Journal of Informatics, Computer Science, Data Science And Arificial Intelligence
Volume 4 No. 1 June 2026, ISSN : 3031-9145

JICDA
ISSN : 3031-9145
Journal of Informatics, Computer Science, Data Science, and Artificial Intelligence

POSITION-BASED PLAYER PROFILING USING UNSUPERVISED LEARNING ON 2026 WORLD CUP PERFORMANCE DATA

Ichsan Firmansyah [1], Karina Andriani [2]
[1,2] STMIK Triguna Dharma
[1] ichsanfirmansyah1989@gmail.com, [2] ndrianikarina@gmail.com

ARTICLE INFO

ARTICLE HISTORY:
Received : 29 June 2026
Revised : 20 June 2026
Accepted : 10 June 2026

Keywords:
Unsupervised Learning;
K-Means Clustering;
Player Profiling;
Football Analytics;
2026 World Cup;
Performance Data

ABSTRACT

This study employs unsupervised learning to profile football players based on performance data from the 2026 FIFA World Cup, addressing a gap in the literature concerning player categorization in international tournament contexts. The dataset comprises 28 performance metrics from 512 players across 32 national teams. Following data preprocessing, K-Means clustering and principal component analysis were applied to identify natural groupings. The optimal cluster count, determined through elbow method, silhouette score (0.42), Davies-Bouldin index, and Calinski-Harabasz index, was established at k=4. This resulted in four distinct player archetypes: Defensive Anchors, characterized by strong defensive contributions and high passing accuracy; Utility Players, demonstrating versatility and high work rate but lower overall ratings; Star Forwards, exhibiting exceptional offensive output and the highest player ratings; and Supporting Attackers, showing moderate attacking contributions with limited playing time. A cross-tabulation with traditional positions revealed substantial but imperfect alignment, with goalkeepers nearly perfectly classified, while midfielders showed the most dispersed distribution. The findings validate the multidimensional nature of football performance and demonstrate the efficacy of unsupervised learning for objective player evaluation, offering practical applications for talent identification, tactical planning, and player development.

INTRODUCTION

The 2026 FIFA World Cup represents the pinnacle of international football, showcasing the world's finest talent across 32 nations competing on the global stage. This premier tournament provides an unprecedented opportunity to analyze player performance at the highest level of competition, where the margin between success and failure is remarkably narrow (Bar-On & Reiche, 2026). In modern football analytics, the ability to objectively evaluate and categorize player performance has become increasingly critical for talent identification, tactical planning, and strategic decision-making (Goes et al., 2021).

Traditional approaches to player evaluation have predominantly relied on subjective assessments from coaches, scouts, and pundits, often influenced by individual biases, limited match observation, and heuristic-based judgments (Wakelam et al., 2022). These conventional methods, while valuable, frequently fail to capture the multidimensional nature of football performance, where players contribute to team success through diverse and often non-intuitive combinations of skills and attributes (Teixeira et al., 2025). The complexity of modern football demands more sophisticated analytical frameworks that can process vast amounts of performance data and identify meaningful patterns that may not be immediately apparent to human observers (Alves et al., 2025).

The emergence of advanced data collection technologies and comprehensive performance tracking systems has revolutionized football analytics, enabling the capture of granular, real-time data on every aspect of player activity during matches (Araújo et al., 2021). These technological advancements have generated unprecedented volumes of performance metrics, creating both opportunities and challenges for the analytical community (Pu et al., 2024). While the availability of detailed performance data has expanded considerably, the methodological approaches for extracting actionable insights from this wealth of information remain an active area of research and development (Beato et al., 2025).

Machine learning and unsupervised learning techniques, in particular, offer promising avenues for discovering hidden patterns and natural groupings within complex performance datasets (Leist et al., 2022). Unlike supervised approaches that require predefined labels, unsupervised methods can identify inherent structures and relationships in data without prior assumptions about category membership (Verma et al., 2022). This characteristic makes unsupervised learning particularly well-suited for player profiling applications, where the optimal categorization of playing styles and performance patterns may not align with traditional positional classifications (Barrera et al., 2021).

Several studies have demonstrated the effectiveness of clustering techniques in sports analytics, including applications in basketball player categorization (Sarlis & Tjortjis, 2024), soccer tactical analysis (Huang et al., 2026), and performance benchmarking across different leagues and competitions (Akhanli & Hennig, 2023). For instance, K-Means clustering has been successfully employed to identify distinct playing styles among football players based on technical and physical performance indicators (Pacifico, 2025). Similarly, principal component analysis (PCA) has proven valuable for dimensionality reduction, enabling the visualization of complex player performance spaces and the identification of key performance dimensions that differentiate player types (Pino-Ortega et al., 2021).

However, significant gaps remain in the current literature regarding player profiling in international tournament contexts (Lepschy et al., 2021). Most existing research has focused on domestic league performances, where the competitive environment and tactical contexts differ substantially from the unique pressures and dynamics of World Cup competition (Morgans et al., 2024). The intensity, stakes, and limited preparation time characteristic of international tournaments create distinct performance patterns that warrant specific investigation (Stafylidis et al., 2025). Furthermore, the diversity of playing styles, tactical philosophies, and competitive standards represented across different footballing nations adds additional complexity to the analysis of World Cup performance data (Martinez-Mejorado & Ramirez-Marquez, 2023).

The 2026 World Cup dataset presents a particularly valuable opportunity for such analysis, as it encompasses comprehensive performance metrics from the highest level of international football, including detailed offensive, defensive, physical, and disciplinary statistics (Esh et al., 2026). This rich dataset enables the application of advanced analytical techniques to identify natural groupings of players based on their performance profiles, potentially revealing distinct player archetypes that transcend conventional positional classifications. Such analysis could provide valuable insights for coaches, scouts, and team managers seeking to optimize player selection, tactical deployment, and strategic planning.

