Football analytics for better betting: Pitch partitioning, possession sequences, expected goal model and player evaluation on Dawson model

Aladár Kollár

Budapest University of Technology and Economics

Abstract

One of the most significant developments in the sports world over the last two decades has been the use of mathematical methods in conjunction with the massive amounts of data now available to analyze performances, identify trends and patterns, and forecast results. Football analytics has advanced significantly in recent years and continues to evolve as it becomes a more recognized and integral part of the game. Football analytics is also used to forecast game outcomes, allowing bettors to make educated guesses. This article describes mathematical concepts related to football analytics that enable a better betting strategies. We explain how the pitch is partitioned into different zones and we define possession sequences. Furthermore, we explain what an expected goals model is and which expected goals model we use in this research. Furthermore, we define two general characteristics of a player evaluation method, each corresponding to one of the equations of the Dawson model. Based on these characteristics, we describe the developments of several general approaches for evaluating players in the context of the Dawson model.

Keywords: *Betting, Dawson model, Football, xG, Pitch partitioning, possession sequences, expected goal model and player evaluation*

Correspondence:

Aladár Kollár

Budapest University of Technology and Economics
Twitter, Linkedin
Author at MightyTips.hu, Sportfogadás.
Betting Tipster at MightyTips.com

Introduction

The impact of data analytics has grown in every facet of our lives over the last two decades: in businesses of all sizes [1], but also in healthcare, media, and sports. Football, until a few years ago, was believed to be immune to this trend [2]. Now, new users in the major football leagues are prospering as a result of the competitive advantage that data analytics investments are beginning to provide: Liverpool, AZ Alkmaar, and Brentford are just a few of the rapidly growing list of successful case studies. Clubs that do not plan to dive on the analytics bandwagon, in some opinions, risk being left behind.

To be more precise, football analytics is the process of developing actionable insights and decisions based on football-related data. The data can range from how many goals a team has scored to various factors such as the distance covered by a single player in a team, the number of passes he has made and how many of those were misplaced, and how many of those created chances for the team to score. Football makes use of both descriptive and predictive analytics. While predictive analytics forecasts the likelihood of an event occurring, descriptive analytics analyzes the information at hand and makes recommendations to increase the likelihood even further [3]. As a result, without good data, analytics is nearly worthless.

The game of football is heavily reliant on data analytics. Current systems are used to forecast game outcomes, which is useful for people betting on their favorite teams. Individuals can make an educated guess as to who will win a match, increasing their chances of earning money. Additionally, data analytics is used to extract hidden information from a game [4]. This enables team managers, coaches, and players to gain a better understanding of their own game, their own mistakes, and their opponents' strategies, weaknesses, and so on. With all of this information about their games, teams can perform significantly better.

Football analytics, in its most primitive form, is typically credited to Charles Reep [5], an Englishman, war veteran, and die-hard football fan [6]. According to legend, during one game in 1950, Reep became enraged by a team's pitiful scoring attempts and began keeping a notebook of observations and trends he noticed . He noted the high ratio of attacks to goals in that game, concluding that even a slight increase in scoring efficiency would result in an additional goal per game. Reep devised a system for recording spatial information for each play, scribbling quickly during the match and allegedly having to spend 80 hours after each game interpreting his shorthand.

Reep became football's first known analytics staff when he was hired on a part-time basis by the Brentford team, which avoided relegation by winning 13 of its 14 matches and doubling its goal total. He would deliver his message to various teams, allegedly trying to instill a culture of attack that was backed up by data [7]□[8]□While he ceased trying to persuade teams in the late 1950s following the collapse of one of his teams, Reep continued to make observations and calculations. His legacy was primarily built on encouraging clubs to cut down on passing and emphasizing the importance of using fewer long passes to propel the ball downfield—a concept based on his observation that the majority of goals are scored on plays involving fewer than four passes.

With the advent of companies utilizing technology and enhanced computing to analyze play on the pitch, analytics would become more complex and prevalent [9]□On the sidelines, the onlooker furiously scribbling player locations and goal shots was replaced by a more technologically advanced system. A fascinating advancement in football analytics has been the establishment of rival companies, each of which uses a unique tracking mechanism or measures distinct actions to inform team decisions. However, the public — and even prospective teams debating which company's number crunching to use — were unaware of the distinctions between these firms [10]□

Opta, for example, was established in 1996 and quickly partnered with the English Premier League to develop infrastructure for providing match data insight. In 1999, the company began collecting real-time data and making it available to individual players in 2000. Five years later, the company might well streamline its tracking innovation in order to provide subscribers with real-time position tracking information.

