ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20251118071140

Original scientific paper
Received: 2025-11-18, Accepted: 2026-01-05

Machine Learning-Based xG forecasting in the Top 5 European Soccer Leagues: A Comparative Analysis

Davronbek Malikov, Jaeho Kim*

Abstract: In the modern soccer world, also referred to as football, analytics has become integral IT serves a role akin to assistant coaches, contributing significantly to the analysis of team and individual player performance during both games and training sessions. Despite the abundance of advanced technologies, there is still a need for concrete measurements in soccer analytics. A prime example is Expected Goals (xG), a widely embraced metric that goes beyond mere scorelines to offer in-depth insights into player and team dynamics, and has proven its value in recent years. The identification of key features from a dataset to predict xG is a critical aspect of machine learning research applied to soccer analytics. However, prior investigations have either overlooked certain crucial features or failed to recognize their significance. This study proposes a novel approach by incorporating features related to coaches, coaching tenure, and tactics, which have the potential to improve predictive accuracy and provide actionable insights for teams and analysts. Using a dataset of 2,917 observations covering the top five European leagues, we applied regression-based machine learning models, employing a preprocessing pipeline and k-fold cross-validation to ensure robust evaluation. The findings reveal that the xG values for the English Premier League (EPL) surpass those of the other four leagues studied, with an average xG of approximately 1.93. This indicates that, on average, teams in the EPL tend to have a higher expected goal count per match than teams in other top European leagues. This superiority can be attributed to various factors, such as the higher average quality of teams and players, tactical nuances, and the overall competitiveness inherent within the EPL compared with the other leagues under scrutiny. In our study, we introduced new features such as coaches, coaching durations (years), and tactics. Incorporating these features enhanced the model performance across *MSE*, *EVS*, and R^2, thereby demonstrating the efficacy of our approach. Lasso and Ridge Regression models achieved improved predictive accuracy, with *EVC* and R^2 reaching up to 96%, while the Decision Tree model showed a nearly 6% reduction in the *MSE*.

Keywords: Expected Goals (xG); football; regression machine learning; Soccer; Top 5 league xG

1 INTRODUCTION

One notable difference setting soccer apart from other sports is the common occurrence of low-scoring games, where each goal is very important and can greatly affect the result of the match [2, 7, 23]. Consequently, the paramount goal of soccer is to successfully convert these opportunities into tangible goals, underscoring the importance of efficient goal conversion in the sport.

The Expected Goals (xG) metric quantifies the likelihood of scoring a goal by analyzing data from similar past shots. Opta is a renowned leader in sports analytics and provides comprehensive and detailed data for various sports [1]. Their extensive historical database, which contains nearly one million shots, offers valuable insights into player performance, team strategy, and match outcomes. By leveraging Opta's data, we can analyze and predict the probability of scoring opportunities, known as xG, with precision and accuracy. The xG values are assigned on a scale of zero to one. A value of zero indicates an impossible chance to score, whereas a value of one suggests that a goal is expected for every given chance [1]. For instance, a chance from the halfway line is less likely to result in a goal than that inside the penalty area. Using xG, numerical values can be assigned to these scenarios. If a chance from inside the box is given an xG of 0.1, this means that a player is expected to score one goal every ten shots in this situation. For example, if a player has 20 chances from inside the box in a single match, their expected goal total would be 20 times 0.1, or 2 xG [1]. The goal results of many matches, including those of the top five European leagues, were predicted. However, a high xG value does not always guarantee a soccer win. Consider a recent Premier League match between West Ham United and Brighton, as shown in Tab. 1. According to the xG statistics, West Ham had a probability of 0.83, whereas Brighton had a probability of 2.32. Logically, with an xG of 2.32, Brighton should have won the game. Surprisingly, however, the match ended in a 0:0 draw, meaning that no goals were scored despite high xG values.

Table 1 Example of xG and actual goals results (EPL match on 02.02.2024)

Team	xG	Actual goals
West Ham	0.83	0
Brighton	2.32	0

In this example, it is evident that the relationship between xG and actual goals is not always direct. Specifically, while a higher xG score indicates a greater likelihood of scoring for a given opportunity, it does not consistently translate to a higher number of goals. Similarly, an xG value of n does not necessarily imply scoring n goals in every instance. For example, consider a scenario in which a player has an xG value of 0.8 for a particular shot. This suggests that, on average, they have an 80 percent chance of scoring at that opportunity. However, in reality, they may fail to convert the chance due to factors such as exceptional goalkeeping, poor finishing, or a last-ditch defensive block. Conversely, a team may have a lower xG value for a match but still manage to score multiple goals. This could occur if they capitalize on fewer but higher-quality chances, or if they benefit from opposition mistakes that lead to goals. The accuracy of soccer xG predictions can be compromised by the selection and identification of key features. Furthermore, in soccer analytics, precise anticipation of goal-scoring opportunities is vital for strategic decision making and effective performance evaluation [15]. By integrating machine learning models for xG prediction, analysts and coaches can not only forecast potential scoring outcomes but also understand which features (player, tactical, and managerial) drive these predictions, providing a framework for explainable and data-driven decision-making in soccer.

However, previous research [17] has encountered limitations in accurately predicting xG owing to the exclusion of meaningful features that significantly influence xG outcomes. For instance, previous studies [17] have primarily focused on player position as the main feature.

However, team tactics, coaches, and coaching duration (years), which are closely related to player position, contribute to the xG prediction in our study. Including these features also enhances the interpretability of the model, as it reveals how managerial and tactical factors influence predicted scoring opportunities. These features are expected to improve xG prediction, because coaching decisions directly influence team strategy, player positioning, and shot selection. Therefore, the research gap lies in the limited incorporation of coaching-related features into the existing xG models. This study addresses this gap by integrating these features to improve prediction accuracy [17].

The primary objective of this study is to address this disparity by developing an advanced xG model that integrates novel and significant features. In light of the gaps identified in prior research, this study aims to develop a comprehensive framework for predicting xG by incorporating both traditional performance indicators and novel contextual features. Specifically, we hypothesize that the inclusion of variables such as coaching factors (e.g., coaching experience and duration) and tactical attributes, in addition to player- and match-level data, would significantly enhance the predictive accuracy of xG models. The overarching objective of this study is to demonstrate that a broader set of explanatory variables leads to more robust and reliable xG predictions, offering deeper insights into player and team performance evaluations. Additionally, our proposal involves comparing xG across the top five European football leagues: the English Premier League (EPL), German Bundesliga (Bundesliga), Spanish La Liga (La Liga), Italian Serie A (Serie A), and French Ligue 1. Such comparative analysis allows clubs and analysts to identify league-specific patterns, informing strategy and resource allocation across different competitive contexts. This new method aims to improve and customize the xG metric to the specific qualities of each league, thereby providing deeper insights into goal-scoring probabilities across various football settings. Our evaluation shows the predicted xGs for the top five leagues, and compares these values with the average actual goals for each league. We also identify the best xG players currently active in these top leagues. Overall, this study not only improves predictive accuracy for xG but also demonstrates how integrating novel features within ML models can support explainable and strategic decision-making in soccer.

