English Premier League Data Analysis: Insights and Recommendations

Xinyao Li*

Beijing Normal University (BNU) and Hong Kong Baptist University (HKBU), United International College (UIC), Zhuhai, Guangdong, 519087, China

*Corresponding author: 1469460334@qq.com

Abstract. This research explores the tactical and statistical patterns of team performance in the English Premier League (EPL). The study focuses on key metrics such as Expected Goals (xG), Expected Goals Against (xGA), and pressing intensities (PPDA and OPPDA) to understand their impact on team success. Analyzing data from Manchester City (MCI) and Liverpool (LIV) for the 2020-2021 and 2021-2022 seasons, the study identifies critical factors that differentiate top-performing teams. The results underscore the role of strategic adjustments in enhancing both offensive and defensive capabilities in the EPL.

Keywords: Expected Goals; Expected Goals Against; Pressing Intensity; Opponent Pressing Intensity; Team Performance; English Premier League; Tactical Strategy; Manchester City; Liverpool; Football Analytics.

1 Introduction

As an enthusiast of football, my leisure time is frequently spent engaging in matches with friends and following the sport. And this passion has inspired the initiation of this project, which entails a comprehensive analysis of data from the English Premier League (EPL) for the 2020-2021 and 2021-2022 seasons. The objective is to apply the knowledge acquired from my academic pursuits to conduct a series of analyses, aiming to distill critical insights from the data. These insights are intended to enhance our understanding of team and player performances [1] [2] (Razali et al., 2017; Muszaidi et al., 2022). In this technologically advanced era, it is important to utilize machine learning algorithms effectively to facilitate improvements in strategies. Through our analysis, we aim to uncover factors contributing to certain teams' success and others' underperformance, thereby offering data-driven recommendations for future enhancements. In terms of the data we will be using, there is a variety of teams from the English Premier League: Arsenal (ARS), Aston Villa (AVL), Bournemouth (BOU), Brentford (BRE), Brighton (BHA), Burnley (BUR), Chelsea (CHE), Crystal Palace (CRY), Everton (EVE), Fulham (FUL), Leeds (LEE), Leicester (LEI), Liverpool (LIV), Luton (LUT), Manchester City (MCI), Manchester United (MUN), Newcastle United (NEW), Norwich (NOR), Nottingham Forest (NFO), Sheffield United (SHU), Southampton

A. K. Draman Mud et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Social Sciences (ICBDSS 2024)*, Advances in Computer Science Research 116,
https://doi.org/10.2991/978-94-6463-562-1_5

(SOU), Tottenham (TOT), Watford (WAT), West Bromwich Albion (WBA), West Ham (WHU), and Wolverhampton Wanderers (WOL). In terms of the variable included in the data, we had wide range of variables, but there are some we are particular interested. Firstly, Player and Team variables are foundational, providing the name of players and their respective teams. The Role variable offers insights into the positional play, which, when combined with Cost and Selection (Sel.), can reveal the perceived value and popularity of players in fantasy football realms. Performance metrics such as Goals.scored, Assists, Clean.sheets, and Goals.conceded directly reflect on-field contributions, crucial for evaluating player effectiveness. Player_xG (Expected Goals) and Player_xA (Expected Assists) are advanced metrics providing a deeper understanding of a player's offensive potential beyond traditional statistics. Team-level variables like Team_xG (Team Expected Goals) and Team_xGA (Team Expected Goals Against) offer a macro perspective, useful for assessing team strategies and overall strength.[3] Team_PPDA (Passes Per Defensive Action) and Team_OPPDA (Opponent PPDA) could provide an understanding of the team's defensive and pressing style. Our goal is to learn more about the patterns of successful teams so that, in light of their achievements, we can advise other teams.

2 Analysis

2.1 Expected Goals Analysis

Our analysis is on the team's expected goals (xG) for every week of play, with a focus on cumulative xG data, which will give us a complete picture of a team's performance throughout a season, increasing or decreasing after each game week. This analysis is essential for determining if a team performs consistently and effectively over an extended period of time. From the Figure 1 and Figure 2 provided, it is evident that teams like Manchester City (MCI) and Liverpool (LIV) have outperformed others in both Season 20-21 and Season 21-22. A notable observation is Manchester City's dramatic increase in xG after game week 22, which markedly sets them apart from other teams. This trend highlights their offensive strength and strategic effectiveness in the latter half of the season, underscoring their dominance in the league.