Data alone is insufficient to achieve a competitive advantage; even more critical is the capacity to interpret it. This is becoming increasingly obvious in the modern era, as databases continue to expand and the responsibility of data scientists becomes more critical [11]□

Rather than embracing the positive aspects of Reep's work (data collection) and focusing on how the data is interpreted, the football establishment dismissed the experiment and its underlying concept, namely that using data, it is possible to improve one's understanding of the game and gain a competitive edge [12]□

The problem of data interpretation persists. For instance, if football clubs are inundated with massive streams of numbers but lack the internal expertise to interpret them and retrieve actionable insights, data becomes nearly useless [13].□It's analogous to giving a person who is unfamiliar with financial markets all the prices, ratios, and indicators for stocks, currencies, and commodities: the data alone does not make that person an authoritative trader. Football teams require data to make sound decisions, but they also require analytics to make sense of it.

In a sports environment, analytics has a variety of on-field applications, including trying to manage both individual and team performance Coaches might use data to optimize their players' exercise programs and nutrition plans in order to maximize fitness [14].□ Additionally, analytics are frequently used to develop tactics and team strategies. With thousands of games of data to analyze, experts can look for patterns in formation, counter strategies, and other critical variables across a large sample size [15].□

.With increasing access to unfathomable amounts of data and technology, an increasing number of teams across a variety of sports have begun utilizing analytics to their advantage. Different analytics techniques, most notably predictive analytics, have risen to prominence in recent years, particularly when it comes to predicting player and team performance and managing teams based on analytics. “Because the majority of professional sports teams operate as businesses, they are constantly looking for ways to increase revenue and cut costs across the organization [16].□ Certain sports analysts concentrate exclusively on the distribution and sale of sports tickets and team merchandise. Additionally, modern marketing and fan outreach efforts heavily rely on analytics to forecast consumer behavior and recognize opportunities to boost brand engagement.

Pitch partitioning

We described in detail how the normalized (x, y)-values of the actions are converted to realistic (x, y)-coordinates on a standard-sized pitch in the data description. Multiple stages of this research project necessitate the partitioning of the pitch [17].□The pitch is divided into a number of zones of equal size. In most of the research projects, only partitionings of equal-sized zones are considered, as this simplifies the interpretation and application of the results [18]□[19]□.Due to the pitch's length of 105 meters and width of 68 meters, the zones have a length of 105 meters and a width of 68 meters. Let xstep denote the zone's length and ystep denote the zone's width. The total number of zones in this case is equal to n = 105/xstep times 68/ystep.

