

Research Article

Leander Forcher*, Leon Forcher, Stefan Altmann and Alexander Woll

Decoding attacking success in soccer – a data-driven analysis of expected possession value in the Bundesliga

<https://doi.org/10.1515/jqas-2025-0053>

Received March 31, 2025; accepted April 17, 2026;

published online June 9, 2026

Abstract: Understanding the key factors driving attacking success represents a critical challenge in soccer match analysis. A promising approach to address this issue is the application of expected possession value (EPV) models. Therefore, this paper aims to develop an EPV model with high explainability to provide detailed practical insights into the keys to attacking success. Tracking and event data of the Bundesliga season 2022/23 were analyzed (306 matches). From three main categories (match performance offense, defense, & match situation context), 21 features were carefully selected by professional match analysts. Afterward, machine learning classifiers were used (e.g. Random Forest, XGBoost) to predict the goal probability of possessions. The selected EPV model showed satisfactory prediction performance (xGBoost: Accuracy = 0.99, Recall = 0.06, F1-Score = 0.10, AUC = 0.85, logloss = 0.05, ECE = 0.01). The most important features in predicting attacking success were the distance (1st) and angle (2nd) to the goal, the offensive space control in the final third (3rd), and the relative pitch position (4th, defined by the number of outplayed opposing

formation lines). By applying the presented EPV model and interpreting its most important features in individual match situations or over a whole match highly practice-relevant information about the tactical match performance of players can be gained.

Keywords: performance analysis; tactical match performance; elite soccer; soccer analytics; machine learning

1 Introduction

With the increasing availability of a wide variety of data in professional soccer, the effort to evaluate this data has intensified. The ones to reveal the most significant information out of the data points may achieve a competitive advantage. In detail, such information can be used to support numerous decisions and may help to increase the match performance of soccer players (Forcher et al. 2024d).

Understanding the match performance of soccer players is one of the main interests from a sports scientific perspective. To better understand a player's match performance, four different facets of performance can be distinguished including physical, technical, tactical, and psychological match performance (Forcher 2024a,b). Thereby, the goal-oriented behavior of players can be referred to as tactical match performance which describes the effectiveness of players' actions in relation to a shared tactical objective, such as scoring a goal (Forcher 2024a).

Those tactical objectives differ depending on the match situation and the ball possession and, therefore, different playing phases of the game are distinguished in which different tactical objectives are pursued: defensive play, offensive play, offensive transition, defensive transition, and set plays (Bauer et al. 2023; Escher 2020; Hewitt et al. 2016). For instance, in offensive play, a team's tactical objective is to control the ball and eventually score. Thereby, the tactical match performance of an individual player describes the quality of a player contributing to this tactical objective (e.g.

The study was conducted at the Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Karlsruhe, Germany.

***Corresponding author: Leander Forcher**, Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Engler-Bunte-Ring 15, 76133 Karlsruhe, Germany; and TSG 1899 Hoffenheim, Horrenberger Str. 58, 74939 Zuzenhausen, Germany, E-mail: leander.forcher@kit.edu
<https://orcid.org/0000-0002-6428-8643>

Leon Forcher, TSG 1899 Hoffenheim, Horrenberger Str. 58, 74939 Zuzenhausen, Germany, E-mail: leon.forcher@tsg-hoffenheim.de

Stefan Altmann, Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Engler-Bunte-Ring 15, 76133 Karlsruhe, Germany; and TSG ResearchLab gGmbH, Zuzenhausen, Germany, E-mail: stefan.altmann@kit.edu

Alexander Woll, Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Engler-Bunte-Ring 15, 76133 Karlsruhe, Germany, E-mail: alexander.woll@kit.edu

by playing a successful pass into a dangerous area to create a scoring opportunity).

Recently, research on tactical match performance mainly focussed on the on-ball behavior of soccer players in offensive play (Forcher et al. 2022a) including the analysis of passes (Forcher et al. 2021; Goes et al. 2021a; Power et al. 2017) or shots (Anzer and Bauer 2021; Gonzalez-Rodenas et al. 2019).

Yet, with the availability of highly accurate tracking data (Sarmiento et al. 2018), which includes continuous information on the positions of all players on the pitch the off-ball behavior of the players becomes quantifiable (Goes 2023; Herold et al. 2022). Furthermore, the complex interactions within teams (Headrick et al. 2012; Laakso et al. 2017) and between teams (Goes et al. 2021b) can be analyzed in great detail. Accordingly, tracking data has been increasingly used to analyze match performance in soccer (Anzer 2022; Bauer 2022; Goes et al. 2021b) due to its potential to provide highly objective information, thereby enhancing insights into soccer match performance.

