENTS

The full list of 1988 players, their athlete IDs, and their assigned tactical clusters is available in a digital spreadsheet as supplementary material. Access link: https://drive.google.com/file/d/108XbDpnNK3bXQX8cowuUsvryF5a0G001/view?usp=drive_link.

REFERENCES

[1] Popovych, I., et al.: *Operationalization of tactical thinking of football players by main game roles.* Journal of Physical Education and Sport **21**(5), 2480-2491, 2021, http://dx.doi.org/10.7752/jpes.2021.05334,

[2] Andrienko, G., et al.: *Constructing Spaces and Times for Tactical Analysis in Football*. IEEE Transactions on Visualization and Computer Graphics **27**(4), 2280-2297, 2019, http://dx.doi.org/10.1109/TVCG.2019.2952129,

[3] Brîndescu, S.; Datcu, F.-R. and Buda, I.-A.: *Study on the efficiency of advanced pressing in the Premier League.* Baltic Journal of Health and Physical Activity **13**(Spec.Iss.1), 115-122, 2021, http://dx.doi.org/10.29359/BJHPA.13.Spec.Iss1.11,

[4] He, Q.; Araújo, D.; Davids, K.; Kee, Y.H. and Komar, J.: *Functional adaptability in playing style: A key determinant of competitive football performance.* Adaptive Behavior **31**(6), 545-558, 2023, http://dx.doi.org/10.1177/10597123231178942,

[5] Herberger, T.A. and Litke, C.: *The Impact of Big Data and Sports Analytics on Professional Football: A Systematic Literature Review.* In: Herberger, T.A. and Dötsch, J.J., eds.: *Digitalization, Digital Transformation and Sustainability in the Global Economy*. Springer Proceedings in Business and Economics. Springer, Cham, pp.147-171, 2021, http://dx.doi.org/10.1007/978-3-030-77340-3_12,

[6] Lüdin, D.; Donath, L.; Cobley, S.; Mann, D. and Romann, M.: *Player-labelling as a solution to overcome maturation selection biases in youth football.* Journal of Sports Sciences **40**(14), 1641-1647, 2022, http://dx.doi.org/10.1080/02640414.2022.2099077,

[7] Oyewole, G.J. and Thopil, G.A.: *Data clustering: Application and trends*. Artificial Intelligence Review **56**(7), 6439-6475, 2023, http://dx.doi.org/10.1007/s10462-022-10325-y,

[8] Fernández Martínez, D.; Casals Toquero, M.; Oliver, M.; Plensa, M. and Manisera, M.: *Reporting of clustering techniques in sports sciences: A scoping review*. Electronic Journal of Applied Statistical Analysis **17**(3), 653-675, 2024, http://dx.doi.org/10.1285/i20705948v17n3p653,

[9] Sinaga, K.P. and Yang, M.-S.: *Unsupervised K-Means Clustering Algorithm.* IEEE Access **8**, 80716-80727, 2020, http://dx.doi.org/10.1109/ACCESS.2020.2988796,

[10] Andaur Navarro, C.L., et al.*: Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review*. BMJ **375**, No. n2281, 2021, http://dx.doi.org/10.1136/bmj.n2281,

[11] Shen, Q., et al.: *Classifying Player Profiles in Elite Women's Football: A K-Means Clustering Analysis of Physical and Technical Data from the 2023 FIFA Women's World Cup.* Football Studies **1**, No. 100026, 2026, http://dx.doi.org/10.1016/j.footst.2026.100026,

[12] Sotudeh, H.: *The principles of tactical formation identification in association football (soccer) – a survey.* Frontiers in Sports and Active Living **6**, No. 1512386, 2025, http://dx.doi.org/10.3389/fspor.2024.1512386,

[13] Pejić Bach, M.; Ivec, A. and Hrman, D.: *Industrial Informatics: Emerging Trends and Applications in the Era of Big Data and AI*. Electronics **12**(10), No. 2238, 2023, http://dx.doi.org/10.3390/electronics12102238,

[14] Šuštaršič, A.; Videmšek, M.; Karpljuk, D.; Miloloža, I. and Meško, M.: *Big data in sports: A bibliometric and topic study*.
Business Systems Research **13**(1), 19-34, 2022,
http://dx.doi.org/10.2478/bsrj-2022-0002,

[15] Chang, V.; Hall, K. and Doan, L.M.T.: *Football results prediction and machine learning techniques.*
International Journal of Business and Systems Research **17**(5), 565-586, 2023,
http://dx.doi.org/10.1504/IJBSR.2023.133178,

[16] Al-Asadi, M.A. and Tasdemır, S.: *Predict the value of football players using FIFA video game data and machine learning techniques*.
IEEE Access **10**, 22631-22645, 2022,
http://dx.doi.org/10.1109/ACCESS.2022.3154767,

[17] Nikić, A.; Topalović, A. and Pejić Bach, M.: *From data to decision: machine learning in football team management.*
In: *47th MIPRO ICT and Electronics Convention (MIPRO)*. IEEE, Opatija, pp.1059-1064, 2024,
http://dx.doi.org/10.1109/MIPRO60963.2024.10569835,

[18] Teixeira, J.E., et al.: *Data mining paths for standard weekly training load in sub-elite young football players: A machine learning approach.*
Journal of Functional Morphology and Kinesiology **9**(3), No. 114, 2024,
http://dx.doi.org/10.3390/jfmk9030114,