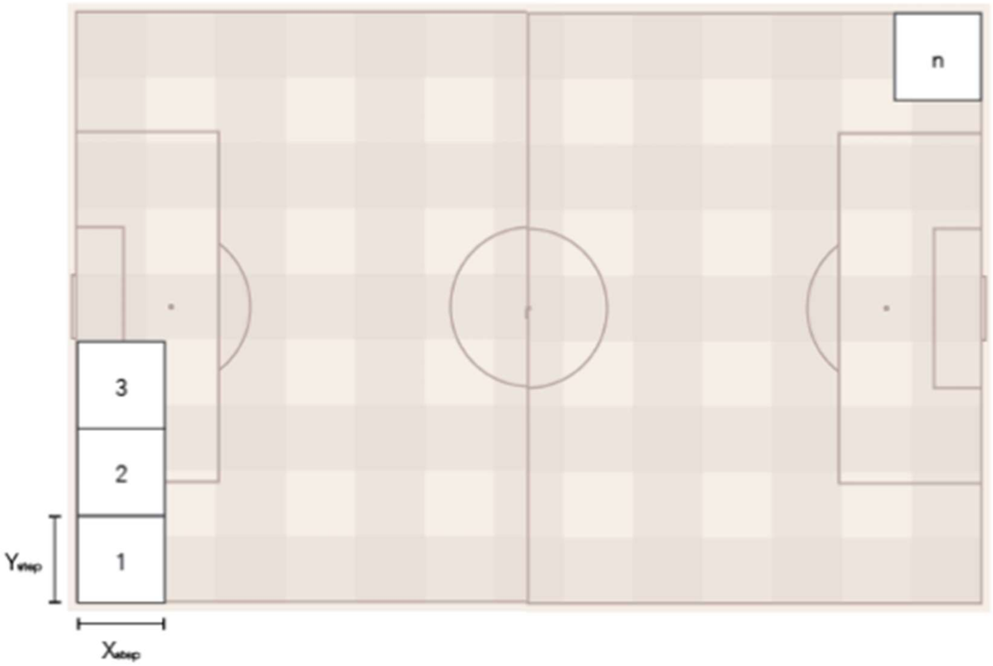


Figure 1: Pitch partitioning: the pitch is divided into n equal-sized zones, which are sequentially numbered from bottom to top, left to right.

We determine the corresponding zone number i for each action using the coordinates (x, y) using below equations:

$$x_{adj} = \lfloor \frac{x}{x_{step}} \rfloor \cdot x_{step}$$

$$y_{adj} = \lfloor \frac{y}{y_{step}} \rfloor \cdot y_{step}$$

$$i = \frac{68}{y_{step}} \cdot \frac{x_{adj}}{x_{step}} + \frac{y_{adj}}{y_{step}}$$

This approach significantly improves the speed with which the corresponding zone for a huge number of points on the pitch can be determined [20].

Possession sequences

A football match can be thought of as a series of sequential actions such as passes, shots, the ball leaving play, throw-ins, fouls, and free kicks. Ai = [a_1,..., a_{Ni}] denotes the action sequence for match I where Ni denotes the total number of actions in match i. This match's action sequence can now be divided into possession sequences [21]□ [22].□ Possession sequences begin and end with the occurrence of one of the following:

- The start or end of a match's period (first half, second half, overtime)
- The ball is removed from play
- The ball is touched by the opposing team (1 touch is enough)
- There has been a goal scored.

Thus, the match I can be defined as a sequence of possession sequences Sj, j = 1,..., Mi, Ai = [S1,..., SMi], where Mi denotes the match i's number of possession sequences. Additionally, the possession sequence Sj is denoted by [akj+1,..., akj+lj], where lj denotes the number of actions in the possession sequence Sj and kj denotes the number of actions in the possession sequence Sj [23]□[24].□

We assign a value uj to each of these possession sequences to represent the sequence's outcome. Possession sequences have one of the following outcomes:

