BMC Sports Science, Medicine and Rehabilitation

RESEARCH Open Access

Winning determinants between top and second division in Chinese professional football leagues: an interpretable machine learning approach

Bo Yuan[1], Jiaxuan Zhu[1], Pengyu Pan[2], Dingmeng Ren[3], Zheng Liang[4], Honglin Song[1] and Tianbiao Liu[1*]

Abstract
This study investigated the impact of winning determinants in two professional soccer leagues. The sample was composed of 1,440 Chinese Super Football League (CSL) and Chinese Football Association China League (CFACL) matches (CSL matches = 720; CFACL matches = 720) during the 2017–2019 seasons. The study employed eXtreme Gradient Boosting (XGBoost) to assess the importance of 25 indicators exhibiting significant differences ($p < 0.05$) in their association with match outcomes, and the SHapley Additive explanations (SHAP) was utilized to interpret these findings. The results showed that scoring performance indicators, such as Shots On Target Inside Box (SOTIB), Shots, and Shots On Target (SOT), significantly influenced outcomes in both the CSL ($\boldsymbol{S_G}$=37.854%) and CFACL ($\boldsymbol{S_G}$=38.934%), with SOTIB being the most impactful. Additionally, this study found that defensive feature clearances were highly influential in both leagues, ranking second only to SOTIB of variable importance. Meanwhile, defensive feature fouls were a more significant factor in determining match outcomes in the CFACL than in the CSL. In both the CSL and CFACL, players must prioritize precision in shooting within the penalty area rather than merely increasing the frequency of shots. For CFACL teams, if consistent high-quality passing is unattainable, effective use of set pieces (e.g., free kicks) could serve as an alternative strategy to organize attacks. These findings can assist coaches in formulating tailored tactical strategies suited to the distinct demands of each league level.

Keywords Soccer, Match performance, Winning determinants, Machine learning, Performance analysis, SHAP

Introduction

For a long time, the number of goals scored is crucial for determining match outcomes, but the factors that influence the match outcomes are complex. The performance indicators that predict success in soccer matches have long been a focus of international soccer research [1]. Previous research on soccer match-winning factors has preliminarily analyzed the impact on match outcomes from various perspectives, such as running distance [2], possession [3], pass success ratio, and shooting efficiency [4]. Recent research has emphasized the integration of multiple methods and variables for more comparative

*Correspondence:
Tianbiao Liu
ltb@bnu.edu.cn
[1]College of Physical Education and Sport, Beijing Normal University, Beijing, China
[2]Institute of Exercise Training and Sports Informatics, German Sport University Cologne, Cologne, Germany
[3]China Football College, Beijing Sport University, Beijing, China
[4]Department of Sport, Hebei University of Technology, Tianjin, China

analysis, rather than isolating a single factor in determining match outcomes [5, 6]. Previous studies have classified team performance indicators into three distinct categories: variables related to goal scoring, variables related to passing and organizing, and variables related to defending [5–7]. This tripartite framework provides a more precise and granular set of descriptors for analyzing match performance dynamics. For instance, one study noted that technical performance is crucial to match outcomes [8, 9]. Another investigation related to winning discriminants in the Bundesliga found that defensive errors and goal efficiency are both significant factors, and cross is a negative indicator [10].

Furthermore, disparities in match performance may exist across different leagues and teams, reflecting the distinct playing styles and strategic priorities inherent in each league [11]. For instance, some researchers investigated the technical tactical and running performance in Span's first and second divisions league and found that technical-tactical performance variables have a greater influence on a team's success than running performance [12]. Another research on the Spanish soccer leagues over eight seasons found that technical-tactical variables, like passes and successful passes in the first division, are more than in the second division [13].

Additionally, several studies have investigated match performance in European leagues, researchers have a common view that soccer leagues with different cultural backgrounds displayed different playing characteristics, as the English Premier League showed more tackles, and long passes than Italian Serie A, and Spanish La Liga showed more passes than English Premier League [14]. Previous research on CSL indeed found disparities compared with the literature of European leagues. For instance, some researchers compared the actual playing time between the CSL and the English Premier League and suggested that while non-anthropogenic factors influenced both leagues, the actual playing time in the CSL was more influenced by anthropogenic factors [15]. Therefore, it is crucial to compare the relationships and impacts of winning determinants across different leagues. Machine learning algorithms offer the capability to process intricate datasets and generate real-time insights [16]. For example, XGBoost is excellent at handling diverse and complex data structures to make accurate predictions, while SHAP can directly identify the most critical features from matches [17]. In recent years, the SHapley Additive exPlanations (SHAP) model has been applied to soccer match performance analysis [17–19], demonstrating strong predictive capabilities in assessing team performance, such as predicting match outcomes [17, 18] and forecasting player injuries [17, 20].Furthermore, the winning determinants in soccer across different leagues are inherently complex, requiring consideration of multiple dimensions [5, 21], including technical-tactical performance, running performance, physical conditioning, and external factors. As each of these factors interacts in intricate ways, it is essential to employ machine learning to uncover the patterns and insights that contribute to winning outcomes.

In summary, previous studies have compared differences in match performance indicators across teams of different competitive levels. However, these indicators may not fully capture the critical factors influencing team success, as teams at distinct levels face opponents of varying strengths. This necessitates a focused comparison of performance metrics among teams within the same competitive tier. Furthermore, while certain performance indicators exhibit significant differences in winning matches for teams at the same level, their relative importance to match outcomes remains unclear. In football, the relationship between performance data and match results is not strictly linear. Conventional linear regression may lack precision in analyzing such relationships. In contrast, the XGBoost and SHAP models enable robust analysis and prediction of non-linear data, effectively identifying the significance of performance indicators and elucidating complex data relationships [19]. Additionally, we have incorporated considerations of cultural differences and disparities in league competitiveness. By conducting comparative analyses within Chinese football leagues and contextualizing findings against China's football development stage and top-tier league performance, this study provides a more nuanced discussion of the results.

Based on the above-mentioned context, the aims of this study were to: [1] analyze the variables influencing the outcomes of games between the Chinese Football League (CFL) and Chinese Football Association Champions League (CFACL) teams; and [2] integrate XGBoost and SHAP to assess the contributions of goal scoring performance, passing and organizing performance, and defending performance to the success of teams in both leagues.