To provide a more detailed numeric view of the team's performance consistency across the 2020-2021 and 2021-2022 seasons, **Table 1** highlights the key metrics of Expected Goals (xG) and Expected Goals Against (xGA) for Manchester City (MCI), which will provide a granular view into the team's performance consistency, highlighting any notable improvements during different stages of each season, as well as shedding light on the efficacy of their offensive and defensive strategies.

In the 20-21 season (S21), there was a consistent uptick in both Team_xG and Team_xGA as the season unfolded. This trend aligns with expectations, considering that accumulating games naturally lead to increased totals. However, the 21-22 season (S22) presents a distinct narrative: Team_xG surged more dramatically, signaling a significantly bolstered attack that outstripped the prior season's performance. Concurrently, Team_xGA was notably reduced across almost all corresponding gameweeks, pointing to a fortified defense. The augmented xG values in the 21-22 season are indicative of Manchester City's offensive enhancement, while the diminished xGA values across the same period underscore their defensive advancements. Such data implies that Manchester City likely implemented effective tactical adjustments after the 20-21 season, which paid dividends in amplifying their offensive prowess and defensive robustness. This strategic evolution is a testament to the team's ability not only to create and capitalize on scoring opportunities but also to adeptly stymie their opponents' attacks.

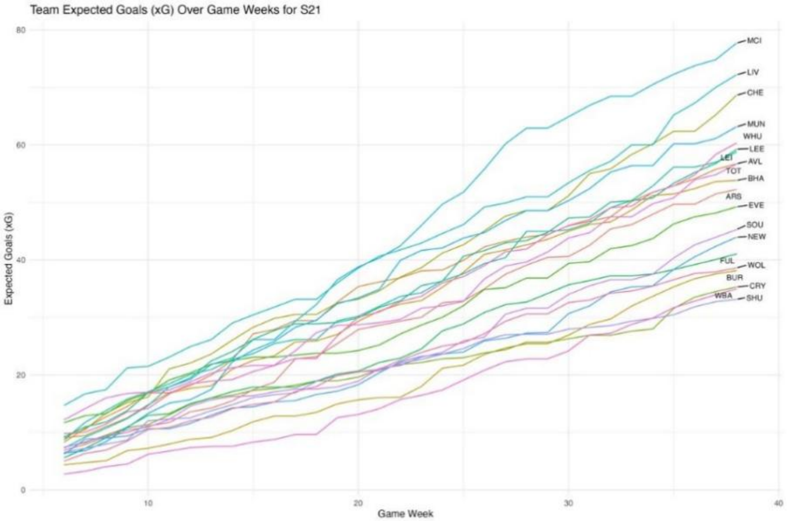

Fig. 1. xG over game weeks in the season 20-21

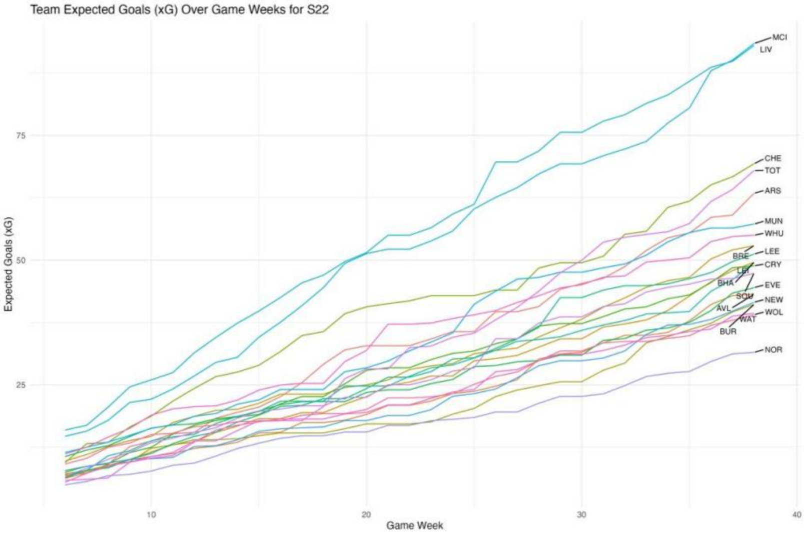

Fig. 2. xG over game weeks in the season 21-22

Table 1. xG and xGA over two seasons for MCI

S21	MCI		S22	MCI
Game Week	Team_xG	Team_xGA	Team_xG	Team_xGA
6	7.26	7.26	14.69	2.69
7	9.16	7.95	15.73	3.64
7	10.74	9.14	17.89	4.71
7	12.41	9.9	21.46	5.88
7	14.67	10.32	22.12	6.93
7	17.93	10.61	24.16	7.69
7	19.21	11.2	26.77	7.96
7	21.8	11.41	29.49	8.13
7	22.96	12	30.51	8.9
7	26.15	12.2	34.54	9.58
7	26.15	12.2	37.54	9.66
7	29.33	12.89	40.86	9.88
7	32.24	13.07	44.5	10.06
7	36.55	13.72	49.35	12.73
7	38.79	14.29	51.3	13.57
7	40.33	14.44	52.2	14.18
7	42.45	14.49	52.2	14.18
7	46	15.67	53.83	15.18
7	49.69	17.14	55.81	15.64
7	51.78	17.47	60.26	16
7	55.8	19.72	62.57	18
7	60.19	23.44	64.51	18.77
7	62.91	23.89	67.25	19.14
7	62.91	23.89	69.29	19.61
7	64.9	24.17	69.29	19.61
7	66.88	24.35	70.91	19.67
7	68.45	25.15	72.26	20.97
7	68.45	25.15	73.79	21.09
7	70.49	25.51	77.47	21.51
7	72.28	26.13	80.48	22.55
7	73.76	28.31	87.93	23.96
7	74.83	29.54	90.08	24.96
7	77.72	30.61	93.4	25.21

2.2 Player Contributions