This research addresses the identified gaps by applying unsupervised learning techniques, specifically K-Means clustering and principal component analysis, to profile players based on their performance data from the 2026 FIFA World Cup. The primary objectives of this study are twofold: first, to identify natural groupings of players based on multidimensional performance characteristics using K-Means clustering; and second, to visualize and interpret these clusters through PCA dimensionality reduction, enabling the discovery of meaningful player archetypes that may complement traditional positional classifications. For academic purposes, this study contributes to the growing body of literature on sports analytics and the application of machine learning techniques to football performance data. For practitioners, the findings may inform more objective and data-driven approaches to player evaluation, squad selection, and tactical planning. For the broader football community, this research demonstrates the potential of unsupervised learning techniques to reveal hidden patterns in player performance, potentially challenging conventional wisdom about player categorization and role definitions in modern football.

RESEARCH METHODS

Data Collection and Dataset Description

The dataset utilized in this research was obtained from Kaggle, a prominent platform for data science competitions and publicly available datasets (Rambe, 2026). The dataset comprises performance statistics for players participating in the 2026 FIFA World Cup, capturing comprehensive metrics across multiple dimensions of football performance. The dataset includes 28 distinct features covering player identification, demographic characteristics, physical attributes, offensive contributions, defensive actions, disciplinary records, physical performance metrics, and overall player evaluation.

The complete feature set is organized into the following categories:

1. Identification and Demographic Attributes
   The dataset begins with identification and demographic attributes, which include a unique Player ID serving as the primary identifier for each individual, Player Name for easy reference, and National Team affiliation indicating the country represented in the tournament. The Position variable captures the player's primary playing position, categorized into the standard football classifications of Goalkeeper, Defender, Midfielder, and Forward. Additionally, demographic information is provided through Age, Height, and Weight measurements, which offer insights into the physical characteristics of the players participating in the tournament.
2. Match Participation and Playing Time
   In terms of match participation and playing time, the dataset records the number of Matches Played by each player throughout the tournament, alongside the total Minutes Played. These metrics are fundamental for contextualizing other performance statistics, as they account for variations in playing time that naturally occur across positions and tactical decisions by team managers. Players with substantial minutes provide more reliable samples of their performance capabilities compared to those with limited appearances.
3. Offensive Performance Metrics
   The offensive performance metrics constitute one of the most comprehensive

dimensions of the dataset, capturing both traditional goal-scoring statistics and advanced analytical indicators. Goals and Assists represent the most direct contributions to attacking play, while Shots and Shots on Target provide insight into attacking volume and efficiency. Passing statistics are represented through Passes Completed and Pass Accuracy (%), which reflect a player's technical proficiency and decision-making in possession. Key Passes indicate passes that lead directly to scoring opportunities, while Dribbles Completed measures the player's ability to progress the ball through individual skill. Advanced metrics include Expected Goals (xG) and Expected Assists (xA), which evaluate the quality of chances created and the likelihood of scoring, offering a more nuanced assessment of attacking contribution that accounts for the difficulty of opportunities created.

4. Defensive Performance Metrics
   The defensive performance metrics capture the player's contributions to preventing opposition goals and regaining possession. Tackles and Interceptions quantify direct defensive actions, while Clearances measure the ability to relieve pressure in defensive situations. For goalkeepers specifically, Saves records are provided to evaluate shot-stopping ability, while Clean Sheets indicate matches where the goalkeeper and defensive unit prevented the opposition from scoring. These defensive metrics are particularly valuable for profiling players whose primary contributions may not be reflected in offensive statistics.
5. Disciplinary and Physical Metrics
   Disciplinary and physical metrics provide additional dimensions of player behavior and athletic output. Yellow Cards and Red Cards capture disciplinary infractions, which can indicate playing style, aggression, or tactical fouling patterns. Physical performance is quantified through Distance Covered and Sprint Count, reflecting the player's work rate and athletic output during matches. These metrics have become increasingly important in modern football analysis, as physical performance is closely linked to tactical execution and overall effectiveness.
6. Overall Performance Evaluation
   Finally, the overall performance evaluation is captured through Player Rating, a composite metric that aggregates various performance dimensions into a single numerical score. This rating serves as an intuitive summary of the player's overall contribution and provides a valuable reference point for validating and interpreting the results of the clustering analysis. The inclusion of this rating enables comparison between algorithmic cluster assignments and expert-derived performance evaluations, offering a form of external validation

for the clustering results.

Benchmark Studies and Methodological Foundations

This research is positioned within a growing body of literature that applies unsupervised learning techniques to player profiling in professional sports. Several benchmark studies provide the methodological foundation and comparative context for this investigation.

(Pacifico, 2025) conducted a seminal study applying K-Means clustering to profile professional soccer players in European leagues. Their analysis demonstrated that clustering techniques could effectively identify distinct player archetypes based on technical and physical performance indicators, with a Silhouette score of 0.39 indicating moderate cluster separation. Their study established important methodological precedents, including feature selection strategies, data preprocessing approaches, and cluster interpretation frameworks that inform the present investigation.

(Cintia et al., 2021) developed PlayeRank, a data-driven performance evaluation system using machine learning approaches for player ranking in soccer. Their work demonstrated the effectiveness of combining multiple performance dimensions to create comprehensive player profiles. While their approach utilized supervised learning techniques, their feature engineering strategies and performance evaluation frameworks provide valuable methodological insights for unsupervised clustering applications.

