Jurnal Masyarakat Informatika, 17(2), 2026
DOI:10.14710/jmasif.17.2.86342
ISSN: 2777-0648

Team Quality and Tactical Variation in English Premier League Clubs: A Comparative Unsupervised Clustering Analysis

Fahri Syaifullah[1)], Amiq Fahmi[2*)]

[1)]*Department of Informatics Engineering, Faculty of Computer Science,Universitas Dian Nuswantoro, Semarang, Indonesia*

[2)]*Department of Information Systems, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia*

[*] *Corresponding author: amiq.fahmi@dsn.dinus.ac.id*

Abstract

Clustering-based approaches to football analytics often rely on proprietary event data, limiting accessibility for independent researchers. This study introduces an unsupervised comparative framework to characterize English Premier League (EPL) team playing styles in the 2025/2026 season using publicly available FBref and Understat statistics. Four clustering algorithms (K-Means, Hierarchical Agglomerative Clustering, Gaussian Mixture Models, and DBSCAN) were applied to a multidimensional feature set spanning offensive, defensive, pressing, and build-up dimensions, with performance evaluated through established validity indices. K-Means produced four exploratory profiles, ordered largely by team quality: Possession Dominant and High Press, Pressing and Transition, Low Block and Direct Play, and Low Block and Vulnerable Defense. Subsampling stability was moderate (median ARI = 0.69 at k = 4), and a feature-group ablation showed that the dominant structure follows team quality, with secondary variation in pressing and build-up. This exploratory study (n = 20, a single snapshot) does not compare its results with proprietary event data and therefore makes no claim of equivalence; it illustrates what can be recovered from public aggregate statistics for team profiling.

Keywords: *unsupervised learning, clustering algorithms, playing styles, English Premier League, sports analytics*

1 Introduction

Football is the most popular sport worldwide, with more than four billion fans, and has evolved into a multi-billion-dollar industry, particularly through prestigious competitions such as the English Premier League (EPL). The complexity of the game, which involves dynamic interactions among players, team strategies, and match conditions, makes performance analysis both a challenge and an opportunity. From an academic perspective, research in this domain contributes to the advancement of data-driven methodologies in sports science, particularly the application of machine learning in football analytics [1-3]. From a practical standpoint, a deeper understanding of team playing styles can assist coaches in designing match strategies, support analysts in preparing scouting reports, and guide club management in strategic decision-making such as player recruitment [4].

The rapid development of tracking technologies and match data recording has resulted in an increasing volume of multidimensional statistics, including ball possession, passing accuracy, shooting efficiency, and game intensity. Leveraging these data through machine learning enables the identification of latent playing patterns that are not easily observable through conventional analysis [5-7].

Received: 02 July 2026; Revised: 30 October 2026; Accepted: 25 November 2026; Published: 27 November 2026
Copyright (c) 2026 The authors. Published by Department of Informatics, Universitas Diponegoro
This is an open access article under the CC-BY-SA license.

Research in football analytics has progressed from predictive approaches to unsupervised exploration of game structures. Early studies predominantly employed supervised learning techniques such as Random Forest, XGBoost, and Neural Networks to predict match outcomes [2,5]. Machine learning has also been applied to player performance, injury forecasting, and player selection in professional football [7-9]. Beyond playing style analysis, machine learning has been widely used to evaluate player contributions through advanced metrics such as Expected Goals (xG), which provide granular assessments of scoring opportunities [6]. Stival et al. [10], for instance, proposed a deep learning pipeline based on graph representation and time-series analysis to predict whether ball possession sequences would progress into dangerous attacking zones, highlighting the importance of spatiotemporal approaches in modern football analytics. Similarly, machine learning has been applied to injury prediction based on workload and physical conditions, demonstrating the broader potential of sports data for both tactical and long-term performance management [3,7]. Beyond predictive modeling, research has progressively shifted toward unsupervised exploration of latent structures within football match data.

In the context of playing style exploration, several studies have shown that team structures can be represented through dimensionality reduction and clustering of match statistics. Methods such as Principal Component Analysis (PCA), K-Means clustering, Hierarchical clustering, Gaussian Mixture Models (GMM), DBSCAN, and Nonnegative Matrix Factorization (NMF) have been widely adopted to extract latent patterns in football data [11-14]. For example, Kong et al. [15] combined PCA and K-Means clustering to identify playing styles in the Chinese Super League, distinguishing between intensity-based and possession-based approaches. Martín-Castellanos et al. [16] and Ruan et al. [17] demonstrated the effectiveness of PCA with Varimax rotation in extracting playing structures across different leagues, while Zhou et al. [14] highlighted the seasonal evolution of playing styles. He et al. [18] further revealed that flexibility in playing styles positively correlates with team performance.

Clustering approaches have also been applied in other contexts. Moffatt et al. [19] used K-Means clustering to classify playing styles in the Australian Football League, whereas Romero Clavijo et al. [13] employed spatiotemporal clustering to analyze offensive and defensive patterns. Li et al. [20] demonstrated that factor analysis and clustering remain dominant approaches in identifying playing styles, despite methodological variations. These findings collectively suggest that unsupervised learning methods, particularly clustering-based approaches, remain the most widely adopted and effective techniques for playing style identification in football analytics.

In this study, playing style is defined as a multidimensional statistical representation of collective team characteristics in offensive actions, ball possession, passing distribution, and overall performance throughout a season. This representation enables the transformation of match data into a latent space that can be analyzed using unsupervised learning methods.

Despite the growing body of literature, several limitations remain. First, most studies have focused on specific leagues such as the CSL or regional competitions, while analyses of highly competitive leagues such as the EPL remain limited in clustering-based approaches. Second, existing studies often rely on single methods such as PCA or K-Means, without providing comprehensive comparisons across multiple clustering techniques. Third, many studies depend on event-based data requiring complex tracking infrastructure, whereas the use of publicly available aggregate data sources such as FBref and Understat has not been fully explored.

This research proposes an unsupervised machine learning framework based on multidimensional statistics to identify EPL team playing styles during the 2025–2026 season using publicly available data from FBref and Understat. The dataset includes shooting and general performance statistics from FBref, as well as advanced metrics from Understat such as Expected Goals (xG), Passes Allowed Per Defensive Action (PPDA), and Deep Completions (DC), which capture pressing intensity and build-up patterns. Four clustering methods such as K-Means, Hierarchical Agglomerative Clustering, Gaussian Mixture Models (GMM), and DBSCAN are compared, with PCA employed only for visualization. Model evaluation is conducted using Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index to ensure objective model selection. PPDA and DC are included as proxies for pressing intensity and build-up play. The contribution of this study is an application-oriented integration of these public metrics with FBref statistics, not a new metric or method. An ablation without PPDA, OPPDA and DC (Section 3.8) changes the k = 4 partition substantially (ARI = 0.37 with the full model), so these tactical features contribute to the four-cluster structure.