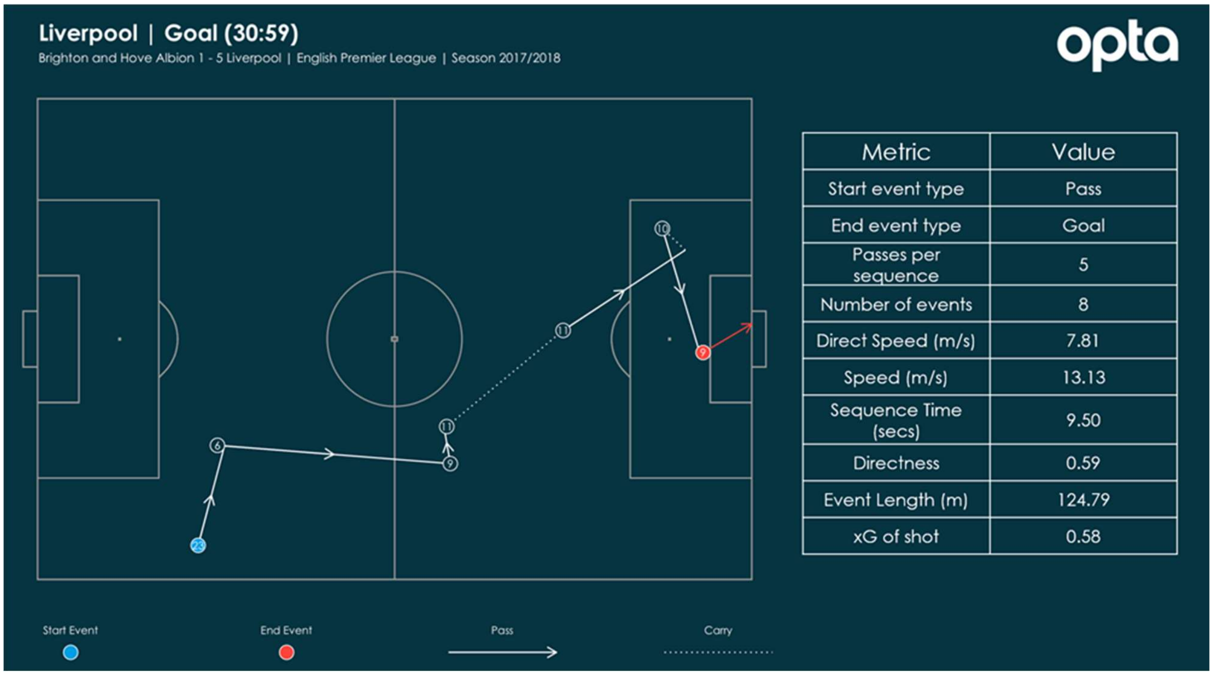
a) Possession is lost

b) A period of play has ended

c) A goal attempt is made (whether successful or unsuccessful)

d) A foul is committed and a free kick is awarded

e) The ball is taken out of play by an opponent and the team that had possession is awarded a corner kick or a throw-in.


Liverpool | Goal (30:59)
Brighton and Hove Albion 1 - 5 Liverpool | English Premier League | Season 2017/2018
opta
23
6
9
11
11
10
9
Start Event
End Event
Pass
Carry

Metric	Value
Start event type	Pass
End event type	Goal
Passes per sequence	5
Number of events	8
Direct Speed (m/s)	7.81
Speed (m/s)	13.13
Sequence Time (secs)	9.50
Directness	0.59
Event Length (m)	124.79
xG of shot	0.58

Figure 2 illustrates Roberto Firmino's goal against Brighton in the 2017/18 Premier League season is depicted in the sequence map below. The sequence's framework reveals the distance traveled, the number of passes involved, and the tempo of the build-up to the eventual shot and resulting goal.

The map illustrates how the sequence leading up to Firmino's goal covered a significant portion of the pitch, beginning deep in Liverpool's own half and ending with just two direct passes into Brighton's half. This demonstrates the Liverpool team's counter-attacking, direct style of play behind Firmino's goal.

When a sequence loses possession, we assign it a value of zero. Sequences that terminate due to the conclusion of a match period are not considered. Additionally, when the ball is removed from play or a foul is committed, the outcome of this sequence is zero, as the sequence has concluded and a new one has begun. When a sequence results in a free kick, corner kick, or throw-in, the sequence's value is also zero. A goal attempt is the most valuable outcome of a possession sequence. Studies value goal attempts using Caley's Expected Goals metric, which is a frequently used metric in the world of football analytics [22]. This metric considers a variety of factors in order to determine the probability that a particular scoring attempt will result in a goal. In the following section, we will discuss the expected goals model in greater detail.

Expected goals model

Sam Green (2012) proposed a unique answer to the difficulty of realistically modeling the game in his essay "Assessing the performance of Premier League goalscorers [25]." Green devised an analytical model based on goalscoring possibilities from specific goalscoring situations called anticipated goals (also known as xG). The purpose of this model was to attribute a probability to the conversion of a goalscoring chance into a goal. If a goal scoring opportunity is given an xG value of 0,3 for example, the effort is converted into a goal 30% of the time with average finishing. We may derive an outcome that indicates how a football match should have ended with average finishing by summing up all the probabilities from the chances made during the game [26].

The distance to the goal and the angle of the shot to the goal are two of the most frequently used parameters in the Expected Goals model. It turns out that these two parameters have a significant effect on the likelihood that a goal attempt from a particular position will result in a goal [27].Some studies, on the other hand, added additional parameters to the model, including the match situation (open play, corner, free kick, or penalty), whether the goal attempt is a rebound and what type of rebound, the attempt type (header, shot from dribble, or shot from pass), and the goal differential between the two teams at the time of the goal attempt [26][1]. Few analyst also takes the match time, league, and body part used to execute the goal attempt into account when calculating the expected goal value of a goal attempt.