(Stafylidis et al., 2025) conducted a comprehensive analysis of match statistics related to winning in the 2014 FIFA World Cup, establishing important benchmarks for World Cup performance analysis. Their examination of key performance indicators at the tournament level provides contextual grounding for interpreting clustering results in the specific competitive environment of international football.

(Gu et al., 2025) provided a systematic review of unsupervised learning approaches to player classification in elite sport, identifying key methodological considerations, common pitfalls, and best practices. Their review highlights the importance of appropriate feature selection, careful preprocessing, and robust validation strategies—all of which inform the methodology employed in this research.

This study extends these benchmark works by applying unsupervised learning specifically to World Cup data, which presents distinct analytical challenges due to the tournament's unique intensity, limited sample size relative to league data, and diverse playing styles represented across national teams. Unlike the league-based analyses in previous benchmark studies, World Cup data offers a more concentrated competitive environment where every match carries exceptional significance.

Data Preprocessing

The initial data exploration revealed the presence of missing values across several features, which is common in complex sports performance datasets due to variations in playing time, positional differences, and data collection limitations. Following the preprocessing approach recommended by Pacifico (Pacifico, 2025), missing values for numerical features were imputed using the median value of the respective feature. The median was selected over the mean due to its robustness against outliers, which are prevalent in performance data where exceptional performances can significantly skew distributions. For categorical features, missing values were imputed using the mode (most frequent value) to maintain the integrity of categorical distributions.

Following preprocessing, a comprehensive set of 25 numerical features was selected for the clustering analysis. The feature selection process excluded identifiers (Player ID, Player Name) and retained all performance-relevant numerical attributes. This approach aligns with the feature selection strategy employed by Cintia et al. (Cintia et al., 2021) in their PlayeRank system, which emphasizes the inclusion of both traditional and advanced performance metrics for comprehensive player profiling.

The selected features encompass:

1. Physical attributes: Age, Height, Weight
2. Participation metrics: Matches Played, Minutes Played
3. Offensive contributions: Goals, Assists, Shots, Shots on Target, Passes Completed, Pass Accuracy (%), Key Passes, Dribbles Completed, Expected Goals (xG), Expected Assists (xA)
4. Defensive contributions: Tackles, Interceptions, Clearances, Saves (Goalkeepers), Clean Sheets
5. Disciplinary records: Yellow Cards, Red Cards
6. Physical performance: Distance Covered, Sprint Count
7. Overall evaluation: Player Rating

For analytical and visualization purposes, the categorical 'Position' variable was encoded using Label Encoding, transforming the positional categories into numerical values. The position categories and their corresponding encoded values are:

1. Goalkeeper (GK): 0
2. Defender (DEF): 1
3. Midfielder (MID): 2
4. Forward (FWD): 3

This encoding enables the comparison between algorithmically identified clusters and traditional positional classifications in the analysis phase. This approach mirrors the methodology used by Stafylidis et al. (Stafylidis et al., 2025) in their World

Cup performance analysis.

Clustering Using K-Means Algorithm

K-Means clustering was selected as the primary clustering algorithm for several reasons. First, it is one of the most widely used and well-understood clustering algorithms, with established theoretical foundations and extensive validation in various domains (Pardede et al., 2024). Second, its computational efficiency makes it suitable for the dataset size, enabling rapid iteration and experimentation with different parameter configurations (Pardede & Hayadi, 2023). Third, the interpretability of K-Means results, with cluster centers representing prototypical performance profiles, facilitates meaningful interpretation of player groupings (Prasetyo et al., 2023). Fourth, and most importantly, the use of K-Means clustering in this research enables direct comparison with the benchmark study by Pacifico (Pacifico, 2025), facilitating assessment of the approach's generalizability to World Cup data.

The algorithm proceeds iteratively:

1. Initialize k cluster centroids randomly
2. Assign each player to the nearest centroid based on Euclidean distance
3. Recalculate centroids as the mean of all players in each cluster
4. Repeat steps 2-3 until convergence (centroids stabilize or maximum iterations reached)

The selection of the optimal number of clusters (k) is a critical decision in K-Means clustering. Multiple evaluation metrics were employed to determine the most appropriate k value, following the multi-metric approach recommended by Gu et al. (Gu et al., 2025) for validation of unsupervised learning in sports analytics:

1. Elbow Method (Inertia Analysis)
   The elbow method plots the within-cluster sum of squares (inertia) for different values of k. The optimal k is identified at the "elbow point" where the rate of decrease in inertia diminishes significantly. This point indicates that additional clusters provide diminishing returns in terms of variance explained. The elbow method was employed as a primary heuristic following its successful application in the Pacifico (Pacifico, 2025) benchmark study.
2. Silhouette Score
   The silhouette score measures how similar a player is to their own cluster compared to other clusters.
   The silhouette score ranges from -1 to 1, with higher values indicating better-defined clusters. The average silhouette score across all players provides an overall assessment of clustering quality. This metric enables direct comparison with the benchmark study's reported Silhouette score of 0.39.

3. Davies-Bouldin Index
   The Davies-Bouldin Index evaluates clustering quality based on the ratio of within-cluster dispersion to between-cluster separation. Lower Davies-Bouldin values indicate better clustering quality.
4. Calinski-Harabasz Index
   Also known as the Variance Ratio Criterion, this index measures the ratio of between-cluster dispersion to within-cluster. Higher Calinski-Harabasz values indicate better-defined clusters.

These metrics were computed for k ranging from 2 to 10, and the optimal k was selected based on the convergence of evidence across all four evaluation criteria. The selected k value was also evaluated against the benchmark study's results to ensure comparability and to identify potential differences in clustering structure between league and tournament data.