Method

Sample and data

In this study, the data consisted of 1,440 matches from CSL and CFACL during three seasons (from 2017 to 2019). Each division includes 720 matches (CSL matches = 720; CFACL matches = 720). All the data were obtained from Champdas Football Big Data Company (http://www.champdas.com), and a semi-automatic soc cer match analysis system was developed by Champdas Soccer Big Data Company. The reliability and validity of the Champdas Football Data collection and analysis system have been validated by other researchers [22]. The dataset covers comprehensive match statistics that allowed for in-depth performance analysis between winning, losing and drawing teams in both divisions.

Variables and procedures

Based on prior studies, the data was initially processed using Excel and subsequently imported into R-Studio for further cleaning and analysis. Twenty-five variables were selected as performance data variables and they were divided into three groups according to the available literature: Variables related to goal scoring, Variables related to passing and organizing, and Variables related to defending (Table 1) [7, 23–25]. Afterward, One-way ANOVA was conducted to assess the differences between winning, losing teams and drawing teams. For each indicator, the mean and standard deviation (Mean ± SD) were calculated to determine whether there were statistically significant differences between the three groups ($p < 0.05$). Additionally, effect sizes were computed to measure the magnitude of the difference between the performance of winning, losing and drawing teams. The effect sizes were also reported as partial eta-squared (ηp^2) [23].All variables exhibiting significant differences ($p < 0.05$) in their association with match outcomes will undergo further analysis. Data was analysed with IBM SPSS Statistics. The significance level was set to $p < 0.05$ for all statistical analyses.

Machine learning

The XGBoost algorithm, developed by Chen et al. in 2016 [26], is an enhanced version of the Gradient Boosting Decision Tree(GBDT). It improves the boosting algorithm by combining multiple weak classifiers to form a strong classifier, using a series of optimizations such as a novel sparsity-aware algorithm for handling sparse data, and a weighted quantile sketch for approximate tree learning. These improvements enable XGBoost to achieve high prediction accuracy and efficient computation, making it a popular choice for machine learning tasks involving large-scale datasets.

The XGBoost algorithm is an optimized gradient-boosting framework that combines weak classifiers into a strong learner. Its objective function minimizes both prediction error and model complexity through regularization. The model prediction for $\widehat{y}$ input x_i is expressed as:

$$\hat{y} = \sum_{k=1}^{K} f_k\left(x_i\right), f_k F \#1$$

To address inconsistencies in feature importance evaluation challenges, SHAP, a game theoretic approach, was introduced to interpret the model output [27]. SHAP combines global and local interpretations to provide insights into the 'black box' models [28]. Equation (2) represents the SHAP formula

$$\phi_j = \sum_{S \subseteq \mathcal{M}\{j\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[f\left(S \cup \{j\}\right) - f\left(S\right)\right] \#2$$

This study integrates XGBoost and SHAP to analyze the winning factors in two divisions. The target variable of the model is the match outcomes, and the feature set includes 19 indicators, such as Goals, Clearances, and others. The model parameters are optimized to improve performance through training. Subsequently, the SHAP model is introduced to interpret the results of the XGBoost model.

Construction of group influence formula

This study analyzes 1,102 matches from CSL and CFACL using XGBoost and SHAP models. By formulating Eq. (3), we further investigate the winning factors categorized into goal scoring performance, passing and organizing performance, and defending performance. This approach enables a deeper understanding of the factors influencing outcomes in both levels of competition.

$$S_x = \frac{SHAP_x}{SHAP_{total}} \times 100\%, x \in \{G, O, D\} \# (3)$$

Here, S_x represents the SHAP proportion of each category including goal scoring performance (S_G), passing and organizing performance (S_O), and defending performance(S_D). The $SHAP_x$ value for category x (where x can be goal scoring performance, passing and organizing performance, or defending performance). And $SHAP_{total}$ indicates the sum of SHAP values for all categories.

Results

According to the results displayed in Table 2, in terms of Variables related to goal scoring, the winning teams in CSL showed significantly differences in SOTIB ($p < 0.001$, ES = 0.163), SOT ($p < 0.001$, ES = 0.136), SIPA ($p < 0.001$, ES = 0.058), Shots ($p < 0.001$, ES = 0.033), Penalties ($p < 0.001$, ES = 0.022), and SOTOB ($p < 0.001$, ES = 0.014) compared to the drawing teams and losing teams. In CFACL, winning teams also showed significantly differences than drawing teams and losing teams in SOTIB ($p < 0.001$, ES = 0.161), SOT ($p < 0.001$, ES = 0.154), SIPA ($p < 0.001$, ES = 0.054), Shots ($p < 0.001$, ES = 0.04), SOTOB ($p < 0.001$, ES = 0.024), Penalties ($p < 0.001$, ES = 0.015).

With regard to *Variables related to passing and organizing*, winning teams in CSL exhibited a significantly higher values in ATPA ($p < 0.001$, ES = 0.024), Passes ($p < 0.001$, ES = 0.013), as well as in ATP ($p < 0.001$, ES = 0.011) and Corners ($p < 0.001$, ES = 0.005) compared to drawing teams and losing teams. In CFACL, the winning teams

Table 1 Definitions of the variables for each dimension

Variables related to goal scoring: operational definition	
Shots	An attempt to score a goal, made with any (legal) part of the body, either on or off target
Shots on target (SOT)	An attempt to score a goal, which required intervention to stop the ball going in or resulted in a goal/shot that would have gone in without diversion
Shots Out Penalty Area (SOPA)	A shot from outside the penalty area
Shots On Target Outside Box (SOTOB)	A shot on target from outside the penalty area
Shots Inside Penalty Area (SIPA)	A shot inside the penalty area
Shots On Target Inside Box (SOTIB)	A shot on target inside the penalty area
Penalties	Player fouled within the penalty box leading to a penalty kick
Variables related to passing and organizing: operational definition	
Free Kicks	Number of free kicks awarded
Front Free Kicks	Number of free kicks awarded on the opponent's half of the pitch
Corners	Ball goes out of play for a corner kick
Breakthrough	Number of successful dribbles past opposing players
Pass	An intentional played ball from one player to another
Passes Success Rate	Successful passes as a proportion of the total passes
Key Pass	The final pass assisting a shot (without scoring)
Attacking Third Passes(ATP)	Number of passes of the ball (possessed by the attacking team) in the 35 m area of the opponent's half of the pitch
Attacking Third Successful Pass Accuracy(ATPA)	Number of successful passes of the ball (possessed by the attacking team) as a proportion of the total passes in the 35 m area of the opponent's half of the pitch
Cross	Balls sent into the central area of the box from a wide position of the attacking third
Variables related to defending: operational definition	
Tackle	The action of gaining possession from an opposition player who is in possession of the ball
Interception	A player intercepts a pass between oppositions and prevents the opponent receiving the ball
Clearances	A player kicks or hits the ball away from the goal of his or her own team without a precise target
Foul	Any infringement that is punished as foul play by a referee
Yellow Card	A player is shown a yellow card by the referee
Red Card	A player is sanctioned a red card by the referee
Pass Blocks	Number of blocked passes completed
Shots Blocks	Number of blocked shots completed