To develop a model for valuing football passes, an expected goals model must first be estimated such that all scoring opportunities in the given data can be valued. The passes that resulted in these scoring opportunities can then be valued [3].When only goals are considered, critical information is lost, as players who make flawless passes but have subpar teammates do not receive the credit they deserve. Additionally, when a player makes a very simple pass and his teammate scores from an almost impossible position, the player who assisted the goal receives excessive credit when only goals are considered. The expected goals model only takes into account the location of the goal attempt and uses data from previous seasons to calculate the probability that a goal attempt from a particular position will find the net.

Player evaluation in team sports

Dawson [28] proposed modeling team output W (e.g., wins) as a component of individual performance, L, direct coaching input, C, and other team performance determinants, X:

$$W = f(L, C, X)$$

and the player's performance is determined by his talent T, indirect coaching influence, C, and other variables, Y:

$$L = f(T, C, Y)$$

Despite some vagueness in the notation, this general model tells the summary of the relationship between individual players' skill and the team's match result, namely the fact that the former is reflected in the latter only indirectly through individual performance [12] [18].

Characteristics of player evaluation metrics

Definitions

Consider the following two characteristics of a technique for evaluating players in team sports:

a) The degree to which a relationship exists between a particular aspect of individual performance and team performance as a whole. It is equivalent to the Dawson model's equation.

b) The extent to which the method accounts for the metric's value for a particular player is determined by his skill rather than by factors beyond his control (like the performance of his team mates and opponents or random chance) [29]. It is equivalent to the Dawson model's equation.

It is worth noting that these are features of the evaluation methods themselves, not of the performance on which they are based or of the skill being assessed. To illustrate this, consider ball juggling as an example of a performance that has a negligible effect on the team's outcome in football. A method that acknowledges this weak link between individual ball juggling performance and team performance, for example by granting few points to players who perform it, would have a high characteristic number I [30]. This characteristic refers to the method's ability to establish a link between individual performance and team performance, not to the relationship's strength.

Similarly, regardless of the degree to which the performance itself is dependent on skill, a method for evaluating performance can have a high value for the characteristic number II. It is sufficient that the method makes an attempt to obtain the strength of the relationship between skill and performance rather than trying to treat them as synonymous [31]. Additionally, while one could argue that it is generally desirable to draw a connection between individual and team performance (i.e., to use a method with a high characteristic number I), the choice of a method with a high characteristic number II is more application-dependent. If the analysis's objective is to retrospectively evaluate players' performance, for example, in order to distribute annual awards, there is less reason to acknowledge the randomness inherent in the performance than there is when the performance is used to assess the pure skill that generated it [32].

Player valuation is the primary application for methodologies designed to accomplish the latter. The reason it is critical to value players based on skill rather than past performance is that football clubs should be paying for future performance through wages and transfer fees, and the relationship between past and future performance can be defined more accurately if one recognizes that both involve some inherent skill [33] [34].

Approaches to player evaluation

To put the characteristics of the player evaluation methods discussed in the previous section into context, their four corner situations are listed below and especially in comparison to the Dawson model.

Low I - low II

The most fundamental and historically earliest approach is to evaluate players on the basis of their individual statistics, which are believed to be somewhat predictive of team success. For example, in football, it is the number of goals scored per season and the pass completion rate; in baseball, it is the batting average (the number of hits divided by the number of attempts) [35]. The relationship between individual performance and team output is assumed to exist but is unverified and unquantified. Additionally, these individual statistics are taken at face value without regard for the degree of chance involved in achieving particular values and, consequently, the likelihood that similar values will be obtained in the future. As if only equation were used in isolation and the effect of external variables was ignored, the Dawson model is reduced to:

L = T

High I - Low II

This approach compensates for the above methods' low I score by establishing a link between personal performance and team success [36]. This approach could be expressed as follows in the context of the Dawson model:

$$W = f(L, C, X) \quad \text{and} \quad L = T$$

low I - high II.

This strategy addresses the other disadvantage of low I - low II methods. It acknowledges that factors beyond players' control can have an effect on their performance and explicitly models the effect of covariates and random luck. As a result, a more accurate estimate of a particular skill can be obtained, but its value to the team remains unknown [37]. The Dawson model is reduced to its second equation in this manner:

$$L = f(T, C, Y)$$

High I - high II

Finally, the most comprehensive approach combines the benefits of the previous two approaches by (1) recognizing that talent and performance are not synonymous and (2) connecting individual and team performance [38].