This study has three primary objectives: (1) to identify and analyze EPL team playing styles using unsupervised machine learning, (2) to compare multiple clustering methods within a unified analytical framework, and (3) to evaluate the relationship between playing styles and team performance indicators. This study makes four key contributions. First, it introduces a comparative framework for four unsupervised learning methods in football playing style analysis. Second, it provides empirical profiles of EPL teams based on multidimensional publicly available statistics. Third, it examines what playing-style structure can be recovered from publicly accessible aggregate data from FBref and Understat, and how much of that structure reflects team quality rather than tactics. Fourth, it illustrates a reproducible workflow that analysts can apply to public data.

2 Research Methods

2.1 Research Workflow

This study employed an unsupervised machine learning framework to identify and analyze playing style patterns among English Premier League teams during the 2025/2026 season. The overall research workflow is illustrated in Figure 1.

The procedure consisted of six steps. (1) Statistics for all 20 clubs as of matchday 32 were obtained from FBref and Understat, merged, and the 10 features were standardized (z-score). (2) K-Means was fitted for k = 2–8 and the Silhouette Score, Davies–Bouldin Index, Calinski–Harabasz Index and inertia were computed. (3) With k = 4, K-Means, Hierarchical Agglomerative Clustering (HAC, Ward), GMM and DBSCAN were fitted directly to the standardized 10-feature matrix, not to principal components. (4) The partitions were evaluated with the internal indices and compared using ARI and NMI. (5) Stability and sensitivity were assessed by subsampling and feature-group ablation. (6) K-Means clusters were profiled by their feature means. PCA (two components, 77.7% of the variance) was used only to visualize cluster separability.

2.2 Data Collection and Preprocessing

Data were collected from FBref and Understat. From FBref, the squad shooting table [21], and the squad and opponent standard-statistics tables of the 2025–2026 Premier League were used [22]; from Understat, the EPL league table was used [23]. The dataset covers all 20 clubs as of matchday

32, with data retrieved on 16 April 2026. Because several advanced tactical metrics were unavailable from FBref at the time of retrieval, additional variables were taken from Understat: Expected Goals (xG), Expected Goals Against (xGA), Passes Allowed Per Defensive Action (PPDA), Opponent PPDA, Deep Completions and Non-Penalty Expected Goals Difference.

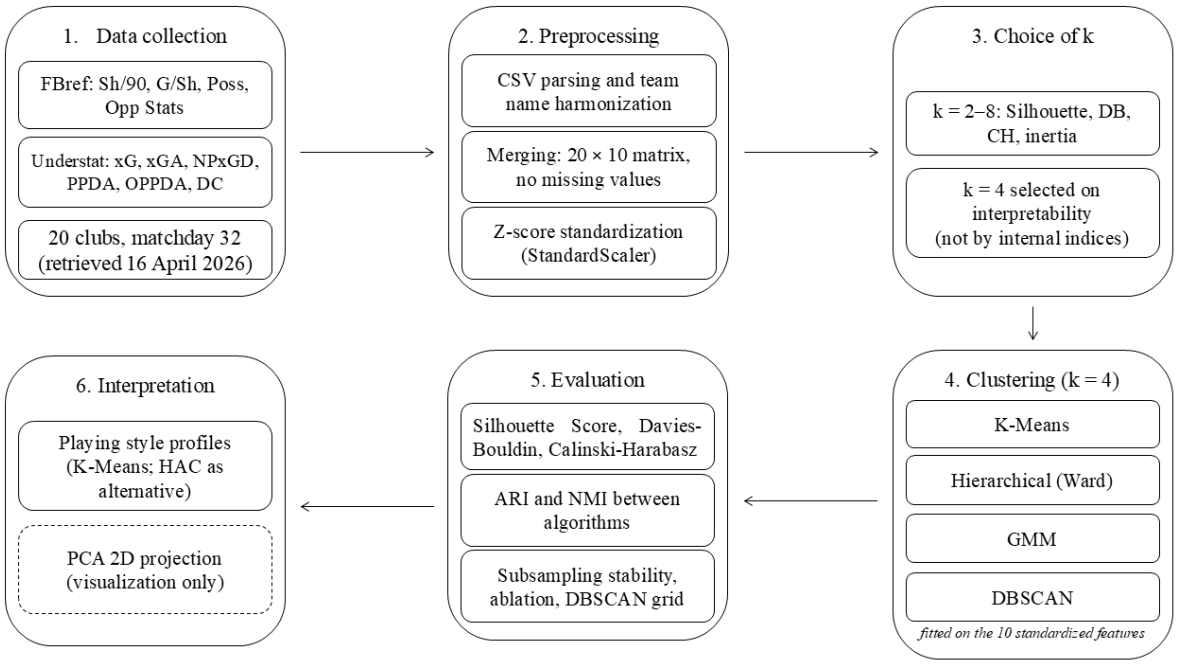

Figure 1 Research workflow for EPL playing style identification using unsupervised learning

After collection, the Understat files, which use semicolons as delimiters and therefore misalign columns when imported into Python, were converted to comma-separated format. Team names in Understat were harmonized to the FBref convention (e.g. “Manchester United” to “Manchester Utd” and “Tottenham” to “Tottenham Hotspur”) to prevent unmatched entries, and the four datasets were merged on the team name column. The resulting dataset contains 20 rows, one per club, with all FBref and Understat features and no missing values.

From the total available columns across the four datasets, 10 features were selected to represent four analytical dimensions: offensive performance, defensive performance, pressing intensity, and build-up play. The features were chosen for their relevance in football analytics literature [11,12,15] and the completeness of the data across all EPL clubs, and are presented in Table 1.

Table 1 Selected Features for Playing Style Identification

No.	Feature	Description	Source	Dimension
1	Poss	Average ball possession % per match	FBref	Possession
2	Sh/90	Total shots per 90 minutes	FBref	Offensive
3	G/Sh	Goals per shot (shooting efficiency)	FBref	Offensive
4	xG	Expected Goals	Understat	Offensive
5	xGA	Expected Goals Against	Understat	Defensive
6	NPxGD	Non-Penalty xG Difference (NPxG minus NPxGA)	Understat	Overall
7	PPDA	Passes Allowed Per Defensive Action	Understat	Pressing
8	OPPDA	Opponent PPDA (how freely opponent can build-up play)	Understat	Pressing
9	DC	Deep Completions	Understat	Build-up
10	Opp Gls p90	Goals conceded per 90 minutes	FBref	Defensive

Before clustering, the selected features were standardized because they exhibit heterogeneous scales and units. For example, PPDA values typically range between 9–14, whereas Deep Completions may exceed 300. Without normalization, variables with larger magnitudes would dominate distance-based clustering calculations. Therefore, all features were standardized using StandardScaler to produce a mean of 0 and a standard deviation of 1. The normalization formula is expressed in Equation (1) as follows:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where x denotes the original feature value, μ represents the feature mean, and σ is the standard deviation.