Analyzing the individual contributions of players during the 2020-2021 and 2021-2022 seasons, **Table 2** lists the top goal scorers for Manchester City. Kevin De Bruyne, Raheem Sterling, and Riyad Mahrez were prominent in the 2020-2021 season, while İlkay Gündoğan led the team's attack in the following season. From the **Table 2**, we can see that in the 20-21 season (S21), Kevin De Bruyne led the team with an impressive 15 goals, showcasing his vital role in the team's attacking prowess. Raheem Sterling followed with a solid 12 goals, and Riyad Mahrez contributed significantly with 10 goals, rounding out a formidable attacking trio. However, in the following 21-22 season (S22), İlkay Gündoğan emerged as the top scorer with 13 goals, indicating his increased influence and perhaps a shift in the team's offensive dynamics. Sterling and Mahrez, while still among the top contributors, saw a decrease in their total goals to 8 each.

Table 2. Top goals scored by players

S21		S22	
Player	Total Goals	Player	Total Goals
Kevin De Bruyne	15	llkay Gü¼ndogan	13
Raheem Shaquille Sterling	12	Raheem Shaquille Sterling	8
Riyad Karim Mahrez	10	Riyad Karim Mahrez	8

Recall that the Figure 1 for Expected Goals (xG) across game weeks revealed that Sheffield United (SHU) consistently registered at the lower end of the spectrum in the 20-21 season, while Norwich (NOR) occupied the bottom position in the 21-22 season from the Figure 2. This consistent underperformance in xG prompted an investigation into the total goals conceded by these teams to understand defensive vulnerabilities. The total goals conceded metric, when attributed to individual players, requires careful interpretation, particularly in relation to their positions on the pitch. Traditionally, the number of goals conceded is a statistic most relevant to defensive positions, including the goalkeeper and defenders, as they are directly involved in preventing the opposition from scoring.

In contrast to the goal-scoring prowess, **Table 3** shows the defensive struggles faced by players from Sheffield United (SHU) and Norwich (NOR) in the 2020-2021 and 2021-2022 seasons, respectively. This table lists the players associated with the highest number of goals conceded, shedding light on the defensive vulnerabilities of these teams. As shown in the **Table 3**, players like Aaron Ramsdale, Enda Stevens, Grant Hanley, and Max Aarons are in positions where their primary responsibilities involve defending their goal. The high number of goals conceded associated with these players could reflect the overall defensive struggles of their respective teams. It could indicate that the team as a whole often found itself under pressure, leading to a higher number of goals being conceded while these particular players were on the field.