The remainder of this paper is organized as follows: a review of related works, a section on data processing that outlines the steps taken to prepare the data for our study, an experiment section detailing our methodology, and a results section presenting the outcomes of our analysis by comparing recent results. Furthermore, the Top 5 xG section introduces a new category of xG designed for the Top Leagues, accompanied by a comparative evaluation. The paper concludes with a final section that summarizes our findings and insights.

2 BACKGROUND AND RELATED WORK

In this section, we present an outline of xG and its advantages for soccer teams, including coaches. In the following section, we integrate the relevant studies that encompass prior research.

2.1 Overview and Expected Goals

In recent years, football has undergone significant changes both on and off the pitch, with new tactics such as counterpressing and zonal marking altering its dynamics [13]. Alongside these tactical evolutions, advanced analytics has become increasingly prevalent, with xG emerging as a key metric for assessing player and team performance. xG provides a more nuanced understanding of how effectively teams convert their scoring opportunities and create high-quality opportunities. Moreover, the increasing use of data, particularly xG, has transformed how football matches are perceived and analyzed. While xG has become a central aspect of tactical discussions, it has also sparked controversy among fans. As mentioned earlier, xG is now an essential part of modern football analysis, influencing coaching decisions, player scouting, and even fan discussions on the game. For example, soccer team coaches consider xG metrics, and using xG statistics, coaches can prepare their teams according to the xG statistics for each player [6]. Let us look at the opinions of one of the coaches of the soccer team: "The xG of the two penalties is huge; one I didn't think was and the other was very soft," argued Newcastle manager Eddie Howe [12].

2.2 Related Work

Although the xG metric is gaining popularity in sports analytics [30], it is a relatively recent notation with a limited number of published papers [2, 26]. This can be partly attributed to the challenges associated with the collection of the experimental data. Privacy concerns make data collection a complex task, and the adoption of modern technologies for live data collection is not universally feasible for every soccer team because of the financial limitations they may face [16, 21]. Consequently, the widespread implementation of xG metrics poses challenges in sports analysis. While prior studies have explored xG computation using various machine learning and statistical methods, most focus predominantly on event and positional features, leaving managerial and tactical factors largely unexamined. This analytical comparison identifies a gap in feature selection that motivates our approach.

In soccer analytics, a closer examination of the historical progression of xG literature brings forth the notable work of the expected goals philosophy by James Tippett as a pioneering contribution [30]. Moreover, Tippett's research not only explains the development of the xG metric but also underscores its practical relevance to football. His study focused on the significant impact of xG on performance assessment and player recruitment (scouting), emphasizing its pivotal role in shaping strategic decisions within the complex domain of soccer analysis. Essentially, the philosophy of expected goals is a foundational piece that

explores the developmental journey of expected goals and their application within the intricacies of football.

Furthermore, in the computation of xG metrics, a combination of machine learning and statistical methods is required. Among these methods, logistic regression provides interpretable coefficients but may oversimplify complex interactions, whereas tree-based models and gradient boosting capture nonlinear relationships with higher predictive power. Neural networks offer flexibility for high-dimensional data but often lack interpretability. This comparison highlights that while predictive accuracy is important, understanding feature contributions remains a challenge. This collection of methodologies underscores the flexibility of choosing models for xG computation, showcasing the various paths available to enhance precision and efficacy in predictive analysis, including applications in player salary prediction [2, 20, 23, 31, 32, 36].

Moreover, the significance of shots as crucial elements in predicting xG metrics is emphasized. Several football studies reinforce the role of shots as effective proxies for success in sports [20, 34]. The conventional metrics of shots and shots on the target may not fully capture the intrinsic value of shot attempts. This research analyzes spatiotemporal patterns, specifically focusing on the 10-second build-up leading to shot attempts. The novelty of this study lies in introducing the concept of defender proximity to the shooter with the aim of enhancing the validity of shot assessments. Utilizing this approach, this study aims to evaluate shot quality and shed light on whether the success of a team in a game can be attributed to dominance or luck. Moreover, shooting type is an important feature for predicting the xG score [3, 32, 35] and distinguishing among four distinct shooting types: open-play footed shots, headers, free kicks, and penalty shots. This study constructed a multilayer perceptron to forecast shot outcomes using a dataset of approximately 10,000 shots [20]. Given its significance in elucidating the unpredictability of goal probability, shot location is a prevalent topic of discussion in almost all studies related to xG [19, 21, 27, 33, 34]. Previous studies primarily rely on event and positional features such as shot location, defender proximity, and type of shot [3, 20, 32, 35]. Few studies consider team-level contextual factors, and none explicitly incorporate managerial or tactical variables, indicating a gap that our study addresses.

Moreover, the techniques outlined earlier are not the only means of forecasting the outcomes using xG. For example, match–outcome prediction is an alternative approach in this context. In contrast to the anticipated goal outcomes, this subject has undergone a thorough examination in previous studies [8, 35]. The most frequently analyzed factor when forecasting match results is the concept of home advantage, which is a phenomenon observed in numerous sports. Home advantage was introduced as a categorical variable denoting home/away conditions. All three models consistently demonstrated that teams had a statistically higher likelihood of winning when playing at home than when playing away [18]. Moreover, an alternative approach employed randomness in match outcomes, with a substantial sample size of 7,304 matches spanning four seasons in the top five leagues. The variables considered in this analysis included distance, angle, rule setting, and body parts used in the shots [3]. An important factor to consider when examining match-outcome predictions is the current form of football clubs [14]. The study includes a variable defining the form of a team, representing the average result over the most recent n games, with 1 indicating a win, 0.5 for a draw, and 0 for a loss. Among the 24 form variables tested, evenly split between the home and away teams (12 each), 20 coefficients proved significant. Match outcome prediction studies focus on factors like home advantage, team form, or randomness in results [3, 8, 14, 18, 35]. Compared to xG-focused studies, these works provide insights into broader match-level outcomes but do not inform feature-level interpretability for shot-level probabilities. In contrast, our approach integrates tactical and coaching features to improve both prediction and explainability. These coefficients align with the expected sign direction–positive for the home team and negative for the away team–across varying values of n.