2.3 Proposed Method

This study proposes a comparative framework by implementing four unsupervised machine learning methods to identify the playing styles of EPL clubs. The selected methods represent different clustering approaches, including distance-based, hierarchical, probabilistic, and density-based clustering. All methods were applied to the same standardized feature matrix consisting of 10 normalized features. The number of clusters was set to k = 4 on interpretability grounds, although the Silhouette Score favors k = 2 (Section 3.2). The technical description of each method is presented as follows.

To ensure the reproducibility of the experiments and provide a transparent basis for comparison, the complete set of experimental parameters for all evaluated methods is summarized in Table 2.

Table 2 Experimental Parameters

Parameter	Value
K-Means: n_clusters	4
K-Means: n_init	10
K-Means: random_state	42
HAC: n_clusters	4
HAC: linkage	Ward
GMM: n_components	4
GMM: covariance_type	full (default)
GMM: random_state	42
DBSCAN: eps (ε)	2.0
DBSCAN: min_samples	2
PCA: n_components	2
PCA: random_state	42
Feature scaling	StandardScaler (μ=0, σ=1)
k search range	2 to 8
GMM: n_init; reg_covar	1; 1e-6 (default). Sensitivity run: 20; 0.01
HAC and DBSCAN: metric	Euclidean
DBSCAN sensitivity grid	ε ∈ {1.5, 2.0, 2.5, 3.0, 3.5, 4.0}; min_samples ∈ {2, 3}
Stability	300 subsamples of 80% (16 of 20), sampling seed 1
Software	Python 3.12.7, scikit-learn 1.5.2, NumPy 2.4.4, pandas 2.3.3, SciPy 1.17.1

2.3.1 K-Means

K-Means partitions the EPL clubs into $k = 4$ clusters based on similarity in the feature space. The algorithm iteratively assigns each team to the nearest centroid and updates centroid positions until convergence. To reduce sensitivity to random initialization, the model was executed with n_init $= 10$, and the solution with the lowest objective function value was selected, as defined in Equation (2).

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2 \quad (2)$$

where J denotes the total within-cluster sum of squared distances, C_i represents the i-th cluster, x is a data point, and μ_i denotes the centroid of cluster C_i.

2.3.2 Hierarchical Agglomerative Clustering

Hierarchical Agglomerative Clustering applies a bottom-up strategy in which each team initially forms an individual cluster. At each iteration, the two closest clusters are merged until the desired number of clusters is reached. This study employs Ward linkage, which minimizes the increase in within-cluster variance during the merging process, resulting in compact and well-separated clusters, as formulated in Equation (3).

$$d(A \cup B, C) = \frac{(n_A + n_C)d(A, C) + (n_B + n_C)d(B, C) - n_C \cdot d(A, B)}{n_A + n_B + n_C} \quad (3)$$

where n_A , n_B , and n_C denote the number of observations in clusters A, B, and C respectively.

2.3.3 Gaussian Mixture Model

Gaussian Mixture Model (GMM) is a probabilistic clustering method that estimates the probability distribution of each data point across multiple clusters. Unlike K-Means, which performs hard assignment, GMM allows a team to belong to multiple clusters with different membership probabilities. This characteristic is useful for identifying teams with transitional or mixed playing styles, as expressed in Equation (4).

$$p(x) = \sum_{k=1}^{K} \pi_k \, N(x \mid \mu_k, \Sigma_k) \quad (4)$$

where π_k denotes the mixing coefficient of component k, and $\mathcal{N}(x \mid \mu_k, \Sigma_k)$ represents the Gaussian distribution with mean μ_k and covariance matrix Σ_k.

A full-covariance GMM (scikit-learn default; n_init = 1, reg_covar = 10^{-6}) was fitted; sensitivity to tied, diagonal and spherical covariance, and model selection by BIC, are reported in Section 3.6.

2.3.4 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) identifies clusters based on local data density without requiring the number of clusters to be specified in advance. Data points located in low-density regions are classified as noise, enabling the detection of teams with statistically distinct playing styles. In this study, the parameters were set to $\varepsilon = 2.0$ and $min_samples = 2$ to accommodate the relatively small dataset size, as defined in Equation (5).

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

where $N_\varepsilon(p)$ denotes the ε-neighborhood of point p, D represents the dataset, and $dist(p,q)$ is the Euclidean distance between points p and q.

2.3.5 Principal Component Analysis

Principal Component Analysis (PCA) was employed exclusively for visualization. Since the dataset contains 10 features, direct visualization of cluster separability in the original feature space is impractical. PCA projects the data into a lower-dimensional space while preserving most of the variance. In this study, the first two principal components explain 77.7% of the total variance, which is considered sufficient for exploratory visualization, as quantified in Equation (6).

$$\text{Explained Variance Ratio} = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \quad (6)$$

where λ_i denotes the eigenvalue of the i-th principal component and $\sum \lambda_j$ represents the total variance in the dataset.

2.4 Testing & Evaluation Metrics

Since this study employs an unsupervised learning approach, no ground truth labels are available to serve as a reference for evaluation. Consequently, conventional supervised learning metrics such as accuracy, precision, recall, and confusion matrix are not applicable in this context. Instead, the clustering results were evaluated using three internal validation indices, namely the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, each of which assesses cluster quality from a distinct perspective.

Prior to clustering analysis, the optimal number of clusters was determined using the Elbow Method, Silhouette Score, and Davies-Bouldin Index across $k = 2$ to $k = 8$. Although $k = 2$ produced the highest statistical performance, $k = 4$ was selected to provide more tactically interpretable playing style categories based on domain knowledge in football analytics.

After determining $k = 4$, all clustering methods, namely K-Means, Hierarchical Agglomerative Clustering, Gaussian Mixture Model (GMM), and DBSCAN, were evaluated using the same validation metrics. K-Means was used for playing style interpretation (Section 3.5).

Cluster stability was assessed by drawing 300 random subsamples of 80% of the clubs (16 of 20; seed 1), clustering each with K-Means (n_init = 10), and computing the ARI between the subsample solution and the full-data solution restricted to the same clubs; the median and 10th percentile are reported. To separate style from team quality, K-Means (k = 2, 3, 4) was repeated on feature groups (tactical: Poss, PPDA, OPPDA, DC; performance: xG, xGA, NPxGD, Sh/90, G/Sh, Opp_Gls_p90; without xG and xGA; without PPDA, OPPDA and DC), and agreement with the full-feature partition and with a K-Means partition on NPxGD alone was measured with ARI. DBSCAN sensitivity was examined over $\varepsilon \in \{1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ and min_samples $\in \{2, 3\}$. GMM was refitted with tied, diagonal and spherical covariance (n_init = 20, reg_covar = 0.01) and compared by BIC.

2.4.1 Silhouette Score

The Silhouette Score measures the similarity of a data point to its assigned cluster compared with

neighboring clusters. Higher values indicate better cluster separation and cohesion, as computed in Equation (7).

$$s(\mathrm{i}) = \frac{\mathrm{b(i)} - \mathrm{a(i)}}{\max(\mathrm{a(i)}, \mathrm{b(i)})} \tag{7}$$

where $a(i)$ denotes the average distance between data point i and all other points within the same cluster, while $b(i)$represents the average distance between data point i and the nearest neighboring cluster.