Table 3. Top goals conceded by players

S21SHU		S22 NOR	
Player	Total Goals Conceded	Player	Total Goals Conceded
Aaron Ramsdale	54	Teemu Pukki	65
Enda Stevens	45	Grant Hanley	54
John Fleck	43	Max Aarons	53

However, the inclusion of Teemu Pukki, a forward, in the context of goals conceded is unconventional, as forwards are not typically involved in defensive duties. Pukki's high number associated with goals conceded likely does not directly reflect on his personal defensive performance but rather on the team's collective inability to defend effectively. It may imply that during the periods Pukki was on the pitch, Norwich frequently found themselves in a defensive position, perhaps as a result of chasing games where they were trailing, which in turn could lead to conceding more goals.

2.3 Goals vs. Expected Goals Analysis

The Figure 3, illustrating Goals versus Expected Goals (G vs xG), indicates that both teams perform relatively in line with the expected metrics. This alignment suggests that the number of goals scored by each team is consistent with the quality of chances they created. Notably, in the 21-22 season, both teams exhibit instances where the actual goals scored surpass the expected goals. This overperformance may be attributed to exceptional finishing skills or perhaps a favorable sequence of match events. It is particularly evident for Manchester City, whose data points frequently reside above the expected value line, highlighting their clinical efficiency in front of goal.

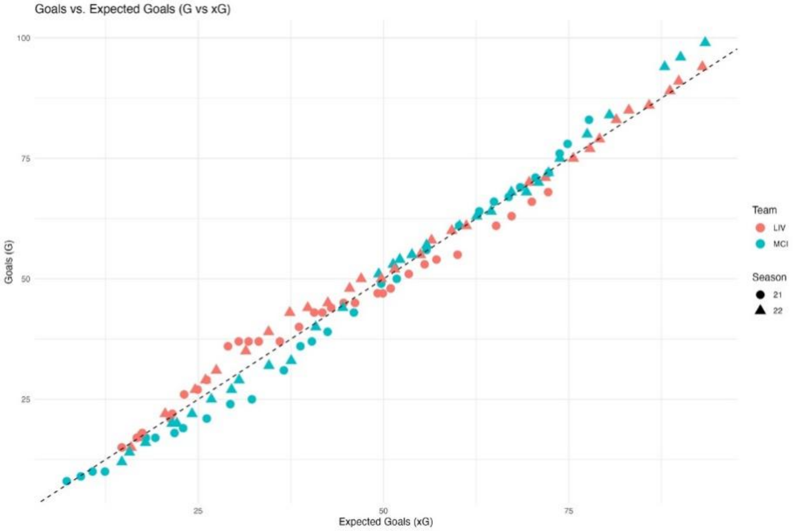

Fig. 3. Goals Scored vs Expected Goals for two seasons and two teams

2.4 Goals Against vs. Expected Goals Against Analysis

The Figure 4, depicting Goals Against versus Expected Goals Against (GA vs xGA), provides an assessment of defensive prowess. Manchester City's data points predominantly lie below the expected value line, especially in the 21-22 season, suggesting a robust defensive record that exceeds expectations based on the quality of chances they conceded. This could reflect a combination of strategic defensive organization and outstanding goalkeeping. The consistency in which both teams maintain points below the line across two seasons could also indicate effective defensive strategies or standout performances by defensive players and goalkeepers.