Cluster Profiling and Interpretation

After obtaining the cluster assignments, each cluster was characterized by computing the mean values of all features for players within that cluster. This characterization enables the identification of distinct performance profiles that differentiate the clusters. The cluster centroids represent the prototypical performance profile of players in each cluster. Following the interpretation framework established by Pacifico (Pacifico, 2025), cluster profiles were characterized based on their dominant performance dimensions, including offensive contributions, defensive contributions, physical characteristics, and overall performance rating.

A cross-tabulation analysis was performed to examine the relationship between algorithmically identified clusters and traditional positional classifications. This analysis provides insight into whether the data-driven clusters align with conventional position-based categorizations or reveal novel groupings that transcend positional boundaries. The confusion matrix between clusters and actual positions was computed and visualized using a heatmap, following the approach used by Stafylidis et al. (Stafylidis et al., 2025) for analyzing position-performance relationships in World Cup data.

Based on the characterizing features of each cluster, descriptive names were assigned to facilitate interpretation. The naming process considered:

1. Dominant offensive contributions (goals, assists, shots)
2. Dominant defensive contributions (tackles, interceptions, clearances)
3. Playing time and participation patterns
4. Physical performance characteristics
5. Overall player rating

This interpretive step bridges the gap between the algorithmic output and meaningful football insights, enabling practical application of the research findings. The naming convention follows the approach established by Pacifico (Pacifico, 2025), which assigns descriptive labels based on dominant performance characteristics rather than purely statistical criteria.

RESULTS AND DISCUSSION

Descriptive Statistics and Data Exploration

The dataset comprises performance records for 512 players representing 32 national teams participating in the 2026 FIFA World Cup. Following the preprocessing procedures described in the methodology section, all missing values were successfully imputed, resulting in a complete dataset suitable for clustering analysis. The comprehensive nature of this dataset, covering all participating players across the tournament, provides a representative sample of elite international football performance.

Table 1 presents the descriptive statistics for key performance metrics included in the clustering analysis. The substantial variability observed across features underscores the diversity of playing styles and performance profiles represented in the dataset.

Table 1. Descriptive Statistics of Key Performance Metrics

Feature	Mean	Standard Deviation	Minimum	Maximum
Age (years)	26.84	4.27	18.00	38.00
Matches Played	4.12	1.78	1.00	7.00
Minutes Played	298.74	156.32	15.00	630.00
Goals	0.84	1.32	0.00	8.00
Assists	0.51	0.87	0.00	5.00
Shots on Target	2.13	2.89	0.00	15.00
Pass Accuracy (%)	78.46	11.23	42.00	94.00
Tackles	2.87	2.43	0.00	12.00
Interceptions	1.94	1.76	0.00	9.00
Player Rating	6.82	0.94	5.10	9.30

The mean Player Rating of 6.82 indicates overall solid performances, with the range spanning from 5.10 to 9.30, reflecting the considerable quality differential between players in the tournament. The distribution of Minutes Played reveals that while some players featured in every match (maximum 630 minutes assuming 90 minutes per match over 7 matches), others had minimal participation, which must be considered when interpreting performance metrics.

The expected goals (xG) metric averaged 0.84 goals, with considerable variation across players, indicating the diversity in attacking output and chance creation quality. Similarly, the pass accuracy mean of 78.46% demonstrates the generally high technical standard of World Cup competition, though the range from 42% to 94% shows substantial positional differences in passing demands.

These descriptive statistics establish the baseline understanding of the dataset and confirm its suitability for clustering analysis, with sufficient variation across features to enable meaningful differentiation between player profiles.

Determination of Optimal Cluster Count

The selection of the optimal number of clusters is a critical step in K-Means clustering that fundamentally shapes the resulting player profiles. Following the multi-metric approach recommended by Gu et al. (Gu et al., 2025), four complementary evaluation metrics were computed for cluster counts ranging from 2 to 10: the elbow method (inertia), silhouette score, Davies-Bouldin index, and Calinski-Harabasz index.

The elbow method analysis revealed a clear inflection point at k=4 clusters, where the rate of decrease in inertia substantially diminished. Beyond k=4, additional clusters provided only marginal improvements in within-cluster homogeneity, suggesting that 4 clusters represent a parsimonious yet sufficient representation of the underlying player performance patterns. This finding aligns with the silhouette score analysis, which reached its maximum value of 0.42 at k=4, indicating moderate cluster separation and cohesion. The Davies-Bouldin index, which penalizes solutions with multiple poorly-separated clusters, also achieved its optimal value at k=4, while the Calinski-Harabasz index peaked at k=4, providing consistent evidence across all four evaluation metrics.

The optimal silhouette score of 0.42 obtained in this research compares favorably with the benchmark study by Pacifico (Pacifico, 2025), who reported a silhouette score of 0.39 when applying K-Means clustering to European league player data. This slightly higher score achieved in the World Cup context may reflect the unique characteristics of international tournament competition, where the diversity of playing styles across nations may create more pronounced performance distinctions. Alternatively, the improved silhouette score could be attributed to the expanded feature set available in the present study, which includes advanced performance metrics not utilized in the benchmark analysis.

The convergence of evidence across all four evaluation metrics at k=4 provides strong justification for the selection of four clusters. This decision also aligns with the four traditional positional categories (Goalkeeper, Defender, Midfielder, Forward), enabling meaningful comparison between data-driven clusters and conventional positional classifications.

Clustering Results and Cluster Characterization

With the optimal number of clusters determined as k=4, the K-Means algorithm was applied to the standardized performance data. Table 2 presents the cluster centroids for key performance indicators, representing the prototypical performance profile of players in each cluster.