The research process is outlined in Fig. 1. First, match data from three categories—*Variables related to goal scoring*, *Variables related to passing and organizing*, and *Variables related to defending*—were collected from the CSL and CFACL. After data collection, One-way ANOVA was applied for preprocessing and cleaning to identify key variables exhibiting significant differences ($p < 0.05$) in their association with match outcomes. Following data preprocessing, the XGBoost model was applied for outcome prediction, with SHAP integrated to quantify feature-level contributions to model interpretability. Finally, an analysis of the model results was conducted to examine the impact of these metrics on match outcomes

also surpassed drawing teams and losing teams in Passes ($p<0.001$, ES = 0.022), and ATPA ($p<0.001$, ES = 0.015), with similar trends observed in Key Passes ($p<0.001$, ES = 0.01), ATP ($p<0.001$, ES = 0.01), Break Throws ($p<0.001$, ES = 0.01). In addition, in CSL, winning teams committed fewer Crosses ($p<0.001$, ES = 0.009), Free Kicks ($p<0.001$, ES = 0.008), Front Free Kicks (($p<0.001$, ES = 0.005). Similarly, CFACL winning teams committed fewer Crosses ($p<0.001$, ES = 0.01), Free Kicks ($p<0.001$, ES = 0.016), Front Free Kicks (($p<0.001$, ES = 0.012).

Regarding *Variables related to defending*, winning teams in CSL displayed significantly higher values in Clearances ($p<0.001$, ES = 0.023), Tackles ($p<0.001$, ES = 0.013), Red Card ($p<0.001$, ES = 0.013), Pass Blocks ($p<0.001$, ES = 0.007) and Interceptions ($p<0.001$, ES = 0.007) compared to drawing teams and losing teams. Similarly, CFACL winning teams showed better performance in Clearances ($p<0.001$, ES = 0.016), Foul ($p<0.001$, ES = 0.015), and Shots Blocked ($p<0.001$, ES = 0.004) than losing teams. Furthermore, in CFACL, drawing teams received more Red Cards ($p<0.001$, ES = 0.003), indicating lower discipline compared to winning teams and losing teams. Similarly, CSL drawing teams showed a higher incidence of Red Cards ($p<0.001$, ES = 0.013).

Figure 2 is a typical SHAP Summary plot to explain the results of XGBoost, which displays the impact of various features on the model's prediction outcomes. This chart includes two sections, displaying the feature impacts in both CSL and the CFACL. The X-axis (SHAPE Value) represents the extent to which each feature affects the model's predictions. The Y-axis (feature) represents all the features for each league, ranked by importance. The color represents different values of the feature, with yellow indicating lower feature values and purple indicating higher feature values.

In the CSL, SOTIB (v = 0.272) showed the highest SHAP value, identifying them as the primary predictor for match outcomes. Other features such as Clearances (v = 0.145), Key Passes(v = 0.096), and ATPA (v = 0.073) also have important effects on match outcomes. Clearances, in particular, indicate a strong defensive contribution, where increased clearance frequency shifts the model's predictions positively. Meanwhile, Cross (v = 0.069), Red Card (v = 0.031), Foul (v = 0.027), and PSR(v = 0.027) exhibit highly levels of importance, their SHAP value on the match outcomes vary. For instance, PSR exert a stronger positive influence on the outcome compared to Red Card. Moreover, while Tackles (v = 0.012), Free Kicks (v = 0.012), Front Free Kicks (v = 0.01), Shots Blocks (v = 0.008), Yellow Card (v = 0.007), and SOTOB (v = 0.007) rank lower on the Y-axis in terms of importance, the significance of the research is also important. Notably, for disciplinary metrics, Red Card