2.4.2 Davies-Bouldin Index

The Davies-Bouldin Index evaluates clustering quality by comparing within-cluster dispersion and between-cluster separation. Lower values indicate better clustering performance, as expressed in Equation (8).

$$DB = \frac{1}{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{k}} \max_{\mathrm{j} \neq \mathrm{i}} \left(\frac{\sigma_{\mathrm{i}} + \sigma_{\mathrm{j}}}{\mathrm{d}(\mu_{\mathrm{i}}, \mu_{\mathrm{j}})} \right) \tag{8}$$

where σ_i and σ_j denote the average distances of cluster members to their respective centroids, while $d(\mu_i, \mu_j)$ represents the distance between cluster centroids.

2.4.3 Calinski-Harabasz Index

The Calinski-Harabasz Index evaluates clustering performance based on the ratio between between-cluster dispersion and within-cluster dispersion. Higher values indicate more compact and well-separated clusters, as formulated in Equation (9).

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)} \tag{9}$$

where SSB denotes the between-cluster dispersion, SSW represents the within-cluster dispersion, k is the number of clusters, and n is the total number of observations.

In addition to internal validity indices, inter-algorithm consistency among K-Means, HAC, and GMM was evaluated using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). ARI measures the similarity between two clustering assignments corrected for chance, while NMI quantifies the mutual dependence between partitions normalized to the range [0, 1]. Higher values for both metrics indicate stronger agreement between algorithms; agreement does not by itself establish stability. DBSCAN was excluded from this analysis due to the presence of noise points, which are incompatible with pairwise label comparison metrics.

2.4.4 Elbow Method

The Elbow Method was employed to determine the optimal number of clusters prior to the comparative evaluation of the clustering algorithms. It plots the inertia, i.e. the K-Means objective J in Equation (2), against k, and the optimal k is identified at the point where the decrease in inertia begins to slow significantly, forming an elbow-like pattern in the curve.

3 Results and Discussion

3.1 Descriptive Statistics

The descriptive statistics of the 10 selected features across all 20 EPL clubs are presented in Table 3.

The features vary considerably across clubs (Table 3): possession ranges from 41.80% to 60.30%, PPDA from 9.24 to 14.42 (lower values indicate more intense pressing), and Deep Completions from 130 to 329. xG and xGA have identical league means (48.28) and the mean NPxGD is approximately zero because every chance created by one club is conceded by another.

To examine multicollinearity, a Pearson correlation matrix was computed for all 10 features (Figure 2). xG and xGA correlate strongly with NPxGD (r = 0.95 and r = −0.92, respectively), as expected because NPxGD is the difference between non-penalty xG and non-penalty xGA; xG and xGA are also negatively correlated with each other (r = −0.77). Possession is strongly correlated with DC (r = 0.91), Sh/90 (r = 0.86) and PPDA (r = −0.78). To test whether this redundancy drives the result, xG and xGA were removed (NPxGD retained). The resulting k = 4 partition agreed closely with the full-feature solution (ARI = 0.83; Silhouette = 0.244), although agreement was lower at k = 2 (ARI = 0.62) and k = 3 (ARI = 0.56).

Table 3 Descriptive Statistics

Feature	Min	Max	Mean	Std (n − 1)
Possession %	41.80	60.30	50.06	5.89
PPDA	9.24	14.42	11.47	1.71
Deep Completions	130.00	329.00	213.40	54.17
xG	30.29	65.85	48.28	10.92
xGA	26.43	69.81	48.28	8.59
NPxGD	-34.27	34.43	-0.001	17.13
Shots/90	9.28	15.97	12.42	2.01
Goal/Shot	0.07	0.12	0.10	0.02
Goals Conceded/90	0.72	1.84	1.31	0.28
OPPDA	9.23	15.60	11.50	2.05

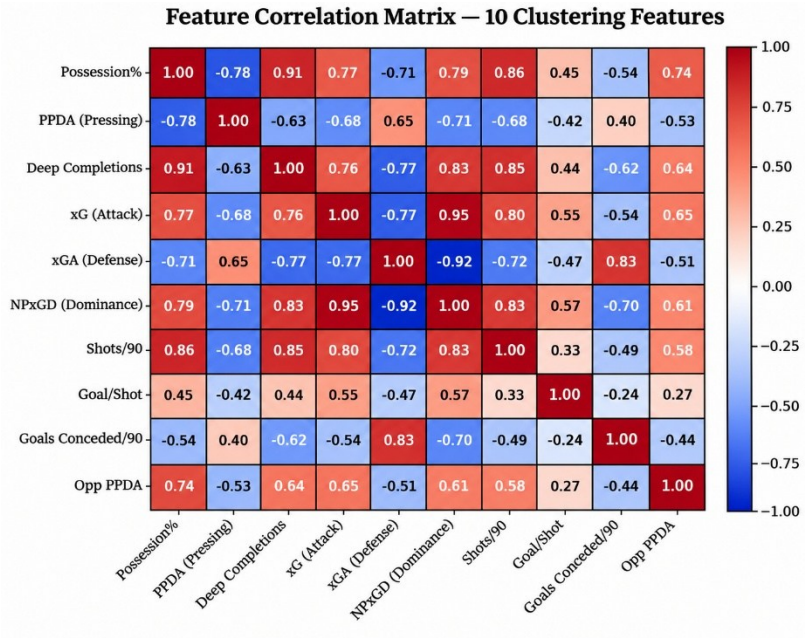

Figure 2 Correlation Matrix of 10 Features

3.2 Optimal Number of Clusters

The results of the optimal k analysis across k = 2 to k = 8 are presented in Figure 3. The Silhouette Score decreases from 0.3136 at k = 2 to 0.1337 at k = 8, suggesting that k = 2 produces the most statistically compact clusters. Similarly, the Elbow Method shows a gradual decrease in inertia with the most notable inflection observed between k = 2 and k = 3. At k=2, the resulting clusters merely separated high-performing clubs from the rest, which reflects overall team quality rather than distinct playing styles. This binary separation fails to capture the tactical variation between, for example, pressing-oriented mid-table clubs and defensively passive ones. Because k = 2 and k = 3 largely reproduce quality tiers, k = 4 was selected as a finer partition that is easier to interpret, although no internal index favors it. Subsampling stability (Table 4) gave a median ARI of 0.69 for k = 4, comparable to k = 2 (0.75) and k = 3 (0.78); k = 4 is therefore not less stable than the alternatives, but its stability is moderate and its selection rests on interpretability rather than on internal indices.