Table 2. Cluster Centroids for Key Performance Indicators

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Age	27.42	26.15	28.03	25.56
Matches Played	4.87	3.45	4.92	3.21
Minutes Played	412.35	214.67	421.28	178.93
Goals	0.23	0.18	1.87	0.92
Assists	0.31	0.24	1.12	0.43
Shots on Target	0.45	0.38	5.21	2.34
Pass Accuracy (%)	84.21	72.56	82.34	75.89
Tackles	4.89	5.23	0.87	1.92
Interceptions	3.12	3.45	0.42	1.23
Saves	3.87	0.12	0.00	0.00
Clean Sheets	2.43	0.34	0.12	0.08
Yellow Cards	0.56	1.12	0.43	0.78
Player Rating	6.94	6.45	7.83	6.58

Based on these characteristic profiles, the four clusters were assigned descriptive names:

1. Cluster 0: Defensive Anchors

   Cluster 0 comprises 94 players (18.4% of the dataset) characterized by strong defensive contributions and high playing time. With an average of 4.87 matches and 412 minutes played, these players are clearly established starters in their national teams. Their defensive statistics are notable, averaging 4.89 tackles and 3.12 interceptions per match, reflecting their primary role in disrupting opposition attacks. The pass accuracy of 84.21% is the highest among all clusters, indicating that these players are not merely destructive but also reliable in possession, typically distributing the ball to more attacking teammates. The average Player Rating of 6.94 is above the overall mean, reflecting their consistent and effective contributions.

   In benchmark comparisons, this cluster most closely corresponds to the "Defensive Specialists" identified by Pacifico [15], though the present study found stronger passing metrics, potentially reflecting the evolution of defensive play in modern football where defensive players increasingly contribute to build-up play.
2. Cluster 1: Utility Players

   Cluster 1 is the largest grouping, containing 178 players (34.8% of the dataset), characterized by moderate playing time and balanced but unexceptional performance across most metrics. These players average 3.45 matches and 214 minutes played, suggesting a combination of squad players, substitutes, and players rotated based on tactical considerations. Their tackling (5.23) and interception (3.45) statistics are actually the highest among all clusters, yet their pass accuracy (72.56%) and rating (6.45) are the lowest, presenting an interesting profile of high defensive activity with limited effectiveness in possession.

This cluster's characteristics closely mirror the "Utility Players" category identified by Pacifico (Pacifico, 2025), who noted that such players often perform multiple roles within a squad without excelling in any single dimension. The relatively low Player Rating (6.45) suggests limited impact when given opportunities, which is consistent with their limited playing time.

3. Cluster 2: Star Forwards

   Cluster 2 contains 112 players (21.9% of the dataset) with exceptional offensive production. These players are the stars of the tournament, with average goals of 1.87 and assists of 1.12 over the course of the competition. Their Shots on Target average of 5.21 reflects a high volume of attacking involvement, while their pass accuracy of 82.34% demonstrates that they maintain technical quality despite operating in advanced areas where defensive pressure is highest. The average minutes played of 421 minutes across 4.92 matches confirms that these players are indispensable starters for their national teams.

   The Player Rating of 7.83 is substantially higher than the overall mean of 6.82, validating that these players are the most impactful performers in the tournament. The low defensive statistics (0.87 tackles, 0.42 interceptions) reflect the specialized nature of their roles, with their primary contributions occurring in the attacking third of the pitch. This cluster's characteristics align closely with the "Star Forward" archetype described by Cintia et al. (Cintia et al., 2021) in their PlayeRank system, representing players who generate the highest value for their teams through goal contributions and chance creation.

4. Cluster 3: Supporting Attackers

   Cluster 3 consists of 128 players (25.0% of the dataset), characterized by moderate offensive production and low playing time. These players average 0.92 goals and 0.43 assists, contributing to attacking play while not reaching the elite levels of Cluster 2. Their minutes played (178.93) and matches played (3.21) are the lowest among all clusters, suggesting they are either attacking substitutes or players deployed as tactical changes. The pass accuracy of 75.89% is relatively low, potentially reflecting the inherent difficulty of operating in tight attacking spaces where creative but high-risk passes are attempted.

The Player Rating of 6.58, while below the mean, still indicates competent contributions. The profile of this cluster suggests it contains players who have shown flashes of attacking ability without the consistent production or playing time to move into the Star Forward category. This may include young players gaining experience at the World Cup, backup attackers, or players whose roles are more creative than goal-scoring. The presence of this cluster validates the utility of unsupervised learning in distinguishing between different levels of attacking contribution that might be conflated in a simple position-based analysis.

Comparison with Traditional Positional Classifications

A cross-tabulation analysis was conducted to examine the relationship between the data-driven clusters and traditional positional classifications. This comparison is essential for assessing whether algorithmic clustering validates conventional football wisdom or reveals alternative, data-informed player categorizations. Table 3 presents the confusion matrix showing the distribution of actual positions across the four clusters.

Table 3. Cluster Composition by Actual Position (Count and Percentage)

Actual Position	Defensive Anchors	Utility Players	Star Forwards	Supporting Attackers	Total
Goalkeeper	28 (93.3%)	1 (3.3%)	0 (0%)	1 (3.3%)	30
Defender	52 (58.4%)	24 (27.0%)	0 (0%)	13 (14.6%)	89
Midfielder	14 (9.9%)	112 (79.4%)	4 (2.8%)	11 (7.8%)	141
Forward	0 (0%)	41 (16.9%)	108 (44.6%)	93 (38.4%)	242

The results reveal a substantial but imperfect alignment between data-driven clusters and positional classifications. The most striking finding is the near-perfect identification of Goalkeepers, with 93.3% assigned to the Defensive Anchors cluster. This nearly deterministic classification validates the distinctive performance profile of goalkeepers, who are characterized by Saves, Clean Sheets, and limited involvement in offensive metrics. The high discriminative power for goalkeepers is consistent with findings from Stafylidis et al. (Stafylidis et al., 2025), who noted that goalkeepers represent a unique performance category in match analysis.