Generally, the features incorporated into these models are engineered from in-game data and divided into two main categories: event and positional. Event data includes detailed information about all pitch occurrences during a match, such as passes, duels, fouls, and shots. Each data point typically includes details, such as the location of the event on the pitch with x and y coordinates, the ending position of the event (for shots, passes, etc.), the player involved, the match in which the event occurred, the success or failure of the action, and various other variables. A team of individuals manually tags these data points while watching the game. In every machine learning project, the careful selection of features is a crucial aspect, and this also holds true for xG metric prediction. Feature selection can significantly affect the overall xG results. An illustrative study highlighted the importance of features such as goalkeeper positioning, player pressure radius, and opposition between the shot and goal in the development of xG [17]. Synthesizing these studies reveals a consistent gap: managerial and tactical features are largely omitted, limiting interpretability and actionable insights. Our study fills this gap by incorporating coaching experience, duration, and tactical attributes, providing a more comprehensive and explainable xG model.

Furthermore, our objective is to introduce the impact of incorporating new features into the development of xG metrics. Additionally, we present a new classification of xG metrics tailored specifically for the top five European leagues (e.g., La Liga xG).

3 HANDLING OUR DATASET

This section underscores the pivotal role that data play in our research efforts. We initiate the discussion by outlining the complexities of the data collection process, explaining both the origin of our data and pertinent details regarding specific features. The subsequent section expounds on the incorporation of supplementary data, thus enriching the comprehensiveness of our dataset exploration.

3.1 Data Sources

The significance of data in both machine-learning projects and research is evident. The quantity and quality of data play a crucial role in influencing the overall research

outcomes. The data used in this study were collected from two primary sources, FBref [9] and Statsbomb [29]. FBref is a comprehensive football statistics platform that provides detailed data on player and team performance, including metrics such as goals, assists, and passes. Statsbomb is another reputable provider of football data, offering advanced analytics and insights derived from detailed event data [29]. A comprehensive overview and additional details can be found on the publicly available website [28]. Moreover, we primarily utilized Statsbomb data in our dataset; however, to ensure consistency and accuracy, we also incorporated data from the FBref website. We gathered comprehensive and reliable data for our research by comparing and merging these two datasets. Furthermore, we augmented our dataset by incorporating data for new features acquired from a professional sports database, Transfermarkt [5]. In our study, we systematically collected and integrated data from Transfermarkt, focusing on the historical statistics and personal details of the players and coaches. This process involved aligning profiles using unique identifiers to ensure consistency in attributes, such as playing history, positions, team affiliations, and career paths. We chose Transfermarkt for its comprehensive performance data, which are essential for accurate xG forecasting, and for its detailed positional and league-specific insights, which help to account for variations in playing styles across Europe's top five leagues.

Table 2 Personal Details and Tactical Attributes of Coaches

Attribute	Value
Date of Birth/Age	Jun 10, 1959 (64)
Place of Birth	Reggiolo, Italy
Citizenship	Italy
Avg. term as coach	2.29 years
Coaching License	UEFA Pro License
Preferred formation	4-3-3 attacking
Agent	CAA Base Ltd. verified

This integration enhances our analysis by combining match-specific performance data with broader career statistics, providing a holistic approach to xG modeling. Tab. 2 presents new features derived from Transfermarkt data, exemplified by Carlo Ancelotti, Real Madrid's head coach for the 2021–2022 seasons. We enriched our dataset with coaching attributes, such as date of birth, nationality, UEFA coaching license level, average tenure, and preferred tactical formation. Incorporating coaching data is vital because the style, tactics, and experience of a coach significantly influence the xG of a player. By linking each player to their respective coach for a given season, we capture the impact of coaching strategies on individual player performance, providing a more comprehensive xG analysis. For instance, the average tenure of a coach indicates managerial stability, which influences player development and team roles. Preferred formations, such as a 4-3-3 attack setup, reveal the offensive or defensive strategy of a coach, which affects the goal-scoring opportunities of players. Linking players to their coaches allows xG forecasts to account for both tactical and managerial factors. By combining player and coach data, we enhance xG predictions by considering individual performance in broader tactical and managerial contexts.

3.2 Data Preprocessing

Our dataset comprises 2,917 observations, with each observation corresponding to a unique entity within the dataset. Tab. 3 provides an overview of the sample dataset, highlighting the key player attributes and coaching-related features. Despite the modest size of the dataset, 90 distinct features known for their reliability in predicting xG were carefully curated. Among these, xG serves as the primary target variable, representing the expected goals associated with each observation. The dataset includes player-specific attributes, such as name, nationality, position, age, birth year, affiliated squad, and competition level. The dataset primarily covers the 2020–2021 football season, with an average player age of 25.27 years. Feature selection follows domain knowledge and correlation-based criteria, while missing values are handled using mean imputation for numerical features and mode imputation for categorical features to maintain data completeness. A significant proportion of players come from top-tier European leagues, including the Bundesliga, Premier League, La Liga, Serie A, and Ligue 1. In addition, the dataset uniquely incorporates coaching-related factors, recognizing the influence of managerial strategies on xG outcomes. Features such as coach name, average tenure (Avg_duration_years), and preferred tactical approach (Tactic) provide insights into how different coaching styles impact players' xG. For example, players under possession-based attacking coaches may generate higher xG values, whereas those under defensive-oriented managers may limit offensive opportunities. These coaching and tactical variables are operationalized as contextual features that capture managerial influence beyond individual player statistics. Categorical attributes such as coach identity and tactical approach are encoded using label encoding to preserve model compatibility while maintaining relative distinctions across categories. Although label encoding does not impose ordinal meaning, it enables the models to learn systematic differences associated with specific coaching contexts. Average tenure is included as a continuous variable to reflect managerial stability and its potential effect on team cohesion and chance creation. Collectively, these features are intended to model structural and strategic environments that shape players' expected goal generation rather than to imply direct causal effects. To assess the predictive power of the model, machine learning metrics, such as R^2, were used to evaluate the performance across multiple models. Although the dataset provides valuable insights, future studies could benefit from a larger dataset to enhance model accuracy, improve generalizability, and capture deeper patterns within the data.