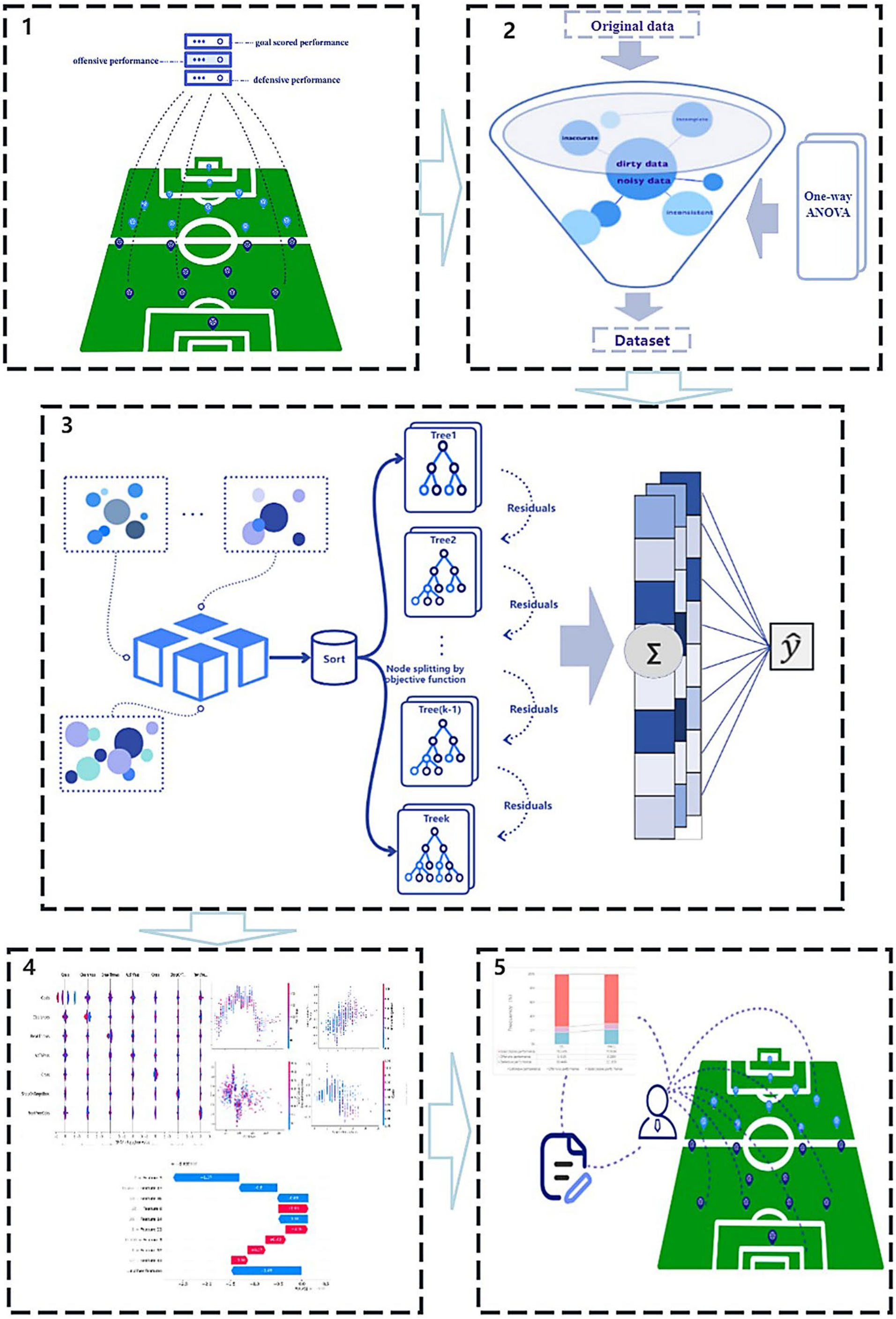

Fig. 1 Procedure of research

and Yellow Card have opposing SHAP values, suggesting that these variables may have contrasting effects on match outcomes under different conditions. Similarly, in the CFACL, SOTOB (v = 0.174) remain the most influential feature affecting match outcomes. Additionally, defensive performance metrics such as Clearances (v = 0.106), Foul (v = 0.054), and as Interceptions (v = 0.013) have a more significant impact compared to their rankings in the CSL, suggesting that defense plays a crucial role in this secondary league. Specifically, the greater influence of Foul in the CFACL, compared to the CSL, indicates that Foul may be a more critical factor in determining match winners in this league.

Table 2 Descriptive statistics of match performance indicators among (mean ± SD), One-way ANOVA, P, ES

	CSL						CFACL					
	Mean ± SD			F	p	ES	Mean ± SD			t	p	ES
	Win	Lose	Draw				Win	Lose	Draw			
Shots	9.67 ± 4.87	8.26 ± 4.79	7.57 ± 4.47	73.926	< 0.001	0.033	9.23 ± 4.69	7.68 ± 4.63	7.02 ± 4.27	90.239	< 0.001	0.04
ShotsOnTarget	4.21 ± 2.37	2.44 ± 2.00	2.45 ± 1.92	339.601	< 0.001	0.136	4.03 ± 2.20	2.28 ± 1.92	2.21 ± 1.75	391.173	< 0.001	0.154
ShotsOnTargetObox	1.10 ± 1.19	0.89 ± 1.05	0.79 ± 0.99	30.736	< 0.001	0.014	1.15 ± 1.11	0.84 ± 1.00	0.78 ± 0.92	52.233	< 0.001	0.024
ShotsOnTargetIbox	3.11 ± 1.87	1.56 ± 1.46	1.67 ± 1.50	419.172	< 0.001	0.163	2.88 ± 1.82	1.44 ± 1.44	1.43 ± 1.36	414.86	< 0.001	0.161
ShotsOutPenaltyArea	3.48 ± 2.42	3.65 ± 2.47	3.17 ± 2.30	14.251	< 0.001	0.007	3.68 ± 2.45	3.49 ± 2.43	3.14 ± 2.32	18.588	< 0.001	0.009
ShotsInPenaltyArea	6.19 ± 3.42	4.61 ± 3.18	4.39 ± 3.08	133.195	< 0.001	0.058	5.56 ± 3.22	4.19 ± 3.02	3.88 ± 2.84	124.125	< 0.001	0.054
Penalties	0.20 ± 0.43	0.08 ± 0.29	0.09 ± 0.31	49.507	< 0.001	0.022	0.18 ± 0.40	0.09 ± 0.30	0.09 ± 0.28	33.599	< 0.001	0.015
Pass	265.02 ± 119.64	260.95 ± 116.40	235.62 ± 101.73	27.735	< 0.001	0.013	248.66 ± 112.66	256.34 ± 114.31	219.36 ± 96.00	48.775	< 0.001	0.022
KeyPasses	5.83 ± 3.56	5.78 ± 3.58	5.14 ± 3.38	16.901	< 0.001	0.008	5.51 ± 3.48	5.21 ± 3.42	4.69 ± 3.19	22.464	< 0.001	0.01
PassesSuccRate	0.76 ± 0.07	0.75 ± 0.07	0.75 ± 0.07	14.869	< 0.001	0.007	0.75 ± 0.08	0.74 ± 0.07	0.73 ± 0.09	13.881	< 0.001	0.006
AttThPass	64.32 ± 34.39	63.43 ± 35.00	56.15 ± 31.54	24.667	< 0.001	0.011	56.10 ± 30.54	55.38 ± 30.50	49.59 ± 28.15	21.328	< 0.001	0.01
AttThSucPaAccuracy	0.69 ± 0.08	0.66 ± 0.08	0.66 ± 0.09	53.986	< 0.001	0.024	0.66 ± 0.09	0.64 ± 0.10	0.64 ± 0.10	32.068	< 0.001	0.015
Cross	10.91 ± 6.19	12.17 ± 7.30	10.73 ± 6.53	19.755	< 0.001	0.009	10.65 ± 6.21	11.66 ± 7.29	10.10 ± 6.37	20.69	< 0.001	0.01
Freekicks	11.11 ± 5.44	11.73 ± 5.55	10.53 ± 5.04	17.835	< 0.001	0.008	11.91 ± 5.83	12.91 ± 6.09	11.15 ± 5.25	34.721	< 0.001	0.016
FrontFreeKicks	4.58 ± 2.86	4.88 ± 2.87	4.38 ± 2.72	11.103	< 0.001	0.005	4.77 ± 2.94	5.08 ± 3.04	4.32 ± 2.65	26.652	< 0.001	0.012
Corners	3.47 ± 2.52	3.42 ± 2.53	3.06 ± 2.32	11.235	< 0.001	0.005	3.22 ± 2.43	3.18 ± 2.43	2.80 ± 2.31	13.607	< 0.001	0.006
BreakThrows	8.85 ± 5.90	8.18 ± 5.54	7.58 ± 5.35	18.262	< 0.001	0.008	7.34 ± 5.30	6.58 ± 4.75	6.15 ± 4.55	22.217	< 0.001	0.01
PassBlocks	6.56 ± 3.74	6.14 ± 3.57	5.82 ± 3.46	15.082	< 0.001	0.007	5.89 ± 3.53	5.57 ± 3.45	5.12 ± 3.24	18.664	< 0.001	0.009
ShotsBlocks	2.06 ± 1.89	1.95 ± 1.68	1.79 ± 1.64	8.641	< 0.001	0.004	1.81 ± 1.73	1.72 ± 1.56	1.57 ± 1.55	8.585	< 0.001	0.004
Clearances	14.39 ± 8.62	11.82 ± 6.84	11.95 ± 7.53	51.687	< 0.001	0.023	14.15 ± 8.88	12.02 ± 7.24	11.91 ± 7.89	35.372	< 0.001	0.016
Intercepttions	7.86 ± 4.90	7.39 ± 4.75	6.87 ± 4.17	16.082	< 0.001	0.007	7.38 ± 4.92	7.02 ± 4.75	6.46 ± 4.25	14.657	< 0.001	0.007
Tackles	11.37 ± 5.66	11.13 ± 5.55	9.94 ± 4.91	28.289	< 0.001	0.013	10.12 ± 5.35	10.21 ± 5.36	8.95 ± 4.84	27.179	< 0.001	0.012
YellowCard	1.34 ± 1.24	1.28 ± 1.14	1.22 ± 1.12	3.212	< 0.05	0.001	1.31 ± 1.20	1.21 ± 1.14	1.13 ± 1.09	9.189	< 0.001	0.004
RedCard	0.04 ± 0.19	0.11 ± 0.33	0.07 ± 0.27	28.369	< 0.001	0.013	0.05 ± 0.22	0.09 ± 0.29	0.06 ± 0.25	7.195	= 0.001	0.003
Foul	10.46 ± 5.02	10.18 ± 5.11	9.48 ± 4.65	14.785	< 0.001	0.007	11.62 ± 5.61	10.79 ± 5.30	10.02 ± 4.94	33.461	< 0.001	0.015