Still, especially the large volume and variety of this type of data brings significant challenges, necessitating the use of computational methods for its evaluation. For example, tracking data of a single match can add up to 3 million data points (Memmert et al. 2017). However, this abundance of data also makes it possible to apply advanced artificial intelligence (AI) and machine learning (ML) techniques effectively. ML excels at identifying hidden patterns within complex dynamic systems. In the context of team sports, the collective interactions among players can be interpreted as such a dynamical system (Welch et al. 2021). By leveraging ML approaches, player behaviors linked to success can be analyzed, allowing for the identification of key tactical factors that contribute to a team's performance (Forcher et al. 2025).

Especially in the analysis of offensive play those machine learning models have found their way into the analysis of tactical match performance in soccer. First and foremost, expected goals (xG), which predicts the outcomes of shots (i.e. goal probability), is frequently used to evaluate the offensive performances of individual players or collective teams. Several studies analyzed xG (Lucey et al. 2015; Mead et al. 2023) and overviews of different published xG models in the literature can be found in Cavus and Biecek (2022) and Hewitt and Karakus (2023). For instance, Anzer and Bauer (2021) predicted the outcome of shots based on seven seasons of the first and second German Bundesliga and indicated that synchronized tracking data can complement the information of event data to effectively increase the accuracy of xG models. Another example is

shown by Cavus and Biecek (2022) who presented an xG model based on shots from seven seasons of the top-five European leagues showing a comparatively powerful prediction performance correctly predicting about 30 % of all goals scored.

Overall, xG models provide an objective way to analyze match performance by offering insights into a team's or a player's attacking effectiveness in soccer. For example, the analysis of well-calibrated xG models can help identify players' shooting patterns, such as those consistently taking high-probability shots, and assess their conversion performance by comparing their individual conversion rate to the average rate predicted by the xG model.

Still, goal scoring is hard to predict (<30 % of all goals scored can be predicted correctly (Anzer and Bauer 2021; Cavus and Biecek 2022)) which may be due to the diversity and distinctiveness of match situations and the difficulty of scoring in the low scoring game of soccer.

Further, the interpretability of those models is not always ensured. This concerns, for example, the included features that should display a specified part of match performance to enable their meaningful interpretation (e.g. ball control as part of technical match performance). Additionally, in most models, the defensive side of the game is undervalued (Hewitt and Karakus 2023). Overall, xG is a significant KPI to analyze tactical match performance since it predicts the main objective of offensive play, namely scoring. Nonetheless, by focusing exclusively on shots solely a small subpart of the match of soccer is analyzed.

To value all actions of the game rather than focusing solely on shots, several approaches have been made to analyze the actions that are meaningfully related to the scored goal. This is completed by analyzing sequences of play and connecting them to the tactical goal of scoring. This aim was quantified using Markov chains (Rudd 2011), expected threat (xT) (Singh 2018), or expected possession value (EPV) (Rahimian et al. 2024). For instance, xT describes the increase in scoring probability to which an on-ball action has contributed. It is based on a scoring probability surface and on the location of events in which actions that led to an increase in xT are favored (Singh 2018). While this approach includes the idea of attributing prior actions to the goal it does solely take on-ball event data into account. Furthermore, it overlooks a wide range of tactical and technical KPIs that are essential for a more in-depth analysis of match situations or sequences of play. For instance, the approach could benefit from integrating defensive features such as the pressure on the ball leader or the defensive coverage of close pass options.

For this purpose, the most promising approach is based on modeling EPV. EPV predicts the probability of a possession ending in a scored goal (Fernández et al. 2019, 2021; Olthof and Davis 2025; Rahimian et al. 2024). In doing so, current approaches use neural networks and complex input data (e.g. raw tracking data) (Fernández et al. 2019, 2021). Those input data are then used to predict different submodels such as xG, action selection probabilities (Fernández et al. 2021), or probability pitch surfaces such as the success probability of the action (Rahimian et al. 2024). Afterward, those prediction models and pitch surfaces are implemented by the neural network to predict the probability of the possession outcomes (e.g. scored goal, conceded goal). With it, those EPV models consider off-ball behavior (e.g. by modeling pass option surfaces and space control), enable an evaluation of all actions during a possession (e.g. passes & dribblings), and their sophistication does justice to the difficulty in predicting goals in soccer.

However, the lack of transparency regarding the inner mechanisms and dependencies of these deep learning models is a significant limitation. In many cases, the features used in different models are described only vaguely (Fernández et al. 2021), and details about the performance of the underlying models or the characteristics of the pitch surfaces are not always specified. As a result, the overall performance of these models remains uncertain. While the performance of deep learning models is powerful, their interpretability is hampered.