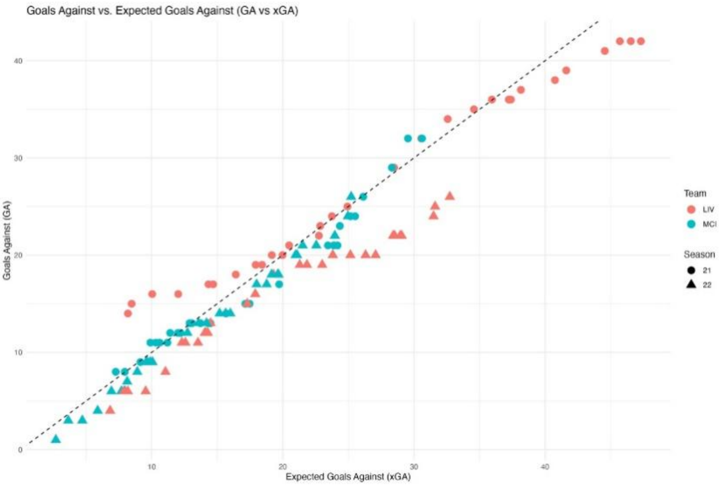

Fig. 4. Goals Against vs Expected Goals Against for two seasons and two teams

2.5 PPDA and OPPDA Analysis

The Figure 5 indicates how aggressively each team presses their opponents, with lower PPDA values signifying more intense pressing. In this case, both Manchester City (MCI) and Liverpool (LIV) appear to maintain a consistent and relatively aggressive pressing strategy across both seasons, as indicated by their lower median PPDA values compared to many other teams. For MCI, there is a noticeable consistency in their PPDA values across both seasons, with a tight interquartile range (IQR) suggesting a disciplined approach to pressing. LIV, while also showcasing a propensity for pressing, displays a slightly higher IQR in the 21-22 season, which could suggest slight variability in their pressing game by game.

The OPPDA value measures how many passes the opposing team is allowed before a defensive action is taken, with lower values indicating more aggressive pressing by a team. We observe that Manchester City (MCI) and Liverpool (LIV) are showing much higher values in the Figure 6 compared to other teams, this indicates that their opponents are able to complete more passes before MCI or LIV make a defensive action. MCI and LIV may be employing a tactical approach that involves sitting deeper and allowing the opposition to have the ball in less threatening areas. This could be a deliberate strategy to maintain defensive shape and stability, inviting opponents forward to create space behind them for counter-attacks.

This tactical profile is consistent with the previous analysis, which highlighted that MCI and LIV have much higher xG and goals scored compared to other teams. Their ability to press and regain possession aggressively creates disruptive attacking opportunities, often catching opponents off-balance and leading to high-quality chances, as reflected in their xG metrics. [4]