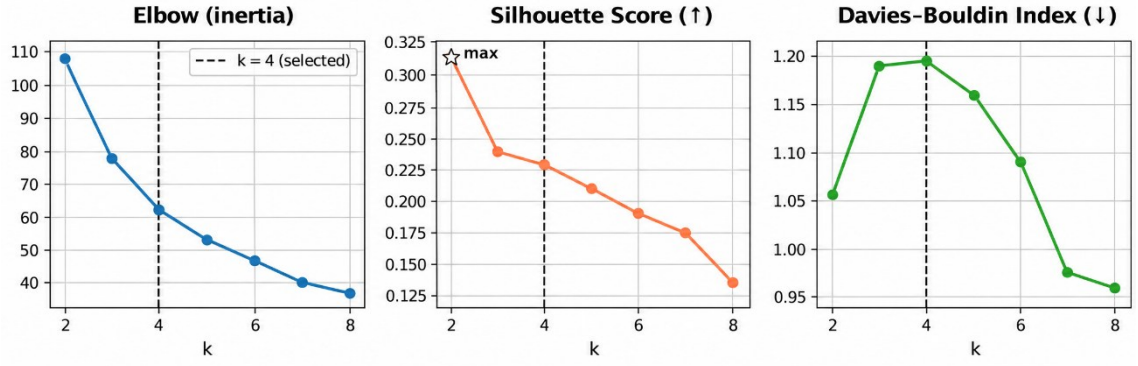

Figure 3 Optimal Number of Clusters

At k = 3 the partition forms top, middle and lower-table tiers (five, nine and six clubs), i.e. an ordering of team quality. k = 4 is not a nested refinement of it: it separates three clubs with the lowest possession and deep completions and the highest PPDA (Burnley, West Ham United, Wolves), and regroups Crystal Palace, Fulham and Leeds United with the lower-table clubs.

Table 4 Comparison of Clustering Quality Metrics Across k = 2, 3, and 4

K	Silhouette Score	Davies-Bouldin	Calinski-Harabasz	Median ARI (P10)	Cluster sizes
2	0.3136	1.0583	15.25	0.75 (0.21)	10/10
3	0.2395	1.1916	13.17	0.78 (0.28)	9/5/6
4	0.2287	1.1939	11.69	0.69 (0.36)	4/7/3/6

3.3 Comparative Evaluation of Clustering Methods

The cluster visualizations using PCA 2D projection for K-Means and Hierarchical Agglomerative Clustering are presented in Figure 4. The evaluation results of the four clustering methods at k = 4 are summarized in Table 5.

DBSCAN (ε = 2.0, min_samples = 2) identified three non-noise clusters (Arsenal and Manchester City; Bournemouth, Brighton, Manchester United and Newcastle United; Everton, Nottingham Forest and Sunderland) and labeled 11 of 20 clubs (55%) as noise, including Chelsea, Liverpool and Aston Villa. Its metrics were computed on the nine non-noise clubs only and are

excluded from the ranking. The result is also highly parameter-sensitive: for ε ∈ {1.5, 2.0, 2.5, 3.0, 3.5, 4.0} and min_samples ∈ {2, 3}, only ε = 2.0 with min_samples = 2 produced three clusters; ε ≥ 3.0 merged all clubs into one cluster, and ε = 1.5 (min_samples = 2) left 18 clubs as noise because only one pair of clubs (Bournemouth and Newcastle United) lay within that distance of each other.

Among the three comparable methods evaluated on all 20 clubs, K-Means achieved the highest Silhouette Score (0.2287) and Calinski–Harabasz Index (11.69), while HAC produced the lowest Davies–Bouldin Index (1.1670); GMM ranked last on all three indices. The differences between K-Means and HAC are small (0.0061, 0.11 and 0.0269, respectively), and neither method dominates across all indices. K-Means and HAC assign 19 of 20 clubs identically; Fulham is the only exception (K-Means: Cluster 4; HAC: Cluster 2). The choice between them therefore rests on practical considerations rather than on the indices: K-Means yields directly interpretable centroids for profiling but depends on initialization (mitigated here by n_init = 10), whereas HAC is deterministic and provides a dendrogram of nested relationships but cannot revise early merges. We use K-Means for profiling and regard HAC as an equally valid alternative.

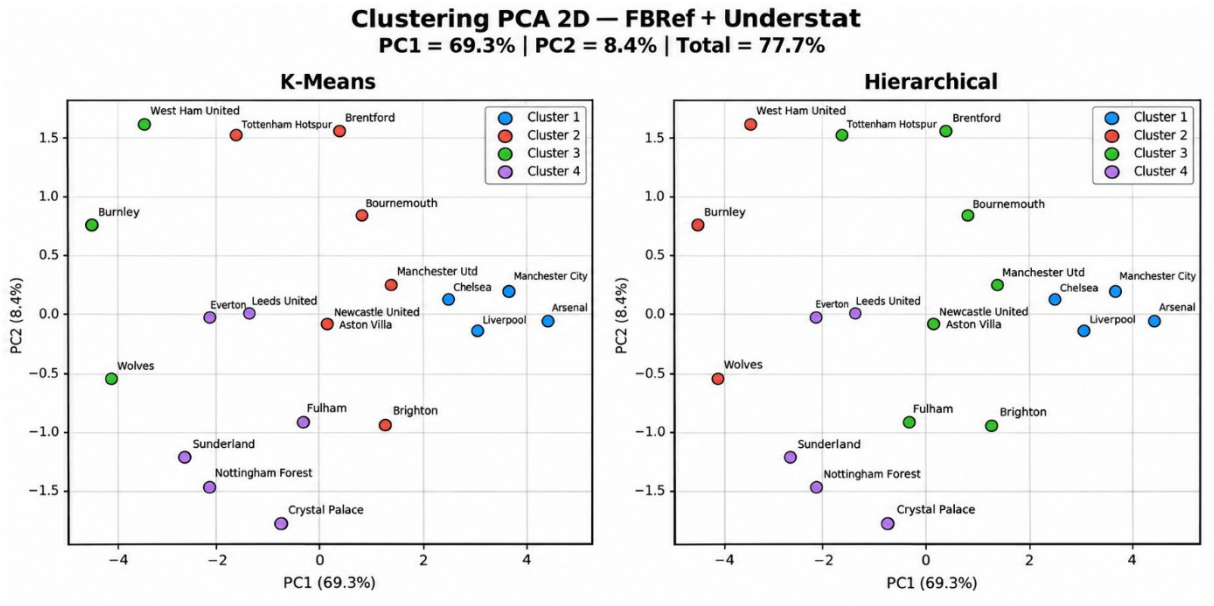

Figure 4 PCA 2D Projection, K-Means and Hierarchical Agglomerative Clustering

Table 5 Evaluation Results

Method	Silhouette	Davies-Bouldin	Calinski-Harabasz
K-Means	0.2287	1.1939	11.69
Hierarchical	0.2226	1.1670	11.58
GMM	0.2140	1.2124	11.20
DBSCAN*	0.4785*	0.6418*	13.90*

*DBSCAN (ε = 2.0, min_samples = 2): 3 clusters, 11 of 20 clubs (55%) labeled as noise; metrics computed on the 9 non-noise clubs only and excluded from the ranking.