We systematically categorized the players into distinct positions: goalkeepers (GK), defenders (DF), midfielders (MF), and forwards (FW). Additionally, we identified players with adaptable capabilities, spanning positions such as Defender-Forward (DF-FW) and Forward-Midfielder (FW-MF). Moreover, coaching details, such as the average number of years a coach spends with one team, were considered crucial features in our dataset. The average coaching tenure within our dataset is 1.92 years, providing valuable insights into the stability and dynamics of coaching roles. Furthermore, it encompasses a diverse array of player

performance metrics within the realm of football, including, but not limited to, goal scoring, assist provision, defensive actions such as tackles and interceptions, passing accuracy, shooting endeavors, and involvement in penalty situations. In our data analysis process, we employed heatmap analysis, which is a visual technique that allows the identification of correlations between different features within the dataset. This method facilitates the exploration of complex relationships and aids in determining the importance of the features. By leveraging heatmap analysis, we sought to identify the most influential features, thereby enhancing the effectiveness of subsequent analyses and modelling efforts.

Table 3 Sample Dataset Featuring Player Information, Coaching Influence, and xG Metrics

Player	Pos	Age	Born	Comp	Coach	Avg_duration_years	Tactic	…	xG
Max Aarons	DF	21	2000	EPL	Dean Smith	2.44	Possession and attacking	…	0.7
Yunis Abdelhamid	DF	33	1987	Ligue 1	Oscar Garcia	0.79	Strong defense	…	1.2
Salis Samed	MF	21	2000	Ligue 1	Pascal Gastien	4.43	Strong defense	…	0.8
Laurent Abergel	MF	28	1993	Ligue 1	Christophe Pelissier	3.35	Possession and attacking	…	2
Charles Abi	FW	21	2000	Ligue 1	Pascal Dupraz	1.66	Strong defense	…	0

Table 4 Final Feature Overview of Features for the Dataset

Abbreviation	Description	Data Type	Note
Coach	Coach name	Object	New feature
Avg_duration_years	Avg_duration_years	Float	New feature
Tactic	Tactic description	Object	New feature
Succ	Success count	Integer	Based on correlation
At	Attempt count	Integer	Based on correlation
Cmp %	Completion percentage	Float	Based on correlation
Gls	Goals scored	Integer	Based on correlation
As	Assist made	Integer	Based on correlation
Pk	Penalty kick	Integer	Based on correlation
xG	Expected goals	Float	Based on correlation
SCA	Shot creation actions	Integer	Based on correlation
Field	Fielding count	Integer	Based on correlation
Def	Defensive actions	Integer	Based on correlation
GCA	Goal creation actions	Integer	Based on correlation
MP	Match played	Integer	Based on correlation
Min	Minutes played	Integer	Based on correlation
Crdy	Yellow cards	Integer	Based on correlation
Crdr	Red cards	Integer	Based on correlation
Sht	Shots taken	Integer	Based on correlation
Pass	Passes made	Integer	Based on correlation
Clr	Clearances	Integer	Based on correlation
Err	Errors committed	Integer	Based on correlation
Blocks	Blocks made	Integer	Based on correlation
Touches	Number of touches	Integer	Based on correlation

Tab. 4 presents the final set of 24 features selected through heatmap analysis with a correlation threshold of ±0.5, ensuring strong relationships with the target variable. This set includes three newly introduced features - coach, average coaching duration, and tactics - whereas the remaining 21 features were identified based on their correlation coefficients surpassing the predefined threshold. Each feature is detailed with its abbreviation, description, and data type, encompassing Object, Float, and Integer.

Categorical features, such as coach name and tactic description, are label-encoded into numerical values for model compatibility. Numerical features, including goals scored, assists made, and penalty kicks, are already in a suitable format for the model input. Integrating both categorical and numerical data provides comprehensive insights into player and team performance, thereby enhancing the predictive accuracy of the model. Moreover, to ensure the robustness of the encoded features, categorical variables such as coach and tactical approaches are transformed using label encoding, converting each category into a unique numerical value. We verified the number of unique categories for both the coaches and tactics.

3.3 New Features for Accurate xG

This study introduces three new features: Coach, Average Coaching Tenure, and Preferred Formation. These three features are considered significant and can assist in predicting xG. The soccer team coach plays a crucial role in shaping the playing style, strategy, and tactics of a team. Coaches have distinct philosophies and approaches to games that directly influence team performance on the field. By incorporating information about the current coach into the predictive model, we can capture the tactical preferences, game plans, and managerial decisions that affect the attacking dynamics of a team. For example, coaches play a pivotal role in shaping the identity and tactical approach of a team. Each coach brings forth a distinct blend of philosophy and strategy, molding the playing style of the team. Klopp and Simeone's contrasting approaches offer fascinating insights into the spectrum of coaching philosophies. Klopp's relentless pursuit of attacking football at Liverpool FC epitomizes his "gegenpressing" philosophy, in which high-intensity pressing and rapid counter-attacks reign supreme during the 2021–2022 season. Conversely, Simeone's

pragmatic approach to Atlético Madrid in the 2021–2022 season prioritizes defensive solidity and disciplined counterattack. Although Simeone's tactics may curb xG opportunities, his strategic prowess elevates Atlético's resilience and competitiveness on the field [25].

Moreover, the average coaching tenure refers to the typical duration in which coaches remain in charge of a team. Coaching stability and continuity have significant implications on team performance and player development. Teams with long-term coaches often benefit from consistent tactical frameworks, well-established play patterns, and strong cohesion. For example, Manchester United experienced coaching instability after the departure of Sir Alex Ferguson in 2013, with several managerial changes over the years [11]. In contrast, Real Madrid enjoyed a relatively stable coaching duration under Zinedine Zidane during the 2016–2018 seasons, which led the team to have multiple Champions League titles. The average coaching period reflects continuity and stability within the coaching setup of a team. Teams with consistent coaching leadership tend to develop a cohesive playing style, build long-term strategies, and foster strong team chemistry, all of which positively influence their xG predictions [22].

In addition, the formation adopted by a soccer team reflects its tactical setup and positional structure during matches. Different formations allocate players to specific roles and positions on the field, thereby affecting their interactions, movements, and attacking patterns. For example, FC Barcelona, particularly under Guardiola in the 2008–2012 seasons, often used 4-3-3 as its preferred starting shape, with an emphasis on intricate passing, positional interchange, and exploiting spaces between opposition lines. In contrast, Bayern Munich under Hansi Flick excelled with a more flexible and dynamic 4-2-3-1 or 4-3-3 formation in the 2019–2021 seasons, emphasizing quick transitions, direct attacking play, and aggressive wing play. The preferred formation shapes the tactical approach, player positioning, and offensive strategies of a team, directly influencing xG predictions. Understanding how teams adapt their formations to exploit their strengths and opponents' weaknesses enhances the accuracy of xG models by capturing the spatial dynamics and strategic variations inherent to different formations [4].