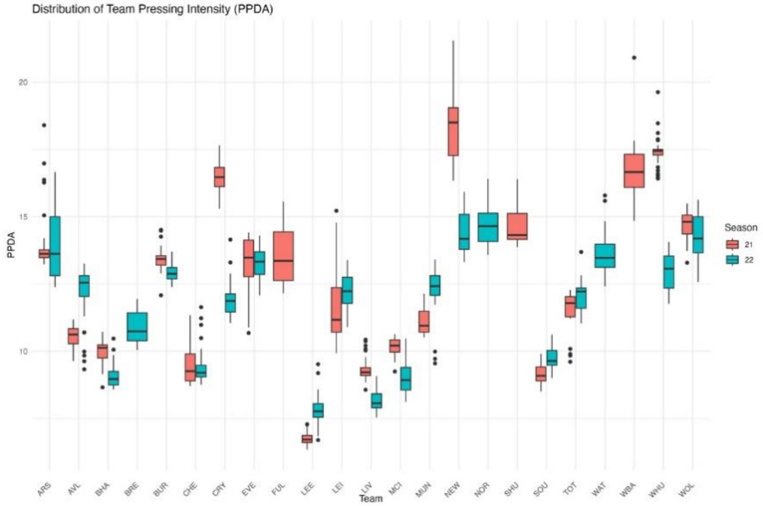

Fig. 5. Team-specific PPDA Distribution Over Seasons

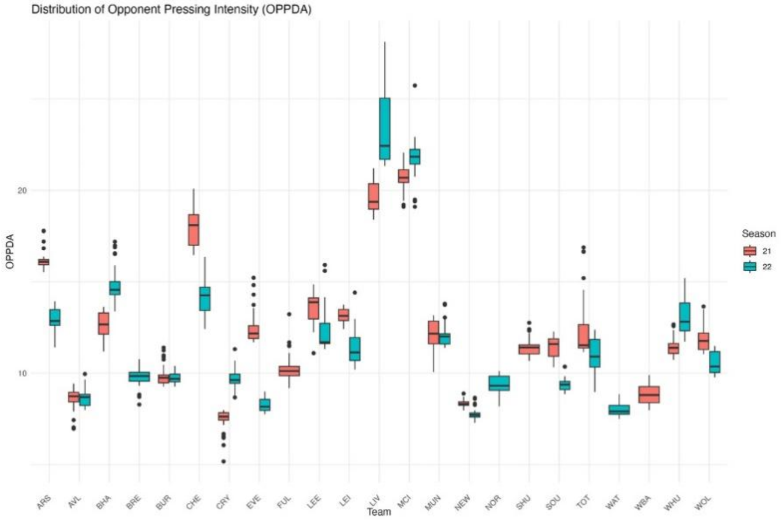

Fig. 6. Team-specific OPPDA Distribution Over Seasons

2.6 Regression Analysis

We would like to also investigate which factor are more important for predicting the Goals scored and Goals conceded, we can build a multivariate regression model, where the dependent variable is the number of goals scored by a team (Team_G) and the independent variables are Team Pressing Intensity (Team_PPDA), Opponent Pressing Intensity (Team_OPPDA), Dangerous Chances created (Team_DC), and Opponent Dangerous Chances (Team_ODC).

Table 4. Regression Results for Goals Scored Based on Pressing Intensity and Dangerous Chances.

	Dependent variable:
	Weekly_G
Weekly_PPDA	0.017
	p = 0.860
Weekly_OPPDA	-0.083
	p = 0.387
Weekly_DC	0.131***
	p = 0.000
Weekly_ODC	0.016**
	p = 0.017
Constant	0.337***
	p = 0.00000
Observations	1,320
R2	0.242
Adjusted R2	0.240
Residual Std. Error	1.246 (df = 1315)
F Statistic	104.993*** (df = 4; 1315)

Note: *p<0.1; **p<0.05; ***p<0.01.

As is shown in **Table 4**, the regression analysis yields insightful findings about the factors influencing a team's goal-scoring prowess. Most notably, the variables representing Dangerous Chances (DC) and Opponent Dangerous Chances (ODC) emerge as statistically significant predictors, distinguished by their low p-values at the 0.05 significant threshold. The coefficient for DC is 0.131, which suggests a strong and direct relationship between the creation of dangerous chances by the team and the number of goals scored. In practical terms, it indicates that for every additional dangerous chance created by the team, there is an expected increase of 0.131 in the number of goals scored. This aligns with intuitive understanding of football tactics, where creating high-quality scoring opportunities is paramount to offensive success. On the other hand, ODC carries a coefficient of 0.016, which, while lower than that of DC, still signifies a positive relationship with goal scoring. This could imply that in games where opponents also create a significant number of dangerous chances, there is a corresponding increase in the team's goal-scoring. This might be reflective of a particular style of play where matches are open and both teams engage in attack-oriented football, leading to more goals overall. Other predictors in the model do not reach statistical significance, suggesting that within the context of this analysis, they do not have a discernible impact on the number of goals scored, at least not in a linear and additive manner as assumed by the model. The model’s R-squared value of 0.242 indicates that approximately

24.2% of the variance in the team's goal-scoring is explained by the model. While this captures a significant portion of the variability, it also suggests that there are other factors, not included in the model, that account for the majority of the variability in goal scoring. These could include aspects such as individual player skills, team dynamics, in-game strategies, or even external factors like weather conditions or player fitness levels.

We would also like to know what are the important factor for the Goals Conceded.