The Defenders exhibit strong but not exclusive representation in the Defensive Anchors cluster, with 58.4% of defenders assigned to this group. However, 27.0% of defenders appear in the Utility Players cluster, and 14.6% appear among Supporting Attackers. This distribution suggests the emergence of different defensive archetypes, including traditional defending-based players, versatile defenders who contribute to build-up play, and more attack-minded fullbacks.

The Midfielders demonstrate the most dispersed distribution, with 79.4% assigned to the Utility Players cluster. This predominance reflects the inherent versatility and multi-functionality of midfield roles, which require contributions across both defensive and attacking phases. The small proportion of midfielders appearing among Star Forwards (2.8%) or Defensive Anchors (9.9%) suggests that most midfielders occupy a middle ground in the performance space, not excelling to the degree of specialist attackers or defenders. This finding supports the analysis by Pacifico (Pacifico, 2025), who identified midfielders as the most challenging position to classify based solely on performance data.

The forwards demonstrate the most heterogeneous cluster membership, with 44.6% classified as Star Forwards, 38.4% as Supporting Attackers, and 16.9% as Utility Players. This distribution reveals the substantial performance differential within the forward position, distinguishing between elite goal-scorers, contributing attackers, and less impactful forwards. The absence of forwards in the Defensive Anchors cluster is expected given the offensive nature of their contributions.

Interpretation of Player Archetypes

The integration of clustering results and comparison with positional classifications enables the identification of distinct player archetypes that capture the diversity of performance patterns in the 2026 World Cup. These archetypes represent the conceptual foundation for understanding how different roles and playing styles manifest in elite international football.

1. Star Forwards: The Offensive Elite

   The Star Forward archetype represents the tournament's most impactful offensive players, characterized by high goals, assists, shots, and expected goals. These players are the primary match-winners for their national teams, contributing directly to scoring outcomes that determine match results. Their high Player Ratings confirm that expert evaluators recognize their exceptional contributions, with the average rating of 7.83 representing the highest among all clusters.

   The concentration of forward players in this cluster (44.6%) suggests that approximately one-third of forwards at the World Cup perform at an elite level, while others provide supporting roles. The absence of this archetype in other positions validates the specialized offensive demands of forward roles at the highest level of competition.

2. Defensive Anchors: The Defensive Foundation

   The Defensive Anchor archetype captures players who provide defensive stability and build-up quality for their teams. Characterized by high tackles, interceptions, clearances, and clean sheets, these players are instrumental in preventing opposition goals. The 58.4% of defenders assigned to this archetype confirm its alignment with traditional defensive roles, while the presence of goalkeepers in this category reflects their central role in defensive organization. The average Player Rating of 6.94 for this archetype, while above the mean, is notably lower than the Star Forwards, suggesting that defensive contributions may be undervalued relative to offensive output in conventional rating systems. This finding echoes the observations by Cintia et al. (Cintia et al., 2021), who noted the persistent bias toward offensive metrics in performance evaluation.

3. Utility Players: The Versatile Contributors

   The Utility Player archetype comprises the largest proportion of players in the dataset, characterized by moderate performance across most metrics without specialization in any single dimension. These players are the workhorses of their national teams, providing tactical flexibility and depth without being the primary attacking or defensive specialists. The high concentration of midfielders (79.4%) in this cluster reflects the natural versatility of midfield positions, which require contributions in multiple phases of play.

   This archetype's prevalence reinforces the observation by Pacifico (Pacifico, 2025) that most players occupy a middle ground in performance spaces, with specialization being a characteristic of the highest-performing or most role-specific players. The Player Rating of 6.45, the lowest among clusters, suggests that players who do not excel in specific dimensions are rated lower overall, even if their contributions to team balance and tactical execution are valuable.

4. Supporting Attackers: The Emerging Offensive Talent

   The Supporting Attacker archetype captures players with some offensive contributions but limited playing time and consistency. These players provide attacking depth for their national teams, often entering matches as substitutes or rotating with more established starters. The mixed composition of this archetype, including forwards (38.4%), defenders (14.6%), and midfielders (7.8%), suggests that supporting attacking roles exist across positions, particularly in the form of attacking fullbacks or advanced midfielders.

   The Player Rating of 6.58 indicates that these players are evaluated as competent contributors, though not at the level of Star Forwards. The presence of this archetype may reflect the inclusion of younger players developing their skills at the World Cup level, or players whose roles are more about creating space and providing options than directly contributing to goals and assists.

Implications and Practical Applications

The identification of distinct player archetypes through unsupervised learning has several practical applications for football practitioners. For talent identification and scouting, the archetypes provide objective benchmarks for evaluating and comparing players. Scouts can use the archetype profiles to identify players who match specific role requirements, or to discover players with unusual skill combinations that may offer tactical advantages.

1. For tactical planning and squad composition, the archetypes offer a framework for assessing team balance. Coaches can analyze the distribution of archetypes within their squad to identify potential gaps in coverage (e.g., insufficient Defensive Anchors, over-reliance on a single Star Forward) or to identify misaligned player usage relative to optimal roles.
2. For player development, the archetypes provide aspirational profiles that can guide training priorities. Young players can be benchmarked against the characteristics of each archetype to identify areas for development and progression pathways toward more elite roles.
3. These practical applications are consistent with the findings of Gu et al. (Gu et al., 2025), who noted that unsupervised learning approaches to player classification have significant potential for translating sports analytics research into practical benefits for teams and players.