Nonetheless, this interpretability is essential to reflect on the possibility of making erroneous scientific claims due to data leakage or the frequently lacking or misleading interpretation of the results of AI methods (Gibney 2022; Kapoor and Narayanan 2023). Therefore, sports scientific application of machine learning approaches to analyze the tactical match performance in soccer should focus on the interpretability of their machine learning approaches (e.g. by using explainable AI methods such as Shapley values) to increase the knowledge gain of studies.

Therefore, this study aims to leverage tracking data and machine learning methods to predict attacking success in elite soccer by developing an EPV model. Thereby this study focuses on the interpretability of the developed EPV model, achieved through two key strategies: (i) KPIs grounded in expert-based technical or tactical features, and (ii) employing explainable AI methods to provide detailed insights into the features and mechanisms underlying the predictions. This procedure allows to reveal detailed practice-relevant knowledge about the dependencies of attacking success and

gives insights into the tactical match performance in soccer. Further, hypotheses were formulated specifically for each feature included (please see Table 1, technical & tactical idea).

2 Methods

This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the local ethics committee (Human and Business Sciences Institute, Saarland University, Germany, identification number: 22-02, 10. January 2022).

2.1 Data

Using an observational study design, official tracking and event data of all matches during the 2022/23 season of the German Bundesliga were analyzed post-event (306 matches).

Tracking data was collected using a semi-automatic multi-camera tracking system (TRACAB, ChyronHego, Melville, NY, USA). This tracking system measures the positions of the ball and all players on the pitch with a sampling frequency of 25 Hz and recently has been validated to be suitable for analyzing soccer-specific performance (Linke et al. 2020).

Event data was officially notated by Sportec Solutions (Sportec Solutions AG, Ismaning, Germany) differentiating about 30 events with over 100 attributes defined by the official match data catalog of the German Soccer League (DFL) (DFL 2014). The event data was synchronized to the tracking data (Anzer and Bauer 2021).

The match dates at which the data was collected are available on the official website of DFL (DFL 2024) and all data was retrieved for the presented analysis on 1/11/2023.

All steps of data preprocessing, analysis, modeling, and visualization were completed using python 3.11.8 and the Pandas, NumPy, Math, Matplotlib, SHAP, and scikit-learn libraries.

2.2 Data processing

To ensure the traceability of the analysis, the main analysis steps are described below. This includes the preprocessing steps of the identification of playing phases, the identification of the tactical formation, and the computation of the coverage of pass options. Based on those preprocessing steps expert-based features were computed.

Table 1: Features for EPV prediction coloured according to the three main feature categories (match situation: grey, offense: orange, defense: blue) which were carefully selected by specifying their technical or tactical aspect of match performance.

Category	Sub-category	Feature	Exact features and expression	Calculation	Technical & tactical idea
Match situation	Absolute pitch position	absolute_pitch_position_lane absolute_pitch_position_third	Categorical: outer lane ($\hat{=}$ 1), half lane ($\hat{=}$ 2), middle lane ($\hat{=}$ 3) Categorical: first third ($\hat{=}$ 1), middle third ($\hat{=}$ 2), final third ($\hat{=}$ 3)	Ball position on the pitch, 5 vertical zones Ball position on the pitch, 3 horizontal zones	The description of the absolute pitch position in thirds and lanes assists in roughly subdividing match situations. In match analysis practice this subdivision of the pitch is used to describe different situational context in which different tactics are deployed. In this context, Forcher et al. (2023) could show that successful behavior in defense differs depending on the pitch position of outer lane or middle lane. It is expected that bringing the ball in the final third and the middle lane may increase the chance of scoring in the following seconds of the possession.
		absolute_pitch_position_distance_to_goal	Numerical: [m]	Distance of the ball to the center of the attacked goal	Anzer and Bauer (2021) showed that the distance to the goal is the most important variable in predicting the goal probability of shots. In line with this study, it is expected that bringing the ball closer to the goal may increase the chance of scoring in the following seconds of the possession.
		absolute_pitch_position_angle_to_goal	Numerical: [°] ($0^\circ - 90^\circ$) $0^\circ \hat{=}$ ball is directly beneath goal, $90^\circ \hat{=}$ ball is in the middle of the pitch directly in front of goal	Angle of the ball towards the center of the attacked goal	Anzer and Bauer (2021) showed that the angle towards the goal is an important variable in predicting the goal probability of shots. In line with this study, it is expected that bringing the ball in a greater angle in front of the goal may increase the chance of scoring in the following seconds of the possession.
	Relative pitch position	relative_pitch_position_horizontal_zones	Categorical: before attackers of defending team ($\hat{=}$ 1), between attackers and midfielders of defending team ($\hat{=}$ 2), between midfielders and defenders of defending team ($\hat{=}$ 3), behind defenders of defending team ($\hat{=}$ 4)	Based on the formation lines of the tactical formation of the defending team in playing phase of defensive play (see methods section, tactical formation), 4 horizontal zones	The relative pitch position is based on the tactical principle of play to outplay opposing formation lines to disrupt the opponents' defensive organization and bring the ball closer to the opponents' goal. This tactical principle was quantified following the idea of Fernandez et al. (2019). It is expected that outplaying opposing formation lines and thus bringing the ball behind opposing formation lines may increase the chance of scoring in the following seconds of the possession.
	History of possession	match_situation_number_actions match_situation_possession_duration	Ordinal: number of actions Numerical: [s]	Number of previous actions of current possession (including passes, tacklings, other ball actions, etc.) Time in which the attacking team is already in ball possession up to the current match situation	One tactical objective of playing phase of offensive play next to scoring a goal, is to control the ball to disrupt the opposing defensive organization by passing (Forcher et al. 2021; Goes et al. 2018). Accordingly, it is expected that a longer duration of the possession of the attacking team with a higher number of passes may increase the chance of scoring in the following seconds of the possession.