3.4 Inter-Algorithm Consistency

To asses whether the three comparable algorithms assign EPL clubs to consistent groupings, inter-algorithms agreement was measured using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

K-Means and Hierarchical Agglomerative Clustering achieved the highest agreement with ARI = 0.828 and NMI = 0.892, indicating strong consistency in club assignment. K-Means and GMM produced ARI = 0.734 and NMI = 0.851, while Hierarchical and GMM yielded the lowest pairwise agreement with ARI = 0.563 an NMI = 0.743. Agreement is high between K-Means and HAC and lower for GMM, whose full-covariance fit is unreliable at this sample size (Section 3.6). Agreement between algorithms shows that the partitions are consistent at k = 4 but does not by itself establish stability or the validity of k = 4.

3.5 Playing Style Profiles

The K-Means clustering with k = 4 identified four distinct playing style profiles among EPL clubs. The complete cluster profiles are summarized in Table 6 and visualized as a heatmap in Figure 5.

Table 6 Playing Style Profiles with values rounded half-up

Cluster	Playing Style	Poss (%)	PPDA	DC	xG	xGA	NPxGD
C1	Possession Dominant and High Press	58.88	9.81	288.00	62.73	38.38	+22.49
C2	Pressing and Transition	51.41	10.25	220.29	50.41	46.93	+3.58
C3	Low Block and Direct Play	42.57	13.65	142.00	34.48	62.02	-25.76
C4	Low Block and Vulnerable Defense	46.35	12.92	191.33	43.07	49.59	-6.29



Figure 5 Cluster Profile Heatmap

Cluster 1, labeled Possession Dominant and High Press, comprises Arsenal, Chelsea, Liverpool, and Manchester City. This cluster has the highest mean possession (58.88%), the lowest PPDA (9.81), the highest Deep Completions (288.00) and xG (62.73), and the lowest xGA (38.38) and goals conceded per 90 minutes (1.05). Its NPxGD of +22.49 indicates superiority in both attacking and defensive phases, and its mean xPTS (59.32) is the highest of the four clusters.

Cluster 2, labeled Pressing and Transition, includes Aston Villa, Bournemouth, Brentford, Brighton, Manchester United, Newcastle United, and Tottenham Hotspur. Its profile lies close to the league average on most features (possession 51.41%, Deep Completions 220.29, xG 50.41, xGA 46.93), with a PPDA of 10.25 indicating more intense pressing than the league means (11.47) but less than Cluster 1. The NPxGD of +3.58 shows that these clubs are only marginally better than the league average overall.

Cluster 3, labeled Low Block and Direct Play, consists of Burnley, West Ham United, and Wolves. These clubs have the lowest mean possession (42.57%), the highest PPDA (13.65), and the lowest Deep Completions (142.00). Their xG (34.48) is the lowest and their xGA (62.02) the highest of the four clusters, giving an NPxGD of −25.76 and the lowest mean xPTS (28.49). The features used here describe possession, pressing and chance quality; they do not measure directness of play, so the "Direct Play" label should be read as an interpretation of low possession and low build-up activity rather than a measured attribute.

Cluster 4, labeled Low Block and Vulnerable Defense, comprises Crystal Palace, Everton, Fulham, Leeds United, Nottingham Forest, and Sunderland. Like Cluster 3, it shows passive pressing (PPDA = 12.92), but with higher possession (46.35%) and Deep Completions (191.33). Its xGA (49.59) is close to the league mean (48.28), so its negative NPxGD (−6.29) stems mainly from low chance creation (xG = 43.07) rather than from an exceptionally weak defense. The "Vulnerable Defense" label should therefore be read as relative to Clusters 1 and 2.

3.6 GMM Soft Assignments

All 20 clubs received a membership probability of approximately 1.0 for a single component. This should not be read as evidence of well-separated styles: a full-covariance GMM with four components on 20 observations in 10 dimensions has 55 free covariance parameters per component but only about five clubs per component, so the fit is underdetermined and posteriors saturate; the same saturation occurred on random Gaussian data of identical size. Refitting with n_init = 20, reg_covar = 0.01 and tied, diagonal or spherical covariance reduced saturation only partly (share of clubs with maximum posterior > 0.99 at k = 4: 100%, 95% and 85%). BIC was lowest at k = 2 under tied covariance (463.0 vs 489.5 at k = 4) but at k = 4 under diagonal (532.5) and spherical (517.8) covariance. GMM soft assignments were therefore not used for profiling.

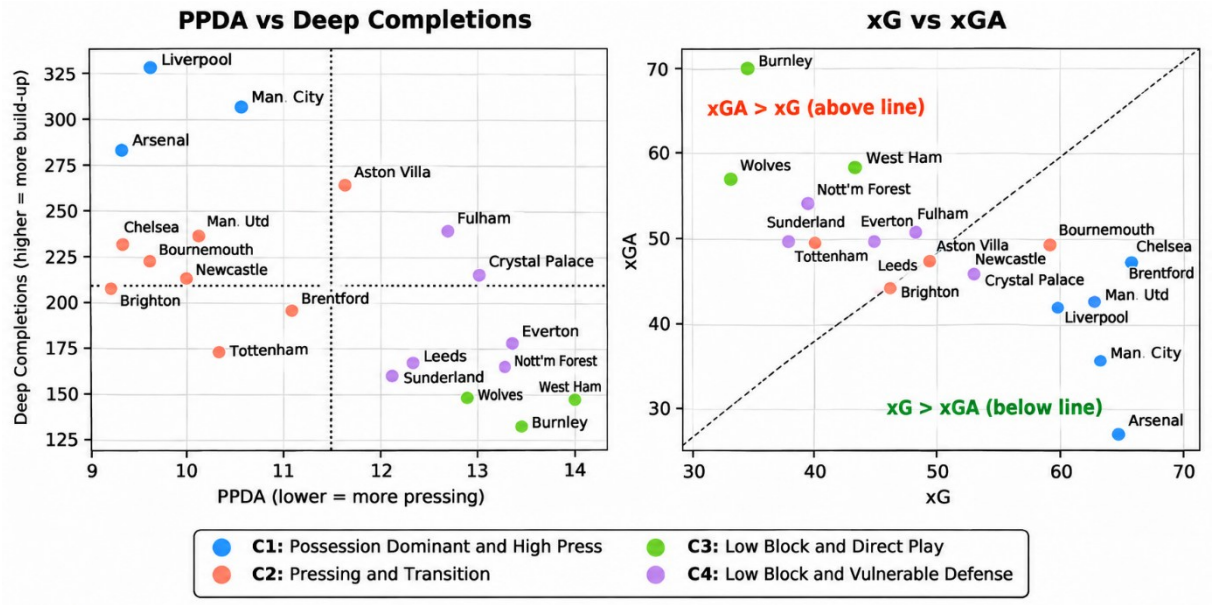

Figure 6 PPDA vs DC & xG vs xGA

3.7 PPDA vs Deep Completions Analysis

Figure 6 presents the scatter plots of PPDA against Deep Completions and xG against xGA, which provide direct visualizations of pressing intensity, build-up play patterns, and the balance between offensive and defensive quality across all 20 EPL clubs.