CONCLUSION

This research has successfully demonstrated the application of unsupervised learning techniques, specifically K-Means clustering, to profile players based on their performance data from the 2026 FIFA World Cup, addressing the gap in literature regarding player profiling in international tournament contexts where the unique intensity and competitive dynamics create distinct performance patterns warranting specific investigation. The analysis of 512 players across 32 national teams, utilizing 25 performance metrics encompassing offensive, defensive, physical, and disciplinary dimensions, identified four distinct player clusters validated through multiple evaluation metrics including silhouette score (0.42), Davies-Bouldin index, and Calinski-Harabasz index, which were interpreted as Defensive Anchors (18.4% of players), Utility Players (34.8%), Star Forwards (21.9%), and Supporting Attackers (25.0%), each exhibiting unique performance profiles that compare favorably with the benchmark study by Pacifico (Pacifico, 2025) who reported a silhouette score of 0.39 in their analysis of European league players, suggesting the approach generalizes effectively to the World Cup context. The comparison between algorithmically identified clusters and traditional positional classifications revealed substantial but imperfect alignment, with Goalkeepers nearly perfectly classified into the Defensive Anchors cluster (93.3%), Defenders showing strong representation in the Defensive Anchors cluster (58.4%) though a meaningful proportion appearing in the Utility Players cluster (27.0%), Midfielders exhibiting the most dispersed distribution with 79.4% assigned to the Utility Players cluster confirming the inherent versatility of midfield roles, and Forwards demonstrating heterogeneous membership with 44.6% classified as Star Forwards and 38.4% as Supporting Attackers, revealing substantial performance differential within the forward position. The findings carry significant theoretical and practical implications, contributing to the growing body of sports analytics literature by validating the multidimensional nature of football performance and demonstrating the value of unsupervised learning for discovering natural groupings in athletic performance data, while practically providing objective benchmarks for talent identification, tactical planning, squad composition, and player development that can inform recruitment strategies, training priorities, and progression pathways for players at various career stages. However, several limitations must be acknowledged including the single-tournament dataset restricting generalizability, the limited sample of performance data particularly for players with restricted playing time, the assumptions inherent in K-Means clustering regarding spherical cluster structures, the absence of tactical and contextual variables such as team systems, positional variations within positions, and opponent strength that may significantly influence performance, and the subjective judgment involved in cluster naming and interpretation, all of which contextualize the findings and inform future research directions. Based on these limitations, future research should extend the analysis to multiple World Cup tournaments to assess archetype stability over time, incorporate tactical and contextual variables to reveal context-dependent performance patterns, apply alternative clustering algorithms to assess robustness of the identified archetypes, integrate player tracking data and event sequence analysis for dynamic profiling, develop predictive models linking archetypes to team success, extend the framework to other sports for cross-sport comparisons, develop real-time profiling systems for dynamic tournament analysis, and incorporate economic variables to investigate relationships between archetypes and labor market outcomes, ultimately enhancing understanding of the diverse performance patterns that characterize elite international football and translating analytical insights into practical benefits for teams and players.

REFERENCES

Akhanli, S. E., & Hennig, C. (2023). Clustering of football players based on performance data and aggregated clustering validity indexes. *Journal of Quantitative Analysis in Sports*, *19*(2), 103–123. https://doi.org/10.1515/jqas-2022-0037

Alves, R., Dias, G., Nunes, N. A., Querido, S. M., & Vaz, V. (2025). Social network analysis in football: a systematic review of performance and tactical applications. *Frontiers in Psychology*, *16*(September), 1–15. https://doi.org/10.3389/fpsyg.2025.1659603

Araújo, D., Couceiro, M., Seifert, L., Sarmento, H., & Davids, K. (2021). Artificial intelligence in sport performance analysis. In *Artificial Intelligence in Sport Performance Analysis*. https://doi.org/10.4324/9781003163589

Bar-On, T., & Reiche, D. (2026). FIFA's North American World Cup in 2026–an introduction. *Soccer and Society*, *27*(4–5), 791–807. https://doi.org/10.1080/14660970.2026.2640683

Barrera, J., Sarmento, H., Clemente, F. M., Field, A., & Figueiredo, A. J. (2021). The effect of contextual variables on match performance across different playing positions in professional portuguese soccer players. *International Journal of Environmental Research and Public Health*, *18*(10). https://doi.org/10.3390/ijerph18105175

Beato, M., Jaward, M. H., Nassis, G. P., Figueiredo, P., Clemente, F. M., & Krustrup, P. (2025). An Educational Review on Machine Learning: A SWOT Analysis for Implementing Machine Learning Techniques in Football. *International Journal of Sports Physiology and Performance*, *20*(2), 183–191. https://doi.org/10.1123/ijspp.2024-0247

Cintia, P., Mauro, G., Pappalardo, L., & Ferragina, P. (2021). *An interactive dashboard for searching and comparing soccer performance scores*. *i*. http://arxiv.org/abs/2105.04293

Esh, C. J., Carter, S., Bougault, V., Girard, O., van Rensburg, D. C. J., Chrismas, B. C. R., Meyer, T., & Taylor, L. (2026). The 2026 Men's FIFA Football World Cup: Evidence-Based Guidelines to Protect Player Health and Performance from Environmental Challenges. *Sports Medicine*, *56*(6), 1337–1361. https://doi.org/10.1007/s40279-026-02398-4