Conclusion

Football has been infiltrated by data in the same way that other aspects of our lives have been infiltrated by data in recent years: intrusively and almost always in contradictory ways. As a result, it is often dismissed in professional situations. As a result, an effort must be made to make sense of the data and to distinguish between what is vital and what is not.

doi:10.20944/preprints202105.0619.v1

Football clubs use data to analyze not only matches, but also individual players. Football teams can determine which players should be traded by analyzing individual players from their own and other teams. Additionally, they are capable of acquiring the best player available from other teams to fill a position on their own team.

This article discusses mathematical concepts relating to football analytics that facilitate the development of more effective betting strategies. We define possession sequences and explain how the pitch is divided into different zones. Additionally, we define what an expected goals model is and the expected goals model that we used in this study. Additionally, we define two broad characteristics of a player evaluation method, each of which corresponds to one of the Dawson model's equations. We describe the evolution of several general approaches for evaluating players in the context of the Dawson model using these characteristics.

References

[1] J. McCullagh, “Data mining in sport: A neural network approach,” *Int. J. Sport. Sci. Eng.*, vol. 4, no. 3, pp. 131–138, 2010.

[2] B. Hutchins, “Tales of the digital sublime: Tracing the relationship between big data and professional sport,” *Convergence*, vol. 22, no. 5, pp. 494–509, 2016.

[3] J. Fernández, L. Bornn, and D. Cervone, “Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer,” 2019.

[4] P. Halvorsen *et al.*, “Bagadus: an integrated system for arena sports analytics: a soccer case study,” in *Proceedings of the 4th ACM Multimedia Systems Conference*, 2013, pp. 48–59.

[5] C. Reep and B. Benjamin, “Skill and chance in association football,” *J. R. Stat. Soc. Ser. A*, vol. 131, no. 4, pp. 581–585, 1968.

[6] R. Pollard, “Charles Reep (1904-2002): pioneer of notational and performance analysis in football,” *J. Sports Sci.*, vol. 20, no. 10, pp. 853–855, 2002.

[7] C. Reep, T. SconnerRoad, S. H. Hillside, M. Lane, and T. Foliot, “THE QUANTITATIVE COMPARISON OF PLAYING STYLES IN SOCCER,” 2011.

[8] O. Larson, “Charles Reep: A major influence on British and Norwegian football,” *Soccer Soc.*, vol. 2, no. 3, pp. 58–78, 2001.

[9] L. Cotta, P. O. V de Melo, F. Benevenuto, and A. A. Loureiro, “Using fifa soccer video game data for soccer analytics,” 2016.

[10] H. Ruiz, P. Power, X. Wei, and P. Lucey, “‘ The Leicester City Fairytale?’ Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 EPL Seasons,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1991–2000.

[11] I. Franks and M. Hughes, *Soccer analytics: successful coaching through match analysis*. Meyer & Meyer Sport, 2016.

[12] G. Liu, Y. Luo, O. Schulte, and T. Kharrat, “Deep soccer analytics: learning an action-value function for evaluating soccer players,” *Data Min. Knowl. Discov.*, vol. 34, no. 5, pp. 1531–1559, 2020.

[13] L. Bornn, D. Cervone, and J. Fernandez, “Soccer analytics: Unravelling the complexity of ‘the beautiful game,’” *Significance*, vol. 15, no. 3, pp. 26–29, 2018.

[14] H. K. Stensland *et al.*, “Bagadus: An integrated real-time system for soccer analytics,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 10, no. 1s, pp. 1–21, 2014.

[15] M. Bacelar, “Monitoring bias and fairness in machine learning models: A review,” *Sci. Prepr.*, 2021.

[16] G. Kumar, “Machine learning for soccer analytics,” *Univ. Leuven*, 2013.

[17] J. L. Herrera-Diestra *et al.*, “Pitch networks reveal organizational and spatial patterns of Guardiola’s FC Barcelona,” *Chaos, Solitons & Fractals*, vol. 138, p. 109934, 2020.