Table 1: (continued)

Category	Sub-category	Feature	Exact features and expression	Calculation	Technical & tactical idea
Tactical match performance offense	On-ball behavior: ball control ball-leading player	offense_onball_ball_control_distance_to_ball	Numerical: [m]	Distance of ball-leading player to the ball	The ball control of the ball-leading player is central to deliberately perform actions on the ball and control the ball during the possession to eventually create a scoring opportunity. Thereby, a small distance of the ball-leading player to the ball, a small difference of the speed of the ball may indicate a high ball control of the ball-leading player. Furthermore, it is assumed that a ball with higher speed is harder to control for the ball-leading player. Accordingly, it is expected that a higher ball control of the ball-leading player may increase the chance of scoring in the following seconds of the possession.
		offense_onball_ball_control_speed_ballleader	Numerical: [km/h]	Speed of ball-leading player	
		offense_onball_ball_control_speed_ball	Numerical: [km/h]	Difference in speed of ball-leading player and ball	
	Off-ball behavior: pass options	offense_onball_ball_control_speed_difference_player_ball	Numerical: [%] (0–100 %)	Expected pass reception (see methods section, pass options & Appendix 1) for the 5 closest pass options to the ball	One tactical objective of playing phase of offensive play next to scoring a goal, is to control the ball to disrupt the opposing defensive organization by passing (Forcher et al. 2021; Goes et al. 2018). In this context, several studies showed that the coverage of close pass options is crucial to regain the ball in defense (Forcher et al. 2022a, 2023). Interpreted conversely, this shows the importance of open pass options in offensive play in order to maintain control of the ball. In line with those findings, it is expected that a higher number of open pass options in ball proximity may increase the chance of scoring in the following seconds of the possession.
		offense_offball_pass_options_close_to_ball_5_mean	mean of expected pass reception value for 5 attacking players closest to the ball		
		offense_offball_pass_options_close_to_ball_5_higher90	Ordinal: number of players (0–5)	5 attacking players closest to the ball which an expected pass reception value of over 90 [%] (0–2)	
		offense_offball_pass_options_dangerous_2_mean	Numerical: [%] (0–100 %)	Expected pass reception (see methods section, pass options & Appendix 1) for the 2 pass options closest to the attacked goal	The main tactical objective of the playing phase of offensive play is to score goals. Therefore, pass options closest to the opposing goal are expected to be the most dangerous pass receivers to create a scoring opportunity in the following actions. Therefore, it is expected that if those dangerous pass options are not covered it may increase the chance of scoring in the following seconds of the possession.
		offense_offball_pass_options_dangerous_2_higher90	Ordinal: number of players (0–2)	2 attacking players closest to the goal with an expected pass reception value of over 90 [%] (0–2)	

Table 1: (continued)