[19] Medić, T.; Pavlečić, A. and Topalović, A: *MLP Neural Networks vs. Logistic Regression: A Comparative Study of Customer Churn Prediction in Bank Marketing*.
ENTRENOVA – ENTerprise REsearch InNOVAtion Journal **11**(1), 2025,
http://dx.doi.org/10.54820/entrenova-2025-0088,

[20] Wing, C.; Hart, N.H.; Fu, S.C.; Nosaka, K. and Ma'ayah, F.: *The use of K-means clustering for the organisation of training groups in Australian football*.
International Journal of Sports Science and Coaching **20**(4), 1642-1650, 2025,
http://dx.doi.org/10.1177/17479541251333925,

[21] Khariri, A.F.; Intan, P.K.; Hafiyusholeh, M.; Asyhar, A.H. and Fanani, A.: *Implementation of K-Means Particle Swarm Optimization for Clustering Football Players in the Top Five European Football Leagues.*
In: Adzkiya, D. and Fahim, K., eds.: *Applied and Computational Mathematics. ICoMPAC 2023*. Springer Proceedings in Mathematics & Statistics **455**. Springer, Singapore, pp.63-75, 2023,
http://dx.doi.org/10.1007/978-981-97-2136-8_6,

[22] Prasetyo, E.; Priyatama, A.S. and Setyatama, F.: *Constellation of Football Players Determination Based on Cost and Performance History Using the K-Means Clustering.*
Digital Zone: Jurnal Teknologi Informasi dan Komunikasi **14**(2), 194-205, 2023,
http://dx.doi.org/10.31849/digitalzone.v14i2.17106,

[23] Demir, E.; Şahin, Y.H. and Üre, N.K.: *How Do Football Teams Play? A Deep Embedded Clustering Approach to Reveal Playing Styles*.
In: Dong, Js.; Sun, J.; Xie, X. and Jiang, K., eds.: *Sports Analytics. ISACE 2025*. Lecture Notes in Computer Science **15925**. Springer, Cham, pp.53-68, 2025,
http://dx.doi.org/10.1007/978-3-032-06167-6_4,

[24] Al-Faiz, M.Z.; Ibrahim, A.A. and Hadi, S.M.: *The effect of Z-Score standardization (normalization) on binary input due the speed of learning in back-propagation neural network.*
Iraqi Journal of Information and Communication Technology **1**(3), 42-48, 2018,
http://dx.doi.org/10.31987/ijict.1.3.41,

[25] Dalatu, P. and Midi, H.: *New approaches to normalization techniques to enhance K-means clustering algorithm*.
Malaysian Journal of Mathematical Sciences **14**(1), 41-62, 2020,

[26] Saputra, D.M.; Saputra, D. and Oswari, L.D.: *Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method.*
In: *Sriwijaya International Conference on Information Technology and Its Applications SICONIAN 2019*. Atlantis Press, pp.341-346, 2020,
http://dx.doi.org/10.2991/aisr.k.200424.051,

[27] Januzaj, Y.; Beqiri, E. and Luma, A.: *Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique.*
International Journal of Online and Biomedical Engineering **19**(4), 174-182, 2023,
http://dx.doi.org/10.3991/ijoe.v19i04.37059,

[28] Dalmaijer, E.S.; Nord, C.L. and Astle, D.E.: *Statistical power for cluster analysis.*
BMC Bioinformatics **23**, No. 205, 2022,
http://dx.doi.org/10.1186/s12859-022-04675-1,

[29] Foss, A.H.; Markatou, M. and Ray, B.: *Distance Metrics and Clustering Methods for Mixed-type Data*.
International Statistical Review **87**(1), 80-109, 2019,
http://dx.doi.org/10.1111/insr.12274,

[30] Rogić, S.; Kašćelan, L. and Pejić Bach, M.: *Customer response model in direct marketing: solving the problem of unbalanced dataset with a balanced support vector machine.*
Journal of Theoretical and Applied Electronic Commerce Research **17**(3), 1003-1018, 2022,
http://dx.doi.org/10.3390/jtaer17030051,

[31] Marčinko Trkulja, Ž.; Primorac, D. and Martinčević, I.: *The influence of consumer motivation on engagement with sports club social media: An intrinsic and extrinsic analysis*.
Business Systems Research **15**(1), 91-109, 2024,
http://dx.doi.org/10.2478/bsrj-2024-0005,

[32] Ma, Z. and Gu, B.: *The influence of firm-generated video on user-generated video: Evidence from China*.
International Journal of Engineering Business Management **14**, 2022,
http://dx.doi.org/10.1177/18479790221118628,

[33] Pejić Bach, M.; Meško, M.; Zoroja, J.; Godnov, U. and Ćurlin, T.: *Usage of simulation games in higher educational institutions teaching economics and business*.
ENTRENOVA – ENTerprise REsearch InNOVAtion Journal **6**(1), 27-36, 2020,

[34] Excel4soccer: *ESPN Soccer Data. Version 1*.
https://www.kaggle.com/datasets/excel4soccer/espn-soccer-data,

[35] Decroos, T.; Bransen, L.; Van Haaren, J. and Davis, J.: *Actions Speak Louder than Goals: Valuing Player Actions in Soccer.*
In: *KDD'19* Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, New York, pp.1851-1861, 2019,
http://dx.doi.org/10.1145/3292500.3330758,

[36] Penkova, T.G.: *Principal component analysis and cluster analysis for evaluating the natural and anthropogenic territory safety.*
Procedia Computer Science **112**, 99-108, 2017,
http://dx.doi.org/10.1016/j.procs.2017.08.179.