Finally, Tab. 5 shows the impact of coaches. Table 5 provides a comparative analysis of the statistical data, showing the differences in performance metrics for Liverpool before and after the tenure of Jürgen Klopp since 2015 [10]. Before Klopp's arrival, Liverpool played 350 games, winning 171, drawing 84, and losing 95 games. Their goals for and against were 586 and 384, respectively. This resulted in an average point per game of 1.7, win rate of 48.9%, and loss rate of 27.1%. The goals per game and goals against per game were 1.67 and 1.09, respectively. During this period, Liverpool won only one trophy–the League Cup.

However, under Klopp's guidance, Liverpool's performance transformed remarkably. In the same number of games (350), Liverpool won 210, drew 81, and lost 59 games. Their goals for surged to 727, and their goals against reduced significantly to 352. This yielded an impressive point per game average of 2.03, with the win rate increasing to 60% and the loss rate decreasing to 16.8%. Notably, the number of goals per game increased to 2.07, while goals against per game decreased to 1.005. Klopp's influence is further underscored by the significant increase in trophies won, with Liverpool clinching four prestigious titles: Premier League, Champions League, Super Cup, and FIFA Club World Cup. This stark contrast in Liverpool's performance before and during Klopp's tenure provides compelling evidence of the pivotal role played by coaches in shaping the performance and xG of a team. Klopp's tactical acumen, motivational prowess, and ability to instill a winning mentality not only elevated Liverpool's performance but also significantly impacted their ability to create and capitalize on goal-scoring opportunities. Thus, the data clearly illustrates how a coach can significantly influence team success, highlighting the importance of coaching in soccer.

Table 5 Comparison of Liverpool's Statistics before and with Klopp

Statistic	Before Klopp: until 2015	With Klopp: from 2015
Games	350	350
Won	171	210
Drawn	84	81
Lost	95	59
Goals for	586	727
Goals against	384	352
Points per game	1.7	2.03
Win rate	48.9%	60%
Loss rate	27.1%	16.8%
Goals per game	1.67	2.07
Goals against per game	1.09	1.005
Trophies won	1 (League Cup)	4 (Premier League, Champions League, Super Cup, FIFA Club World Cup)

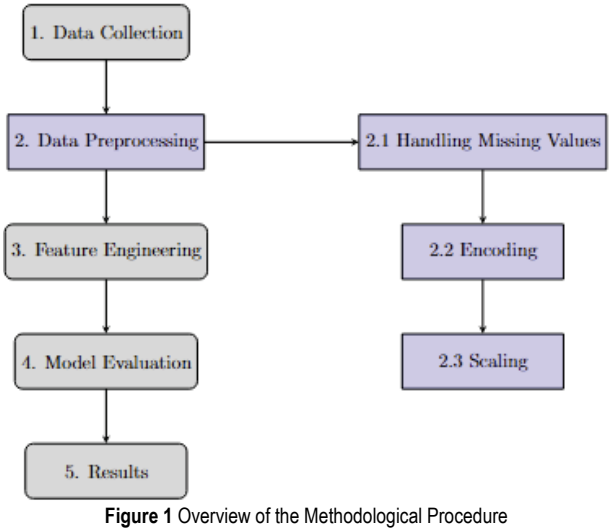

Figure 1 Overview of the Methodological Procedure

3.4 Overall Procedures for Our Analysis

This section describes the step-by-step process used in the analysis, as illustrated in Fig. 1. The methodology begins with data collection, in which Transfermarkt serves as the primary data source. After gathering the necessary data,

preprocessing is performed to ensure data quality and model readiness. Missing values are handled using the mean for numerical features and the mode for categorical features. Categorical variables are encoded using label encoding, whereas numerical features are subjected to standard scaling to maintain consistency across different feature ranges. Following preprocessing, feature engineering is conducted to extract meaningful information that enhances the predictive performance. Various machine learning models are then applied, as detailed in Section 4, in which the model comparison and evaluation are explained. Finally, the methodology concludes with the results presented in Section 5, which provides an in-depth analysis of model performance and findings.

4 MODEL COMPARISON AND EVALUATION

In this section, we commence by providing an introduction to various machine-learning models. Subsequently, we employed a comparative evaluation approach to select the most suitable model for our study, prioritizing the attainment of the highest score.

4.1 Model Introduction

In this study, we employed supervised learning methodologies because of the availability of both input and output (target) values. Within the realm of supervised learning, models typically fall into two categories: regression and classification. Notably, our study forgoes the utilization of classification machine learning models because our objective does not involve categorizing target variables or predicting binary outcomes, such as the presence or absence of a goal. Instead, we focused on predicting the xG score, which is a continuous numerical variable. Hence, we opted for regression machine-learning models to address this objective. These include the decision tree, which is valued for its interpretability and capability to capture complex relationships between features and target variables. The random forest model is well known for its ability to mitigate overfitting and enhance the predictive accuracy through ensemble learning. Ridge regression introduces a penalty term to address multicollinearity in the dataset and improve model stability. Lasso regression facilitates feature selection by penalizing the size of the model coefficients. Finally, the XGBoost regressor, a potent gradient boosting algorithm, is known for its ability to enhance predictive performance by iteratively improving weak models. The inclusion of managerial and tactical features is motivated by their influence on team strategy, player positioning, and shot selection. By integrating these features, the models aim not only to predict xG more accurately but also to provide interpretable insights into which factors drive expected goals, connecting the methodology to a decision-making framework.

4.2 Comparative Evaluation

Determining the most appropriate machine-learning model is a critical stage in experiments. Consequently, we employed a comparative evaluation approach to select the most suitable model for our dataset based on its highest score.