Abbreviations: CSL, Chinese Super Football League; CFACL, Chinese Football Association China League; SOT, shots on target; SOPA, shots out penalty area; SOTOB, shots on target outside box; SIPA, shots inside penalty area; SOTIB, shots on target inside box; ATP, Attacking Third Passes; ATPA, Attacking Third Successful Pass Accuracy; PSR, Passes Successful Rate

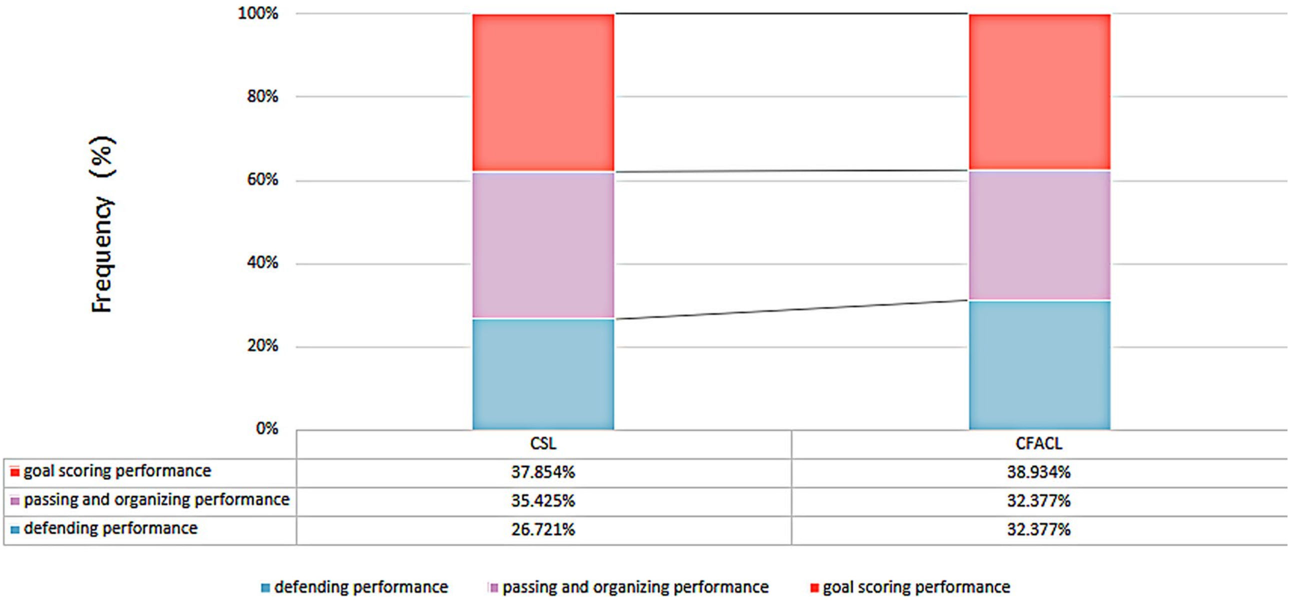

Fig. 2 SHAP Summary Plot. Abbreviations: 1st league (CSL), Chinese Super Football League; 2nd league (CFACL), Chinese Football Association China League; ATPA, attacking third successful pass accuracy; SOTIB, shots on target inside box; BT, Breakthrough; SOT, shots on target; SOTOB, shots in target outside box; SIPA, shots inside penalty area