In the PPDA versus Deep Completions plot, Cluster 1 clubs combine low PPDA (9.37–10.66) with high DC (233–329), although they are not the extremes on both: Brighton has the lowest PPDA in the league (9.24) and Chelsea's DC (233) is below Aston Villa's (266). Cluster 3 clubs occupy the high-PPDA, low-DC region, while Clusters 2 and 4 lie in the intermediate region. In the xG versus xGA plot, Cluster 1 clubs lie in the high-xG, low-xGA region and Cluster 3 clubs in the low-xG, high-xGA region, on opposite sides of the balance line, which is consistent with the quality gradient discussed below.

3.8 Discussion

The results of this study reveal several patterns that warrant deeper interpretation in the context of football analytics literature. The four playing style profiles identified through K-Means clustering reflect mainly a gradient of team quality, with secondary variation in pressing and build-up.

Cluster 1 combines the highest possession, deep completions and xG with the lowest PPDA and xGA. Because these features are strongly associated with team quality, this cluster is better described as the strongest clubs than as a distinct tactical style.

Cluster 2 lies close to the league average on most features, with somewhat more intense pressing (PPDA = 10.25 vs a league mean of 11.47). Its seven clubs show that moderate pressing is not exclusive to top-performing clubs, but the cluster's NPxGD (+3.58) indicates only a marginal advantage overall.

Clusters 3 and 4 both combine passive pressing (PPDA = 13.65 and 12.92) with negative NPxGD. They differ mainly in build-up and chance creation: Cluster 3 has the lowest possession (42.57%) and Deep Completions (142.00) and the weakest xG and xGA, whereas Cluster 4 has higher possession (46.35%) and Deep Completions (191.33) and an xGA (49.59) close to the league mean.

The separation between Clusters 1 and 3 reflects the two extremes of the quality gradient: pressing intensity and build-up are strongly associated with each other and with team quality (Poss–DC $r = 0.91$; Poss–PPDA $r = -0.78$).

To separate style from quality, K-Means was repeated on feature groups, and the results are summarized in Table 7. The leading principal component of the full feature set explains 69.3% of the variance and correlates with xPTS, which was not a clustering feature, at $|r| = 0.95$; the clusters are ordered by mean xPTS (C1 59.3, C2 46.4, C4 39.8, C3 28.5). Performance-only clustering at $k = 2$ closely reproduced a quality-only partition from NPxGD (ARI = 0.80), whereas tactical-only clustering did not (ARI = 0.21), so some tactical variation exists beyond team quality. Removing PPDA, OPPDA and DC left the $k = 2$ partition unchanged (ARI = 1.00 against the full-feature solution), suggesting that the two-group split is carried by performance features. At $k = 4$, however, the tactical-only solution produced a singleton cluster (Fulham) and the performance-only solution was only weakly related to the full solution (ARI = 0.37), so a four-cluster structure is not supported by either feature group alone. The full-feature clusters are therefore best read as a quality gradient with secondary tactical variation, not as four independent playing styles.

Table 7 Feature-Group Ablation (K-Means)

Feature set	k = 2 Sil.	k = 2 ARI vs full	k = 2 ARI vs NPxGD	k = 4 Sil.	k = 4 sizes	k = 4 ARI vs full
Tactical	0.449	0.62	0.21	0.413	5, 8, 6, 1	0.59
Performance	0.292	0.62	0.80	0.314	2, 4, 9, 5	0.37
Without xG, xGA	0.325	0.62	0.21	0.244	5, 4, 8, 3	0.83
Without PPDA, OPPDA, DC	0.295	1.00	0.46	0.280	2, 4, 11, 3	0.37
NPxGD only	0.540	0.46	1.00	0.561	9, 7, 2, 2	0.18

Tactical = Poss, PPDA, OPPDA, DC; Performance = xG, xGA, NPxGD, Sh/90, G/Sh, Opp_Gls_p90. Sil. = Silhouette (not comparable across feature sets); ARI vs full = against the 10-feature partition at the same k; ARI vs NPxGD = against the NPxGD-only partition at k = 2. Cluster sizes follow arbitrary K-Means indexing.

4 Conclusion

This study applied an unsupervised machine learning framework to identify and profile the playing styles of all 20 English Premier League clubs during the 2025/2026 season using publicly available data from FBref and Understat. Four clustering methods, namely K-Means, Hierarchical Agglomerative Clustering, GMM, and DBSCAN, were implemented and evaluated using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index.

The comparative evaluation demonstrated that K-Means and HAC produced near-identical partitions (19/20 clubs) at k = 4, while DBSCAN, despite achieving superior quantitative scores, classified the majority of clubs as noise due to the small dataset size and was therefore unsuitable as the primary method. Four exploratory profiles were identified, ordered largely by team quality: (1) Possession Dominant and High Press, characterized by Arsenal, Chelsea, Liverpool, and Manchester City; (2) Pressing and Transition, comprising seven clubs including Aston Villa and Newcastle United; (3) Low Block and Direct Play, consisting of Burnley, West Ham United, and Wolves; and (4) Low Block and Vulnerable Defense, encompassing six mid-to-lower table clubs.

This study makes four key contributions. First, it introduces a comparative framework for four unsupervised learning methods within a unified analytical pipeline for football playing style analysis. Second, it provides empirical playing style profiles of all 20 EPL clubs based on multidimensional publicly available statistics. Third, it examines what structure can be recovered from public aggregate data and shows that most of it reflects a gradient of team quality. Fourth, it illustrates a reproducible workflow that analysts can apply to public data.

The primary limitation is the small sample (n = 20 clubs, p = 10 features, i.e. two observations per feature), which makes the partitions sensitive to feature selection and algorithm choice, as the stability and ablation analyses show. The data are a single snapshot (matchday 32, retrieved on 16 April 2026), so changes in style over the season are not captured. There is no independent tactical ground truth, and no comparison with proprietary event data was made. Future work should extend the analysis to multiple seasons or rolling windows and test sensitivity to scaling. Squad passing and possession tables from FBref were unavailable at the time of data collection, so PPDA and Deep Completions from Understat were used as proxies for pressing and build-up; more granular passing and possession variables could yield more nuanced profiles in future work.

Declaration of the Use of Generative AI

The authors used Claude (Anthropic) and ChatGPT (OpenAI, GPT-4) as research assistants, including for brainstorming, structural ideation, and language editing. The authors confirm full responsibility for the content and scientific integrity of this manuscript.

Author Contributions

Fahri Syaifullah: Conceptualization, methodology, Software, data analysis, and validation, and writing - original draft. Amiq Fahmi: Supervision, review, and editing. All authors have read and approved the final manuscript.