Table 5. Regression Results for Goals Conceded Based on Pressing Intensity and Defensive Performance

	Dependent variable:
	Weekly_GA
Weekly_PPDA	0.115
	p = 0.230
Weekly_OPPDA	0.179*
	p = 0.063
Weekly_DC	0.012*
	p = 0.083
Weekly_ODC	0.125***
	p = 0.000
Constant	0.411***
	p = 0.000
Observations	1,320
R2	0.233
Adjusted R2	0.231
Residual Std. Error	1.245 (df = 1315)
F Statistic	99.892*** (df = 4; 1315)

Note: *p<0.1; **p<0.05; ***p<0.01

From the regression result for goals conceded shown in **Table 5**, the Opponent Dangerous Chances (ODC) is statistically significant with a coefficient of 0.125 and a p-value well below the 0.05 threshold, it is clear that an increase in the dangerous chances conceded by a team is strongly associated with an increase in the number of goals they concede. This finding underscores the critical importance of maintaining a robust defense that can effectively limit high-quality scoring opportunities for the opposition. Furthermore, the model's R-squared value, standing at 0.233, indicates that approximately 23.3% of the variability in weekly goals conceded is explained by the variables included in the model, which means approximately 76.7% of this variability is influenced by factors not included in the model. These could encompass a range of elements such as individual player errors, variations in opponent quality, specific game situations like set pieces or counter-attacks, or even external factors like weather conditions and player fitness.

3 Conclusion and Discussion

This study aim to detect underlying pattern in the English Premier League (EPL) dataset, in order to accomplish this, we have done a series analysis. First of all, we accessed the team performances in relation to metrics like Expected Goals Against (xGA) and Expected Goals (xG). After doing a thorough analysis, we have found that Liverpool (LIV) and Manchester City (MCI) have both performed exceptionally well and consistently during the 2020–2021 and 2021–2022 seasons. Besides, we included the important scorers in MCI and LIV whose efforts have been essential to these teams' outstanding results. Moreover, we also included clubs at the lower end of the xG spectrum in our research, and we listed players whose defensive errors led to more goals given up. In comparison to the top-performing teams, this aspect of the analysis offered a striking contrast. We focused on the subtleties of both offensive and defensive play, analyzing Opponent Passes per Defensive Action (OPPDA) and Passes per Defensive Action (PPDA) analytics to identify the tactical components that support MCI and LIV's success. Our results show that both teams demonstrate a strategic balance with a high OPPDA and a comparatively low PPDA. This suggests a two-pronged strategy—assertive pressing to reclaim possession and tactical allowing the opposition to pass—that many of their opponents appear to have missed. To cap off our investigation, we conducted regression analysis to determine the variables that significantly impact goals scored and conceded. The results of this analysis were telling; Defensive Challenges (DC) and Offensive Defensive Challenges (ODC) emerged as statistically significant predictors for goals scored. In parallel, ODC stood out as a crucial factor in the equation for goals conceded.

Our analysis reveals some important strategies, especially from top teams like Manchester City and Liverpool. These teams have a special way of playing: they press their opponents aggressively but also let them have the ball at times. This strategy, where they use low PPDA and high OPPDA, has been very successful. Other teams could learn from this and try similar tactics. Another key finding is how important DC and ODC are. These are moments in the game where defending or attacking actions can really make a difference. Teams that want to score more goals or let in fewer goals should focus on these moments in training. For example, teams that struggle to score (low xG) might need better attackers or need to train their players to be more effective in front of the goal. On the other hand, teams that concede a lot of goals should work on their defense. [5]

References

1. Razali, N., Mustapha, A., Yatim, F. A., & Ab Aziz, R. (2017). Predicting football matches results using Bayesian networks for English Premier League (EPL). In *Iop conference series: Materials science and engineering* (Vol. 226, No. 1, p. 012099). IOP Publishing.
2. Muszaidi, M., Mustapha, A. B., Ismail, S., & Razali, N. (2022). Deep Learning Approach for football match classification of English Premier League (EPL) based on full-time results. In *Proceedings of the 7th International Conference on the Applications of Science and Mathematics 2021: Sciemathic 2021* (pp. 339-350). Singapore: Springer Nature Singapore.

3. Kulikova L. I., Goshunova A. V. (2013). Measuring efficiency of professional football club in contemporary researches. *World Applied Sciences Journal*, 25(2), 247–257.
4. Espitia-Escuer M., Garcia-Cebrián L. I. (2020). Efficiency of football teams from an organisation management perspective. *Managerial and Decision Economics*, 41(3), 321–338.
5. Espitia-Escuer M., García-Cebrián L. I. (2016). Productivity and competitiveness: The case of football teams playing in the UEFA Champions League. *Athens Journal of Sport*, 3(1), 57–85.