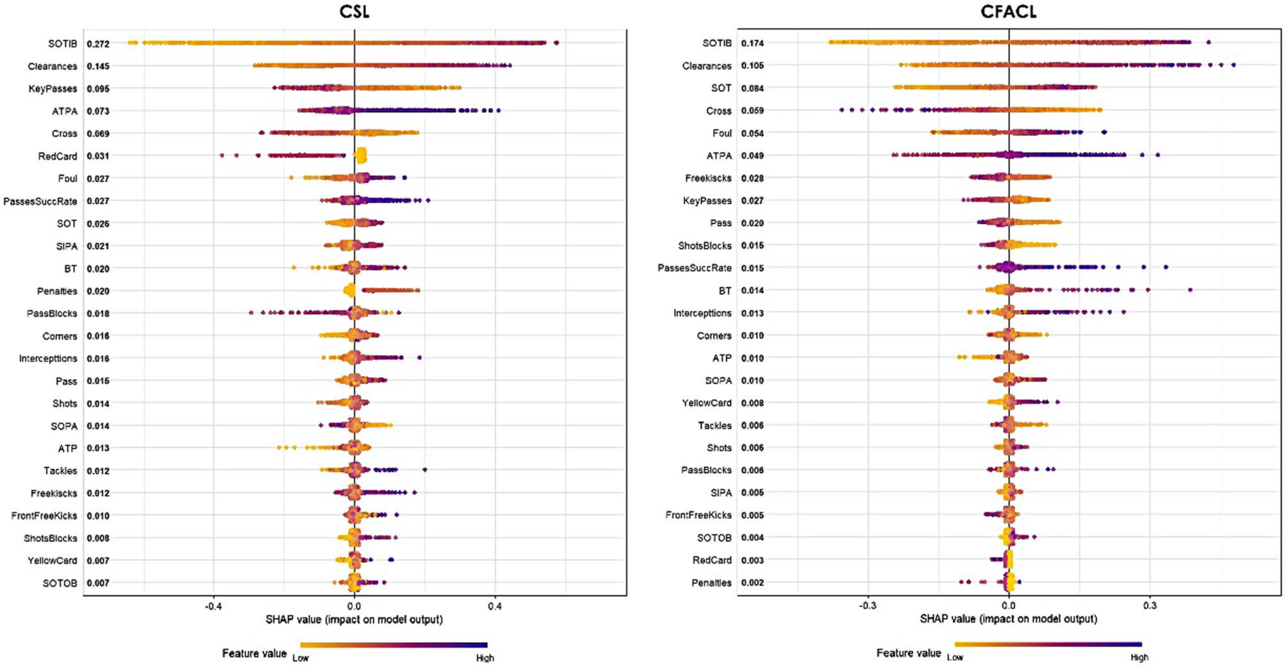

Fig. 3 Group proportion. Abbreviations: CSL, Chinese Super Football League; CFACL, Chinese Football Association China League

Finally, this study, utilizing SHAP values and the group influence model, reveals that the values of S_G、 S_p and S_D in the context of CSL are 37.854%, 35.425%, and 26.721%, In contrast, within CFACL, the corresponding values of S_G, S_p and S_D are 38.934%, 32.377% and 32.377%, respectively. Figure 3 provides a clearer representation of the results.

Discussion

The purpose of this study was to explore the key factors that differentiate winning teams between the CSL and CFACL. While previous research has examined a range of discriminants related to match success [8, 29], few studies have focused on the comparative analysis of winning factors between different divisions, particularly within the context of China's two-tier league. Utilizing

XGBoost and SHAP, we uncovered significant differences in the impact of various features on match outcome predictions across the two divisions leagues.

This research explores the perspective of indicator groups. Through the data obtained from the constructed indicator group model based on SHAP and XGBoost, it is found that, first of all, the score indicators have a significant impact on match outcomes in both CSL (S_G =37.854%) and CFACL (S_G=38.934%). This finding is consistent with previous research [12, 30–32], highlighting that scoring indicators remain the most critical factors for winning across different soccer leagues. Additionally, although the influence of other indicators on match results is relatively minor, a comparison between the two leagues shows that defensive indicators in CFACL (S_D=32.377%) perform better than those in CSL (S_D=26.721%). This may be due to several factors, including the strength disparities in lower leagues that prompt coaches to emphasize defensive strategies against stronger opponents [33], the slower match pace that highlights the importance of defense [34], and the differing playing styles in lower-tier competitions where teams, with fewer top players, rely more on coordinated team defense to secure victories [35].

Goal scoring performance

Additionally, previous studies have further identified that the most critical variables influencing match victories include ball possession, shots on target, and successful passes [36]. A comparison of the top ten key performance indicators impacting match outcomes in this study supports this perspective. The results reveal that metrics such as SOT and ATPA rank highly in both CSL and CFACL matches. However, it is noteworthy that in the CSL, the importance of SIPA's ranking is moderate, whereas in the CFACL, it ranks near the bottom. This discrepancy suggests that, although variables such as total shots play significant roles in both leagues, for the CSL, the number of shots within the penalty area appears to be more crucial for securing victory in matches. This discrepancy may stem from variations in shot quality between the leagues [37], as CSL players generally exhibit higher shot accuracy, leading to more goals from shots inside the penalty area, which consequently has a stronger impact on match outcomes.

Additionally, the study found that SOTIB ranks first in importance among all variables related to goal scoring in both the CSL and CFACL. Research on the German First Bundesliga suggests that shot quality is more important than shot quantity for winning matches [38]. This research also supports this conclusion, indicating that whether in the CSL or CFACL, players need to have precise shots within the penalty area rather than merely increasing the number of shots. This highlights that in top-tier leagues, match outcomes are more influenced by the quality of shots, as shots from different positions are affected by factors such as technical and tactical performance, defensive organization, and coaching strategies [39]. As the level of competition increases in top-tier leagues, the higher standard of defensive organization makes shots within the box more critical and threatening. This finding suggests a potential need for CFACL teams to focus on enhancing shot quality in their training programs. This observation aligns with studies on the first La Liga and the FIFA World Cup [12, 32], which also emphasize the significant impact of shots inside the box on match victories.

Passing and organizing performance

The studies of the UEFA Champions League [40] and the Spanish Professional Football League [1] have found that variables such as ball possession, passes, successful passes, and crosses have a significant impact on match outcomes. Although this research only selected two offensive-related variables, ATPA and Crosses, when identifying data relevant to match results, the SHAP analysis revealed that both variables had high importance in CSL and CFACL. Further analysis of the 2014 FIFA World Cup in Brazil [32]emphasized that, for close games in the same leagues, long passes and crosses—aerial duels—are not always effective. The results of this research also indicate that, although crosses ranked highly in terms of their impact on match outcomes in both the CSL and CFACL, their effect was primarily negative, especially in CSL matches. Therefore, when developing match strategies, coaches should be more cautious in implementing crossing tactics, taking into account the opponent and the league level. This indicates that for CFACL teams, if they are unable to achieve higher-quality passing, they might still be able to organize effective attacks through set-pieces such as free kicks.