We utilized two experimental methodologies to compare the machine-learning models. In the first approach, we utilized a dataset obtained from Statsbomb [28, 29]. Conversely, in the second approach, we enrich our existing dataset by incorporating additional features such as coaches, tactics, and average duration. To achieve this, we sourced data from a professional soccer website called Transfermarkt [5]. Moreover, in both experiments, we evaluated the performance of our models using established metrics, including mean squared error (MSE), explained variance score, and R-squared. These metrics offer comprehensive insight into the accuracy, predictive capability, and overall fit of the models. The MSE quantifies the average squared difference between the predicted and actual values, providing a measure of the accuracy of the model. The explained variance score indicates the proportion of variance in the target variable that the model can explain, with higher scores reflecting a better predictive capability. R-squared measures the proportion of variance in the target variable explained by the model, and offers an indication of its overall fit to the data. Moreover, to ensure robust evaluation, we split the dataset into training and test sets using an 80/20 ratio, with the training set used to fit the models and the test set reserved for assessing the predictive accuracy of unseen data. To improve model performance, hyperparameters were selected using standard practices and preliminary tuning, including Randomized Search. Although the code cannot be shared due to restrictions, all model settings are fully described to ensure transparency and reproducibility. Performance metrics were calculated on the test set to provide a comprehensive assessment of predictive accuracy and model fit. Tab. 6 shows that the comparison between the performance of the regression models under two distinct approaches (without new features and with new features) reveals noteworthy improvements in the predictive accuracy and model performance with new features. Based on an evaluation of several regression models, the Ridge Regression model emerged as the best performer for our dataset when new features were incorporated. Here, we present a detailed analysis of the performance of each model and the rationale behind selecting the Ridge Regression model as the most suitable. The decision tree model shows improvement with the inclusion of new features, where MSE improves by approximately 6%, and explanatory power highlights minor changes in explanatory power and goodness of fit. This highlights the sensitivity of decision trees to additional contextual variables such as coach, coaching duration, and tactics. Ridge and Lasso Regression also show modest gains in fit, whereas Random Forest and XGBoost remain largely unaffected, with only negligible shifts in performance. Although the numerical improvements are modest, the results highlight how contextual variables such as coach, coaching duration, and tactics can shape predictive accuracy and model behavior. This contribution lies not only in performance gains but also in revealing which models are more sensitive to contextual information, offering valuable insights for both modeling and practical decision-making.

Table 6 Performance Comparison of Regression Models

Model	Without new features			With new features		
	MSE	*EVS*	R^2	*MSE*	*EVS*	R^2
Decision Tree	1.51	0.83	0.83	1.42	0.84	0.84
Random Forest	0.58	0.93	0.93	0.61	0.93	0.93
Ridge Regression	0.50	0.94	0.94	0.49	0.96	0.96
Lasso Regression	0.50	0.94	0.94	0.49	0.96	0.96
XGBoost Regressor	0.62	0.93	0.93	0.60	0.93	0.93

Overall, the inclusion of coaching- and tactics-related features improves the predictive performance across models, with the decision tree benefiting the most in terms of accuracy and explanatory capability. These results highlight the importance of feature engineering for enhancing model performance, particularly for models such as decision trees, which rely heavily on feature differentiation. Moreover, the adjustments implemented in approach two lead to all metrics being improved across various regression models compared to approach one. These improvements, reflected in the reduced *MSE* and increased *EVS* and *R*-squared scores, underscore the effectiveness of the approach in enhancing predictive accuracy and model performance.

5 EVALUATION RESULT

The results presented in this section are intended to provide a comparative and descriptive evaluation of predicted xG patterns across leagues and players. Our analysis focuses on identifying associations and performance trends rather than establishing causal relationships or statistical significance. In this section, we present an overview of our comparative evaluation of the top five soccer leagues based on their average xG using the Lasso Regression model, which is known for its superior performance metrics, as discussed earlier. Additionally, our analysis extends beyond a comparison of the top five leagues by incorporating individual-player statistics. The next section discusses the average xG values of the top five players. Tab. 7 shows the rankings of the top five football leagues based on their average xG during the 2020/2021 football season. Bundesliga, renowned for its fast-paced and competitive matches, leads the list with an average xG of 1.93.

Following closely, the Premier League is known for its attacking play style and high-scoring matches, with an average xG of 1.76. Ligue 1, characterized by its flair and technical prowess, maintains a solid position, with an average xG of 1.70. Serie A, renowned for its tactical sophistication and defensive solidity, ranks fourth, with an average xG of 1.65. Finally, La Liga, famous for its emphasis on possession-based football and technical proficiency, completes the top five with an average xG of 1.55. We present a comprehensive analysis comparing our predicted results with the actual goals scored across the top five football leagues during the 2022–2023 season [24]. The relationship between xG and actual goal-scoring outcomes is presented in Tab. 8. Moreover, comparing the predicted xG for the 2020/2021 season with the actual goals scored in the 2022/2023 season presents an opportunity to recognize the two-season gap. This temporal comparison is exploratory in nature and aims to examine the stability of xG-based patterns over time, rather than to claim longitudinal generalization or predictive causality across seasons.

Table 7 Comparison of Average Expected Goals across Leagues during the 2020–2021 Season

League	Average xG
Bundesliga	1.93
Premier League	1.76
Ligue 1	1.70
Serie A	1.65
La Liga	1.55

Table 8 Average Number of Actual Goals per Match in the 2022–2023 Season

Rank	Leagues	Actual goals
1	Bundesliga	3.17
2	Premier League	2.85
3	Ligue 1	2.80
4	Serie A	2.56
5	La Liga	2.51

For example, it offers the unique opportunity to observe the impact of a two-season gap. This comparative analysis allows for the long-term evaluation of the predictive model, assessing its consistency and reliability over time. By examining how well the model predictions hold across different seasons, researchers can gauge their robustness and applicability beyond their immediate performance.

Furthermore, it enables the assessment of performance consistency. Analyzing the predictions made two years prior to the actual outcomes helps determine whether the model consistently forecasts xG values effectively across various periods. Consistent accuracy reinforces confidence in the reliability of the model for decision making in soccer analytics. Finally, this approach validates the utility of the model in practical scenarios. Comparing predictions across seasons demonstrates how effectively the insights of the model translate into real-world applications, such as informing team strategies or evaluating player performance over an extended timeframe. Starting with the Bundesliga, our examination uncovers a notable average xG of 1.93 per match, highlighting a strong inclination to generate scoring opportunities within the league. This aligns seamlessly with the Bundesliga's renowned reputation for dynamic and attacking football, as evidenced by its leading average of 3.17 actual goals per match, demonstrating the efficiency of Bundesliga teams in capitalizing on these opportunities. When transitioning to the EPL, we observe a steady level of goal-scoring potential, with an average xG of 1.76 per match. This corresponds well with the robust average of the league of 2.85 actual goals per match, reaffirming its position as one of the most competitive and goal-rich leagues in the world. In Ligue 1, the typical xG hovers around 1.70 per game, indicating a moderate number of scoring chances. This aligns with the average of the league of 2.80 actual goals per match among participating teams, enhancing the sense of balance and excitement in French football. Moving to Serie A, we note an average xG of 1.65 per match, signaling a slightly reduced likelihood of goal-scoring opportunities compared with other leagues. Serie A's average of 2.56 actual goals per match indicates a marginally decreased goal-scoring rate relative to its xG expectancy. Lastly, La Liga exhibits the

lowest average xG per match at 1.55, suggesting a comparatively lower frequency of goal-scoring chances. However, the league maintains consistent goal-scoring efficiency, as evidenced by its average of 2.51 actual goals per match, showcasing effectiveness, despite fewer expected goal-scoring opportunities. These observed differences should be interpreted as correlational patterns rather than causal effects, reflecting league-specific playing styles and contextual factors captured by the model.