[18] P. Silva, P. Aguiar, R. Duarte, K. Davids, D. Araújo, and J. Garganta, “Effects of pitch size and skill level on tactical behaviours of Association Football players during small-sided and conditioned games,” *Int. J. Sports Sci. Coach.*, vol. 9, no. 5, pp. 993–1006, 2014.

[19] F. Wenk and T. Röfer, “Coordinated pitch observation for a humanoid robot soccer team,” 2011.

[20] R. Maneiro Dios and M. Amatria Jiménez, “Polar coordinate analysis of relationships with teammates, areas of the pitch, and dynamic play in soccer: a study of Xabi Alonso,” *Front. Psychol.*, vol. 9, p. 389, 2018.

[21] M. Kempe, M. Vogelbein, D. Memmert, and S. Nopp, “Possession vs. direct play: evaluating tactical behavior in elite soccer,” *Int. J. Sport. Sci.*, vol. 4, no. 6A, pp. 35–41, 2014.

[22] C. Lago and R. Martín, “Determinants of possession of the ball in soccer,” *J. Sports Sci.*, vol. 25, no. 9, pp. 969–974, 2007.

[23] D. Link and M. Hoernig, “Individual ball possession in soccer,” *PLoS One*, vol. 12, no. 7, p. e0179953, 2017.

[24] M. Merlin, S. A. Cunha, F. A. Moura, R. da S. Torres, B. Gonçalves, and J. Sampaio, “Exploring the determinants of success in different clusters of ball possession sequences in soccer,” *Res. Sport. Med.*, vol. 28, no. 3, pp. 339–350, 2020.

[25] S. Green, “Assessing the Performance of Premier Leauge Goalscorers,” *OptaPro Blog*, 2012.

[26] H. Eggels, R. van Elk, and M. Pechenizkiy, “Expected goals in soccer: Explaining match results using predictive analytics,” in *The machine learning and data mining for sports analytics workshop*, 2016, vol. 16.

[27] A. Rathke, “An examination of expected goals and shot efficiency in soccer,” *J. Hum. Sport Exerc.*, vol. 12, no. 2, pp. 514–529, 2017.

[28] P. Dawson, S. Dobson, and B. Gerrard, “Estimating coaching efficiency in professional team sports: Evidence from English association football,” *Scott. J. Polit. Econ.*, vol. 47, no. 4, pp. 399–421, 2000.

[29] L. Lamas, R. Drezner, G. Otranto, and J. Barrera, “Analytic method for evaluating players’ decisions in team sports: Applications to the soccer goalkeeper,” *PLoS One*, vol. 13, no. 2, p. e0191431, 2018.

[30] B. R. Auer and T. Hiller, “On the evaluation of soccer players: a comparison of a new game-theoretical approach to classic performance measures,” *Appl. Econ. Lett.*, vol. 22, no. 14, pp. 1100–1107, 2015.

[31] M. Manafifard, H. Ebadi, and H. A. Moghaddam, “A survey on player tracking in soccer videos,” *Comput. Vis. Image Underst.*, vol. 159, pp. 19–46, 2017.

[32] L. O. Gavião, A. P. Sant’Anna, G. B. Alves Lima, and P. A. de Almada Garcia, “Evaluation of soccer players under the Moneyball concept,” *J. Sports Sci.*, vol. 38, no. 11–12, pp. 1221–1247, 2020.

[33] J.-P. Renno, J. Orwell, D. Thirde, and G. A. Jones, “Shadow Classification and Evaluation for Soccer Player Detection.,” in *BMVC*, 2004, pp. 1–10.

[34] C. Müller, T. Sterzing, J. Lange, and T. L. Milani, “Comprehensive evaluation of player-surface interaction on artificial soccer turf,” *Sport. Biomech.*, vol. 9, no. 3, pp. 193–205, 2010.

[35] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, and F. Giannotti, “PlayeRank: data-driven performance evaluation and player ranking

in soccer via a machine learning approach," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 5, pp. 1–27, 2019.

[36] J. Lyle, *Sports coaching concepts: A framework for coaches' behaviour*. Routledge, 2005.

[37] C. Anderson and D. Sally, *The numbers game: Why everything you know about soccer is wrong*. Penguin, 2013.

[38] M. Espitia-Escuer and L. I. García-Cebrián, "Performance in sports teams," *Manag. Decis.*, 2006.