Defending performance

Defensive performance plays a decisive role in soccer matches. This study also found some points of interest regarding the defensive metrics of the two-tier leagues. First, in the CSL, the defensive feature Clearances ranked second only to SOTIB in terms of variable importance. This result is consistent with previous research [41], indicating that clearances, as a defensive action, play a pivotal role in maintaining defensive stability and mitigating high-intensity offensive plays in top-tier competitions. Meanwhile, in the CFACL, clearances were equally critical, ranking first among all defensive metrics. This suggests that despite differences in team strength, technical and tactical performance, and match context between leagues of different levels [37], consistent Clearance continues to exert a significant influence on match outcomes in high-level competitions. Furthermore, the research results show that the importance of Clearances ranks below SOTIB in both levels of the league, indirectly indicating that SOTIB consistently have a direct impact on match outcomes across all league levels. Clearances serve as a defensive response to opponents' shots or attacks, primarily aimed at preventing goals through technical handling [31, 32]. While defensive actions can effectively avert crises, they do not directly influence the score. By contrast, SOTIB is a direct determinant of match outcomes. Representing the conversion of offensive plays such as shooting into tangible scores, which are immediately quantified as match points and ultimately exert a decisive influence on the outcome [42]. Consequently, the importance of the clearance metric ranks consistently after SOTIB across both league levels. Additionally, by comparing defensive metrics between the CSL and CFACL, the study found that although the overall importance of defensive metrics is relatively lower in CFACL, other defensive metrics such as Foul and Interceptions have a more significant impact on match outcomes than in CSL games, with the importance of Foul being particularly higher than that of corresponding metrics in CSL matches. This finding aligns with previous studies that have explored the impact of Foul on match outcomes [12, 32, 42]. For example, research on the first and second divisions of LaLiga [12]has similarly indicated that foul play a crucial role in the second divisions of LaLiga. This phenomenon is influenced not only by the physical performance of players and the tactical strategies employed by coaches but also by the match context [12]. Therefore, for CFACL teams, if they are unable to organize an effective defense in a timely manner, employing tactical fouls in appropriate areas might be a crucial factor in reducing the likelihood of losing. As the level of competition increases, referees tend to adopt stricter standards in the assessment of fouls, with a more diversified foul detection system, thereby reducing the frequency of fouls. Supporting this, studies on the 2014 FIFA World Cup have shown that compared to 2006 and 2010, referees applied stricter foul enforcement, leading to a noticeable decrease in serious fouls and injuries [43].

This study acknowledges several limitations. For example, contextual variables such as home vs. away status, team strength, opponent quality, and match dynamics were not considered. Furthermore, the analysis lacked control for covariates. Future research should incorporate these factors to enable a more nuanced exploration of determinants of success across professional football leagues at different competitive tiers.

Conclusion

This study explores the factors influencing match outcomes in the CSL and CFACL, utilizing XGBoost and SHAP to reveal the varying importance of these discriminants across two divisions. The results show that while Shots On Target Inside Box are consistently the key determinant of match outcomes, other factors exhibit significant differences between the two leagues, particularly Penalties. Defensively, Clearances are crucial in both leagues, and Shots Blocks has a more pronounced impact in the CFACL, potentially due to contextual differences. Therefore, in daily training and matches, coaches and players should prioritize improving shot quality and accuracy within the penalty area over merely increasing shot frequency. For CFACL teams, capitalizing on set-piece opportunities (e.g., free kicks and corners) through dedicated training is critical for securing victories. Additionally, adopting strategic defensive measures—such as tactical fouling in key zones and implementing shot-limiting tactics—can significantly enhance the likelihood of success for CFACL teams. These findings offer valuable insights for coaches in tailoring strategies to different league contexts and emphasize the need for tactical adjustments based on league characteristics and opponent conditions. Future research could further explore the influence of home and away factors, and individual player performances.

Abbreviations

XGBoost	eXtreme Gradient Boosting
SHAP	Shapley Additive explanation
CSL	Chinese Super Football League
CFACL	Chinese Football Association China League
SOT	shots on target
SOPA	shots out penalty area
SOTOB	shots on target outside box
SIPA	shots inside penalty area
SOTIB	shots on target inside box
ATP	Attacking Third Passes
ATPA	Attacking Third Successful Pass Accuracy
PSR	Passes Successful Rate

Acknowledgements
The authors would like to thank those who participated in the study.

Author contributions

Jiaxuan Zhu, Bo Yuan and Tianbiao Liu conceptualized this study and wrote the original draft preparation, Pengyu Pan and Honglin Song contributed to the methodology and data visualization, Jiaxuan Zhu managed the dataset, Bo Yuan, Tianbiao Liu, Dingmeng Ren and Zheng Liang reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding
The author(s) reported there is no funding associated with the work featured in this article.

Data availability

The data that support the findings of this study are available from Champdas Football Big Data Company (http:// www. champ das. com), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Champdas

Football Big Data Company. If someone wants to study further with the data, please contact the corresponding author.

Declarations

Ethics approval and consent to participate
The study involves Chinese Super Soccer League (CSL) matches, which are public matches, to analyze match performance—The research project did not involve human participants, human experiments, and human data. The match data are permitted to make scientific research by the Champdas Football Big Data Company. Authors confirm that all methods were carried out in accordance with relevant guidelines and regulations. This study was conducted according to the ethical principles of the World Medical Association Declaration of Helsinki and approved by College of physical education and sports, Beijing Normal University.

Competing interests
The authors declare no competing interests.

Consent for publication
Not applicable.

Received: 22 January 2025 / Accepted: 24 March 2025
Published online: 21 April 2025