Table 9 Players and Average xG (in descending order)

Best xG Player	Average xG
Robert Lewandowski	32.6
Karim Benzema	23.5
Kylian Mbappe	23.4
Ciro Immobile	22.2
Mohammed Salah	21.8

Overall, our analysis underscores the robustness of our predictive model in capturing the dynamics of goal-scoring opportunities, thus providing invaluable insights into team performance and goal-scoring potential across top football leagues. Consequently, the results suggest that our model provides slight improvements in predicting xG relative to actual goal outcomes, offering insights into football dynamics at the highest level. Moreover, Table 9 presents the top five players ranked by average xG as estimated by our predictive model, illustrating how the model captures individual-level goal-scoring potential across the top five European leagues. These players consistently exhibit high predicted xG values, indicating that the model effectively identifies elite attacking performance patterns rather than relying solely on league-level trends. Robert Lewandowski records the highest average xG (32.6), reflecting the model's sensitivity to sustained shot quality and volume, and highlighting his central offensive role for Bayern Munich in the Bundesliga. Similarly, Karim Benzema (23.5) and Kylian Mbappé (23.4) demonstrate high xG values that align with their tactical roles and attacking responsibilities at Real Madrid and Paris Saint-Germain, respectively. Ciro Immobile (22.2) and Mohamed Salah (21.8) further illustrate the model's ability to generalize across different leagues and tactical systems, capturing consistent goal-scoring contributions in Serie A and the Premier League. By linking player-level xG outputs to the model's predictions, this analysis reinforces the internal coherence of the results and demonstrates how the proposed features support reliable xG estimation at both league and individual-player levels. While evaluating the model predictions across the top five European leagues, we observed some discrepancies between the predicted xG and actual goals. For instance, the Bundesliga shows the highest average actual goals per match (3.17), surpassing the Premier League (2.85) and La Liga (2.51). This outcome is influenced by the presence of a high-performing player, Robert Lewandowski, whose season average xG of 32.6 significantly elevates the overall scoring metric of the league. Similarly, although the Premier League and La Liga have strong offensive statistics, the predicted xG does not fully align with the actual goals in some matches, particularly when standout individual performance disproportionately affects league-level aggregates. These deviations indicate that the model may underpredict or overpredict matches with exceptional player contributions or atypical scoring patterns. Such an analysis highlights the fact that high-xG players or dominant performances in a season can skew league-wide statistics, emphasizing the importance of considering individual player influences in future model refinements.

6 LIMITATIONS AND FUTURE WORK

Although our study encountered certain limitations, particularly in data collection, constraints imposed by modern technology restricted us to specific online platforms, thereby reducing data diversity and potentially introducing biases inherent in the original collection methods. The dataset, which consists of nearly 3000 observations across five leagues, is relatively limited for machine learning models. However, representativeness is ensured by including all matches with complete player and event data during the study period, covering diverse teams, player roles, and match situations. The absence of real-time data hindered our ability to capture immediate trends and reduced the timeliness of the study. The temporal gap between the training (2020/21) and validation (2022/23) data may introduce confounding factors such as player transfers, managerial changes, and evolving team compositions. Accordingly, this cross-season comparison is interpreted as a robustness check rather than a controlled longitudinal or causal evaluation of the xG prediction model. The absence of real-time data and player injury information may have overlooked factors affecting the outcomes. Future research could leverage advanced data collection technologies, such as real-time tracking systems and integrate additional performance metrics to improve the predictive accuracy and applicability in football analytics.

7 CONCLUSION

In conclusion, this study introduces novel features to enhance the predictive capabilities of xG models in football analytics. The primary novelty of this work lies in the explicit integration of coaching-related and tactical variables into xG modeling, which are largely absent from existing event- and position-based approaches. Incorporating coach attributes such as average tenure and preferred tactics provides valuable insights into the multifaceted nature of xG prediction. Regression machine learning algorithms, including decision tree, random forest, ridge regression, lasso regression, and XGBoost regressor, have facilitated the development of robust predictive models. Comparative analysis indicates that the ridge and lasso regression models perform comparatively well, showing a lower MSE than other models, whereas the decision tree model demonstrates a 6% reduction in MSE with the inclusion of new features, highlighting the value of feature engineering and regularization techniques in enhancing prediction quality. The findings show that incorporating contextual and managerial factors into machine-learning frameworks can meaningfully extend conventional xG modeling approaches. Additionally, a statistical approach was proposed to compare the top five football leagues–Premier League, Bundesliga, La Liga, Serie A, and Ligue 1–based on xG metrics. The Bundesliga emerged as a standout league, exhibiting the

highest average xG values. The strong correlation between our predicted xG values and the actual goals per match in these leagues during the 2022–2023 season demonstrates the reliability and practical applicability of our model. This study also identified the top five players with the highest xG scores across these leagues, with Lewandowski leading. This comprehensive analysis offers valuable insights into league and player performance dynamics, and contributes to the advancement of football analytics and decision-making processes. The generalizability of the results is constrained by the dataset size, the exploratory nature of the cross-season comparison, and the absence of advanced model interpretability analyses, which are addressed as directions for future research. Overall, this research provides a reference point for informed decision making and supports further exploration in football analytics, acknowledging that model performance has certain limitations.

8 REFERENCES

[1] Analyst, O. (2023). Expected goals (xG). *Opta Analyst.* https://theanalyst.com/eu/2023/08/what-is-expected-goals-xg/

[2] Anzer, G., & Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*, 53. https://doi.org/10.3389/fspor.2021.624475

[3] Brechot, M., & Flepp, R. (2020). Dealing with randomness in match outcomes: How to rethink performance evaluation in European club football using expected goals. *Journal of Sports Economics*, 21(4), 335–362. https://doi.org/10.1177/1527002519897962

[4] Bundesliga. (2020). How Hansi Flick steered Bayern Munich to Klassiker victory over Borussia Dortmund in his first Bundesliga game. *Bundesliga.* https://www.bundesliga.com/en/bundesliga/news/how-hansi-flick-steered-bayern-to-4-0-win-over-dortmund-klassiker-tactics-8142

[5] Transfermarkt. (2023). Coaches: Transfermarkt data. *Transfermarkt.* https://www.transfermarkt.com/carlo-ancelotti/profil/trainer/523

[6] Driblab. (2020). Expected goals (xG): What it is and how it works. *Driblab.* https://www.driblab.com/driblab-en/expected-goals-xg-what-it-is-and-how-it-works/

[7] Eggels, H., Van Elk, R., & Pechenizkiy, M. (2016). Explaining soccer match outcomes with goal scoring opportunities predictive analytics. In *Proceedings of the Conference on Soccer Analytics.* https://doi.org/10.1080/000368400421101

[8] Falter, J.-M., & Perignon, C. (2000). Demand for football and intramatch winning probability: An essay on the glorious uncertainty of sports. *Applied Economics*, 32(13), 1757–1765.