Conflict of interest

The authors declare that they have no conflicts of interest.

Funding

This research received no external funding.

References

[1] M. Rico-González, J. Pino-Ortega, A. Méndez, F. Clemente, and A. Baca, "Machine learning application in soccer: a systematic review," *Biol. Sport*, vol. 40, no. 1, pp. 249–263, 2023, doi: 10.5114/biolsport.2023.112970.

[2] R. Bunker, C. Yeung, and K. Fujii, "Machine Learning for Soccer Match Result Prediction," 2025, *Artificial Intelligence, Optimization, and Data Sciences in Sports,* doi: https://doi.org/10.1007/978-3-031-76047-1_2.

[3] A. Rossi, L. Pappalardo, and P. Cintia, "A Narrative Review for a Machine Learning Application in Sports: An Example Based on Injury Forecasting in Soccer," *Sports*, vol. 10, no. 1, p. 5, Dec. 2021, doi: 10.3390/sports10010005.

[4] A. Ati, P. Bouchet, and R. Ben Jeddou, "Using multi-criteria decision-making and machine learning for football player selection and performance prediction: a systematic review," *Data Sci. Manag.*, vol. 7, no. 2, pp. 79–88, Jun. 2024, doi: 10.1016/j.dsm.2023.11.001.

[5] E. F. E. Atta Mills, Z. Deng, Z. Zhong, and J. Li, "Data-driven prediction of soccer outcomes using enhanced machine and deep learning techniques," *J. Big Data*, vol. 11, no. 1, p. 170, Nov. 2024, doi: 10.1186/s40537-024-01008-2.

[6] J. H. Hewitt and O. Karakuş, "A machine learning approach for player and position adjusted expected goals in football (soccer)," *Frankl. Open*, vol. 4, p. 100034, Sep. 2023, doi: 10.1016/j.fraope.2023.100034.

[7] A. Majumdar, R. Bakirov, D. Hodges, S. Scott, and T. Rees, "Machine Learning for Understanding and Predicting Injuries in Football," *Sports Med. - Open*, vol. 8, no. 1, p. 73, Dec. 2022, doi: 10.1186/s40798-022-00465-4.

[8] S. Manish., V. Bhagat, and R. Pramila, "Prediction of Football Players Performance using Machine Learning and Deep Learning Algorithms," in *2021 2nd International Conference for Emerging Technology (INCET)*, Belagavi, India: IEEE, May 2021, pp. 1–5. doi: 10.1109/INCET51464.2021.9456424.

[9] M. S. Croock, A. N. Rashid, and B. M. Merzah, "Intelligent Classifiers for Football Player Performance Based on Machine Learning Models," *Int. J. Electr. Comput. Eng. Syst.*, vol. 15, no. 2, pp. 173–183, Feb. 2024, doi: 10.32985/ijeces.15.2.6.

[10] L. Stival *et al.*, "Using machine learning pipeline to predict entry into the attack zone in football," *PLOS ONE*, vol. 18, no. 1, p. e0265372, Jan. 2023, doi: 10.1371/journal.pone.0265372.

[11] S. Plakias *et al.*, "Identifying Soccer Teams' Styles of Play: A Scoping and Critical Review," *J. Funct. Morphol. Kinesiol.*, vol. 8, no. 2, p. 39, Mar. 2023, doi: 10.3390/jfmk8020039.
[12] S. Plakias *et al.*, "Identifying Soccer Players' Playing Styles: A Systematic Review," *J. Funct. Morphol. Kinesiol.*, vol. 8, no. 3, p. 104, Jul. 2023, doi: 10.3390/jfmk8030104.
[13] F. A. R. Clavijo, R. Drews, R. A. Denardi, B. Travassos, and U. C. Corrêa, "Identification of football teams styles of play by cluster analysis," *Int. J. Sports Sci. Coach.*, vol. 19, no. 3, pp. 1019–1034, Jun. 2024, doi: 10.1177/17479541231186796.
[14] C. Zhou, C. Lago-Peñas, A. Lorenzo, and M.-Á. Gómez, "Long-Term Trend Analysis of Playing Styles in the Chinese Soccer Super League," *J. Hum. Kinet.*, vol. 79, pp. 237–247, Jul. 2021, doi: 10.2478/hukin-2021-0077.
[15] L. Kong, T. Zhang, C. Zhou, M.-A. Gomez, Y. Hu, and S. Zhang, "The evaluation of playing styles integrating with contextual variables in professional soccer," *Front. Psychol.*, vol. 13, p. 1002566, Sep. 2022, doi: 10.3389/fpsyg.2022.1002566.
[16] A. Martín-Castellanos, M. R. Flores, D. M. Solana, R. L. Del Campo, F. N. Garrosa, and D. Mon-López, "How do the football teams play in LaLiga? Analysis and comparison of playing styles according to the outcome," *Int. J. Perform. Anal. Sport*, vol. 24, no. 1, pp. 18–30, Jan. 2024, doi: 10.1080/24748668.2023.2262813.
[17] L. Ruan, H. Ge, M.-Á. Gómez, Y. Shen, B. Gong, and Y. Cui, "Analysis of defensive playing styles in the professional Chinese Football Super League," *Sci. Med. Footb.*, vol. 7, no. 3, pp. 279–287, Jul. 2023, doi: 10.1080/24733938.2022.2099964.
[18] Q. He, D. Araújo, K. Davids, Y. H. Kee, and J. Komar, "Functional adaptability in playing style: A key determinant of competitive football performance," *Adapt. Behav.*, vol. 31, no. 6, pp. 545–558, Dec. 2023, doi: 10.1177/10597123231178942.
[19] S. J. Moffatt, R. Gupta, N. F. Collier, and B. S. Keller, "Classifying and quantifying team playing styles in the Australian Football League," *Int. J. Perform. Anal. Sport*, vol. 25, no. 3, pp. 562–579, May 2025, doi: 10.1080/24748668.2024.2325269.
[20] Y. Li, S. Zong, Y. Shen, Z. Pu, M.-Á. Gómez, and Y. Cui, "Characterizing player's playing styles based on player vectors for each playing position in the Chinese Football Super League," *J. Sports Sci.*, vol. 40, no. 14, pp. 1629–1640, Jul. 2022, doi: 10.1080/02640414.2022.2096771.
[21] "Premier League Stats: Squad Shooting Stats," FBref. Accessed: Apr. 16, 2026. [Online]. Available: https://fbref.com/en/comps/9/2025-2026/2025-2026-Premier-League-Stats#all_stats_squads_shooting
[22] "2025–2026 Premier League Stats: Squad and Opponent Standard Stats," FBref. Accessed: Apr. 16, 2026. [Online]. Available: https://fbref.com/en/comps/9/2025-2026/2025-2026-Premier-League-Stats
[23] "EPL xG Table and Scorers for the 2025/2026 season," Understat, Accessed: Apr. 16, 2026. [Online]. Available: https://understat.com/league/EPL/2025