References

1. Lago-Peñas C, Lago-Ballesteros J, Dellal A. Gómez MJJoss, medicine. Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. 2010;9(2):288.
2. Thomas V, Reilly TJBJSM. Changes in fitness profiles during a season of track and field training and competition. 1976;10(4):217.
3. Jones P, James N, Mellalieu SDJIJPAS. Possession as a performance indicator in soccer. 2004;4(1):98–102.
4. Hughes M, Franks IJJ. Analysis of passing sequences, shots and goals in soccer. 2005;23(5):509–14.
5. Brito Souza D, López-Del Campo R, Blanco-Pita H, Resta R. Del Coso JJIJopais. A new paradigm to understand success in professional football: analysis of match statistics in LaLiga for 8 complete seasons. 2019;19(4):543–55.
6. Kubayi A, Toriola AJJH. Match performance indicators that discriminated between winning, drawing and losing teams in the 2017. AFCON Soccer Championship. 2020;72:215.
7. Liu T, Garcia de Alcaraz A, Zhang L, Zhang Y. Exploring home advantage and quality of opposition interactions in the Chinese football super league. Int J Perform Anal Sport. 2019;19:1–13.
8. Ma X, Li X, Quan T, Liu H, Liu TJPotIoME, Part P. Journal of Sports Engineering, Technology. The influence of technical performance indicators on the results of the Chinese Football Super League at different stages of the season-based on evidence from the 2010–2019 seasons. 2023:17543371231170065.
9. Andrzejewski M, Oliva-Lozano JM, Chmura P, Chmura J, Czarniecki S, Kowalczuk E et al. Analysis of team success based on match technical and running performance in a professional soccer league. 2022;14(1):82.
10. Lepschy H, Wäsche H, Woll AJIJPAS. Success Factors Football: Anal German Bundesliga. 2020;20(2):150–64.
11. Oliva-Lozano JM, Fortes V, Muyor JMJRSM. When and how do elite soccer players sprint in match play? A longitudinal study in a professional soccer league. 2023;31(1):1–12.
12. Oliva-Lozano JM, Martínez-Puertas H, Fortes V, López-Del Campo R, Resta R, Muyor JMJBS. Is there any relationship between match running, technical-tactical performance, and team success in professional soccer? A longitudinal study in the first and second divisions of LaLiga. 2023;40(2):587–94.
13. Errekagorri I, López del Campo R, Resta R, Castellano JJS. Performance analysis of the Spanish Men's top and second professional football division teams during eight consecutive seasons. 2023;23(22):9115.
14. Sarmento H, Pereira A, Matos N, Campaniço J, Anguera TM, Leitão JJIJPAS. different? English premier league, Spaińs La Liga and Italýs seriés a–What's 2013;13(3):773–89.
15. Zhao Y, Liu TJFP. Factors that influence actual playing time: evidence from the Chinese super league and english premier league. 2022;13:907336.
16. Malamatinos M-C, Vrochidou E, Papakostas GAJC. On predicting soccer outcomes in the Greek league using machine learning. 2022;11(9):133.
17. Song H, Li Y, Zou X, Hu P, Liu TJSR. Elite male table tennis matches diagnosis using SHAP and a hybrid LSTM–BPNN algorithm. 2023;13(1):11533.
18. Moustakidis S, Plakias S, Kokkotis C, Tsatalas T, Tsaopoulos D. Predicting football team performance with explainable AI: leveraging SHAP to identify key team-Level performance metrics. 2023;15(5):174.
19. Plakias S, Kokkotis C, Mitrotasios M, Armatas V, Tsatalas T, Giakas G. Identifying key factors for Securing a champions league position in French Ligue 1 using explainable machine learning techniques. 2024;14(18):8375.
20. Abasi A, Nazari A, Moezy A, Fatemi Aghda SA. Machine learning models for reinjury risk prediction using cardiopulmonary exercise testing (CPET) data: optimizing athlete recovery. BioData Min. 2025;18(1):16.
21. Rohde M, Breuer CJIJFS. Europe's elite football: Financial growth, sporting success, transfer investment, and private majority investors. 2016;4(2):12.
22. Gong B, Cui Y, Gai Y, Yi Q, Gómez M-Á. The Validity and Reliability of Live Football Match Statistics From Champdas Master Match Analysis System. 2019;10.
23. Comfort P, Thomas C, Dos' Santos T, Suchomel TJ, Jones PA, McMahon JJJS. Changes in dynamic strength index in response to strength training. 2018;6(4):176.
24. Liu H, Miguel-Ángel G, Carlos L-P, Sampaio J. Match statistics related to winning in the group stage of 2014 Brazil FIFA world cup. J Sports Sci. 2015;33(12):1205–13.
25. Pan P, Li F, Han B, Yuan B, Liu T. Exploring the impact of professional soccer substitute players on physical and technical performance. BMC Sports Sci Med Rehabilitation. 2023;15(1).
26. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
27. Hu M, Zhang H, Wu B, Li G, Zhou LJSR, SHAP. Interpretable predictive model for shield attitude control performance based on XGboost and. 2022;12(1):18226.
28. Scott M. Su-In LJAinips. A unified approach to interpreting model predictions. 2017;30:4765-74.
29. Han B, Yang L, Pan P, García-de-Alcaraz A, Yang C, Liu TJBSS, Medicine, et al. Influence Removing Home Advant Chin Footb Super Leag. 2022;14(1):208.
30. Lago-Ballesteros J, Lago-Peñas CJJH. Performance in team sports: identifying the keys to success in soccer. 2010;25(1):85–91.
31. Wright C, Atkins S, Polman R, Jones B, Sargeson LJIJPAS. Factors associated with goals and goal scoring opportunities in professional soccer. 2011;11(3):438–49.
32. Liu H, Gomez M-Á, Lago-Peñas C, Sampaio JJJ. Match statistics related to winning in the group stage of 2014 Brazil FIFA world cup. 2015;33(12):1205–13.
33. Delgado Bordonau JL, Domenech Monforte C, Guzmán Luján JF, Méndez Villanueva A. Offensive and defensive team performance: relation to successful and unsuccessful participation in the 2010 Soccer World Cup. 2013.
34. Wallace JL, Norton KIJJS. Sport Mi. Evolution of world cup soccer final games 1966–2010: game structure. Speed Play Patterns. 2014;17(2):223–8.
35. Hewitt A, Greenham G, Norton KJIJPAS. Game style in soccer: what is it and can we quantify it? 2016;16(1):355– 72.
36. Lepschy H, Wäsche H, Woll AJT. How to be successful in football: a systematic review. 2018;11(1).
37. Liu H, Gómez M-A, Gonçalves B, Sampaio JJ. Technical performance and match-to-match variation in elite football teams. Joss. 2016;34(6):509–18.
38. Yue Z, Broich H, Mester JJIJoSS. Coaching. Statistical analysis for the soccer matches. First Bundesliga. 2014;9(3):553–60.
39. Schulze E, Mendes B, Maurício N, Furtado B, Cesário N, Carriço S et al. Effects of positional variables on shooting outcome in elite football. 2018;2(2):93–100.
40. Lago-Peñas C, Lago-Ballesteros J, Rey EJJ. Differences in performance indicators between winning and losing teams in the UEFA champions league. 2011;27(1):135–46.
41. Zenbaba E. Technical performance of Ethiopian male soccer National team. Turkish J Sport Exerc. 2018;20(2):116–21.

42. Freitas R, Volossovitch A, Almeida CH, Vleck VJGJE, Research S. Elite-level defensive performance in football: a systematic review. 2023;53(4):458–70.
43. Junge A, Dvořák JJB. Football injuries during the 2014. FIFA World Cup. 2015;49(9):599–602.

Publisher's note