[9] Sport Website. (2024). Soccer: FBref. *FBref.* https://fbref.com/en/

[10] Planet Football. (2022). Comparing Liverpool's record before and after Jürgen Klopp's appointment. *Planet Football.* https://www.planetfootball.com/quick-reads/comparing-liverpools-record-before-and-after-jurgen-klopps-appointment-in-2015/

[11] Mcevoy, S. (2018). Manchester United and Sir Alex Ferguson with 27 years of stability. *MailOnline.* https://www.dailymail.co.uk/sport/football/article-6507649/Man-United-27-years-stability-theyre-looking-FOURTH-boss-five-years.html

[12] Furniss, M. (2024). Liverpool 4–2 Newcastle stats. *Opta Analyst.* https://theanalyst.com/eu/2024/01/liverpool-4-2-newcastle-stats/

[13] Garratt-Stanley, F. (2022). Jobs in football: What is expected goals (xG)? *Jobs in Football.* https://jobsinfootball.com/blog/what-is-expected-goals-xg/

[14] Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21(2), 331–340. https://doi.org/10.1016/j.ijforecast.2004.08.002

[15] Gu, C., De Silva, V., & Caine, M. (2024). A machine learning framework for quantifying in-game space-control efficiency in football. *Knowledge-Based Systems*, 283, 111123. https://doi.org/10.1016/j.knosys.2023.111123

[16] Herbinet, C. (2018). Predicting football results using machine learning techniques. *MEng thesis, Imperial College London.*

[17] Hewitt, J. H., & Karakus, O. (2023). A machine learning approach for player and position adjusted expected goals in football (soccer). *Franklin Open*, 4, 100034. https://doi.org/10.1016/j.fraope.2023.100034

[18] Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7), 544–553. https://doi.org/10.1016/j.knosys.2006.04.011

[19] Kharrat, T., McHale, I. G., & Pena, J. L. (2020). Plus-minus player ratings for soccer. *European Journal of Operational Research*, 283(2), 726–736. https://doi.org/10.1016/j.ejor.2019.11.026

[20] Lucey, P., Bialkowski, A., Monfort, M., Carr, P., & Matthews, I. (2015). Quality vs. quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In *Proceedings of the Conference on Soccer Analytics.*

[21] Madrero Pardo, P. (2020). Creating a model for expected goals in football using qualitative player information.

[22] Marca. (2021). Zidane departs as Real Madrid's most successful coach. *Marca.* https://www.marca.com/en/football/real-madrid/2021/05/27/60af73fbca4741756c8b46da.html

[23] Noordman, R. (2019). Improving the estimation of outcome probabilities of football matches using in-game information. *Amsterdam School of Economics, Faculty of Economics and Business.*

[24] Okunev, L. (2023). Comparison of the top-5 football leagues by goals scored. *Medium.* https://medium.com/@okunevleo/%D1%81%D1%80%D0%B0%D0%B2%D0%BD%D0%B5%D0%BD%D0%B8%D0%B5-%D1%82%D0%BE%D0%BF-5-%D1%84%D1%83%D1%82%D0%B1%D0%BE%D0%BB%D1%8C%D0%BD%D1%8B%D1%85-%D0%BB%D0%B8%D0%B3-%D0%BF%D0%BE-%D0%B7%D0%B0%D0%B1%D0%B8%D1%82%D1%8B%D0%BC-%D0%B3%D0%BE%D0%BB%D0%B0%D0%BC-288e5ccba7a3

[25] Rajarshi. (2021). Jürgen Klopp vs. Diego Simeone: Head-to-head manager comparison. *Sportco.* https://www.sportco.io/article/jurgen-klopp-vs-diego-simeone-head-to-head-comparison-846574

[26] Robberechts, P., & Davis, J. (2020). How data availability affects the ability to learn good xG models. In *Machine Learning and Data Mining for Sports Analytics: 7th International Workshop (MLSA 2020), Co-located with ECML/PKD.* https://doi.org/10.1007/978-3-030-64912-8_2

[27] Schulze, E., et al. (2018). Effects of positional variables on shooting outcome in elite football. *Science and Medicine in Football*, 2(2), 93–100. https://doi.org/10.1080/24733938.2017.1383628

[28] Somdeep. (2023). Soccer dataset. *GitHub.* https://github.com/somdeep/Statball

[29] StatsBomb. (2023). StatsBomb: Soccer data. *StatsBomb.* https://statsbomb.com/what-we-do/soccer-data/
[30] Tippett, J. (2019). *The expected goals philosophy: A game-changing way of analyzing football.*
[31] Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *PLoS ONE*, 18, e0282295. https://doi.org/10.1371/journal.pone.0282295
[32] Malikov, D., & Kim, J. (2024). Beyond xG: A dual prediction model for analyzing player performance through expected and actual goals in European soccer leagues. *Applied Sciences*, 14(22), 10390. https://doi.org/10.3390/app142210390
[33] Zeng, Z., & Pan, B. (2021). A machine learning model to predict player's positions based on performance. In *Proceedings of icSPORTS*, Valletta, Malta, 28–29 October 2021, 36–42. https://doi.org/10.5220/0010653300003059
[34] Anzer, G., & Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*, 3, 624475. https://doi.org/10.3389/fspor.2021.624475
[35] Malikov, D., Jung, P., & Kim, J. (2025). Predicting soccer player salaries with both traditional and automated machine learning approaches. *Applied Sciences*, 15(14), 8108. https://doi.org/10.3390/app15148108

Authors' contacts:

Davronbek Malikov
Department of AI Convergence Engineering, Gyeongsang National University,
501, Jinju-daero, Jinju-si, 52828 Gyeongsangnam-do, Republic of Korea
davronbekmalikov96@gmail.com

Jaeho Kim
(Corresponding author)
Department of AI Convergence Engineering, Gyeongsang National University,
501, Jinju-daero, Jinju-si, 52828 Gyeongsangnam-do, Republic of Korea
jaeho.kim@gnu.ac.kr