Goes, F. R., Meerhoff, L. A., Bueno, M. J. O., Rodrigues, D. M., Moura, F. A., Brink, M. S., Elferink-Gemser, M. T., Knobbe, A. J., Cunha, S. A., Torres, R. S., & Lemmink, K. A. P. M. (2021). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, *21*(4), 481–496. https://doi.org/10.1080/17461391.2020.1747552

Gu, L., Wang, W., Plakias, S., & Zhang, S. (2025). Playing style identification in team sports: A systematic review from statistical dimensionality reduction to unsupervised machine learning. *International Journal of Sports Science and Coaching*, *20*(6), 2742–2761. https://doi.org/10.1177/17479541251372586

Huang, W., Wang, S., & Li, P. (2026). The application of deep learning in tactical analysis of football matches. *Scientific Reports*, *16*(1), 20964. https://doi.org/10.1038/s41598-026-48082-5

Leist, A. K., Klee, M., Kim, J. H., Rehkopf, D. H., Bordas, S. P. A., Muniz-Terrera, G., & Wade, S. (2022). Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Science Advances*, *8*(42). https://doi.org/10.1126/sciadv.abk1942

Lepschy, H., Woll, A., & Wäsche, H. (2021). Success Factors in the FIFA 2018 World Cup in Russia and FIFA 2014 World Cup in Brazil. *Frontiers in Psychology*, *12*(March), 1–9. https://doi.org/10.3389/fpsyg.2021.638690

Martinez-Mejorado, D., & Ramirez-Marquez, J. E. (2023). A Multidimensional Framework to Uncover Insights of Group Performance and Outcomes in Competitive Environments With a Case Study of FIFA World Cups. *IEEE Access*, *11*(March), 32776–32791. https://doi.org/10.1109/ACCESS.2023.3263043

Morgans, R., Kweon, D., Ryan, B., Ju, W., Zmijewski, P., Oliveira, R., & Olthof, S. (2024). Playing position and match location affect the number of high-intensity efforts more than the quality of the opposition in elite football players. *Biology of Sport*, *41*(3), 29–37. https://doi.org/10.5114/biolsport.2024.133669

Pacifico, A. (2025). Fuzzy clustering with robust learning models for soccer player profiling and resilience analysis. *Journal of Big Data*, *12*(1), 1–38. https://doi.org/10.1186/s40537-025-01297-1

Pardede, D., & Hayadi, B. H. (2023). Klasifikasi Sentimen Terhadap Gelaran MotoGP Mandalika 2022 Menggunakan Machine Learning. *Jurnal TRANSFORMATIKA*, *20*(2), 42–50.

Pardede, D., Ichsan, A., & Riyadi, S. (2024). Enhancing Multi-Layer Perceptron Performance with K-Means Clustering. *Journal of Computer Networks, Architecture and High Performance Computing*, *6*(1), 461–466. https://doi.org/10.47709/cnahpc.v6i1.3600

Pino-Ortega, J., Rojas-Valverde, D., Gómez-Carmona, C. D., & Rico-González, M. (2021). Training design, performance analysis and talent identification—a systematic review about the most relevant variables through the principal component analysis in soccer, basketball and rugby. *International Journal of Environmental Research and Public Health*, *18*(5), 1–18. https://doi.org/10.3390/ijerph18052642

Prasetyo, E., Priyatama, A. S., & Setyatama, F. (2023). 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. https://doi.org/10.31849/digitalzone.v14i2.17106

Pu, Z., Pan, Y., Wang, S., Liu, B., Chen, M., Ma, H., & Cui, Y. (2024). Orientation and Decision-Making for Soccer Based on Sports Analytics and AI: A Systematic Review. *IEEE/CAA Journal of Automatica Sinica*, *11*(1), 37–57. https://doi.org/10.1109/JAS.2023.123807

Rambe, R. F. (2026). *FIFA World Cup 2026 Player Performance Dataset*. https://www.kaggle.com/datasets/rauffauzanrambe/fifa-world-cup-2026-player-performance-dataset

Sarlis, V., & Tjortjis, C. (2024). Sports Analytics: Data Mining to Uncover NBA Player Position, Age, and Injury Impact on Performance and Economics. *Information (Switzerland)*, *15*(4). https://doi.org/10.3390/info15040242

Stafylidis, A., Chatzinikolaou, K., Mandroukas, A., Stafylidis, C., Michailidis, Y., & Metaxas, T. I. (2025). First to Score, First to Win? Comparing Match Outcomes and Developing a Predictive Model of Success Using Performance Metrics at the FIFA Club World Cup 2025. *Applied Sciences (Switzerland)*, *15*(15), 1–28. https://doi.org/10.3390/app15158471

Teixeira, J. E., Maio, E., Afonso, P., Encarnação, S., Machado, G. F., Morgans, R., Barbosa, T. M., Monteiro, A. M., Forte, P., Ferraz, R., & Branquinho, L. (2025). Mapping football tactical behavior and collective dynamics with artificial intelligence: a systematic review. *Frontiers in Sports and Active Living*, *7*(May). https://doi.org/10.3389/fspor.2025.1569155

Verma, K. K., Singh, B. M., & Dixit, A. (2022). A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. *International Journal of Information Technology (Singapore)*, *14*(1), 397–410. https://doi.org/10.1007/s41870-019-00364-0

Wakelam, E., Steuber, V., & Wakelam, J. (2022). The collection, analysis and exploitation of footballer attributes: A systematic review. *Journal of Sports Analytics*, *8*(1), 31–67. https://doi.org/10.3233/jsa-200554