Category	Sub-category	Feature	Exact features and expression	Calculation	Technical & tactical idea
Tactical match performance offense	Off-ball behavior: space control	offense_offball_space_control_final_third_offense_percentage	Numerical: [%] (0–100)	Percentage of space control of attacking team in the final third of the pitch based on voronoi diagrams	One tactical principle of offensive play is to create space. This controlled space is especially valuable in the final third of the pitch in front of the goal and was used as performance metric in several other approaches (Memmert et al. 2017). It is expected that a higher space control of the attacking team in the final third may increase the chance of scoring in the following seconds of the possession.
	Off-ball behavior: box occupation	offense_offball_box_occupation_offense_offball_box_occupation_numerical_superiority_offense	Ordinal: number of players (0–11) Ordinal: difference in number of attackers and defenders (–11 – 11) positive values \triangleq attackers are in numerical superiority negative values \triangleq defenders are in numerical superiority	Number of players in the penalty box in front of the attacked goal	One tactical principle of offensive play is to bring and position attackers in the opposing penalty box and in the best case outnumber the opposing defensive players. This behavior increases the controlled space in front of the goal and thereby increases the chance of successfully receiving a pass and with it, create a scoring opportunity since most goals are scored within the box. Therefore, it is expected that a higher box occupation of the attackers may increase the chance of scoring in the following seconds of the possession.
Tactical match performance defense	Off-ball behavior	defense_offball_defensive_pressure_on_ballleader	Numerical: [%] (0–100)	Defensive pressure on the ball-leading player is quantified using the defensive pressure metric that was used to quantify pass options (compare methods section pass options & Appendix 1) Defending players in shot area the shot area stretches between the ball position and the two goalposts of the attacked goal	One tactical objective of defensive play is to regain the ball, several studies showed that defensive pressure on the ball-leading player is the most important performance indicator to deny the actions of the ball-leader (e.g., passes or dribblings) and regain the ball (Forcher et al. 2022b). Accordingly, it is expected that a smaller defensive pressure on the ball-leading player may increase the chance of scoring in the following seconds of the possession.
		defense_offball_defenders_in_line_of_shot	Ordinal: number of players (0–11)	Goalkeeper in goal is defined by the position of the goalkeeper (goalkeeper in line of shot) and his distance to the goal (closer than 11 meters to the goal)	One of the tactical objectives in defensive play is to protect the own goal. In the final stages of an opposing attack defenders, therefore, place themselves between the ball-leading player and the own goal to block a possible shot (Anzer and Bauer 2021). It is expected that a smaller number of defenders in line of the shot may increase the chance of scoring on the following seconds of the possession.

2.2.1 Playing phases

The playing phases were determined based on the event data. At first, ball possessions were identified using a rule-based approach to calculate possession chains following the example of (Sumpster and Andrzejewski 2022). Based on the defined ball possessions, the different playing phases – offensive play, defensive play, defensive transition, offensive transition, and set plays – were identified. This study focused exclusively on the playing phase of offensive play. Offensive plays were defined as sequences involving at least two on-ball actions (e.g. duel or dribbling) or a successful pass occurring closer to the attacking team’s goal than 10 of the 11 opposing players positioned behind the ball. The offensive phase stopped when the ball went out of play, play was stopped due to an event such as a foul, or the opposing team regained possession of the ball.

During a possession of offensive play, all match situations in which an on-ball action (e.g. pass, dribbling) of the attacking team was performed were analyzed. Accordingly, all features (see methods section Features) were computed based on the tracking data of the identified time-frame of the on-ball action. This procedure was chosen to effectively analyze match situations in which the attacking team has ball control. Otherwise, modeling of pass options would not make sense if computed during a ball move.

Further, only 11 versus 11 match situations were taken into account. Accordingly, all matches with player dismissals were excluded (42 matches). Additionally, all match situations in which not all 22 players and the ball were identified were excluded.

2.2.2 Tactical formation

The tactical formation was identified within three time-windows of a whole match: first half (0–45 [min]), first interval of second half (46–62.5 [min]), second interval of second half (62.6–90 [min]). Those time-windows were chosen to account for in-game formation changes which highly influence the match performance in soccer (Forcher et al. 2022b) and have been found to occur almost exclusively in the second half (95.2 %) (Forcher et al. 2023).

Next to in-game formation changes, previous studies also showed that tactical formations also significantly vary depending on the playing phases (Bauer et al. 2023). Therefore, the tactical formations were solely identified in the considered playing phase of offensive play that was investigated in this study.

To quantify a tactical formation, a KMeans clustering algorithm was used to cluster the longitudinal x -positions

of all players of a team into three formation lines (e.g. 5-3-2) (Forcher et al. 2024a; Goes et al. 2021b).

Overall, this results in 6 formations per team per match (3 time-windows \times 2 formations [defending team & attacking team]). The information about the tactical formation was used to define the relative pitch position of a match situation (see Table 1).

2.2.3 Pass options

The coverage of pass options for the ball-leading player was analyzed to account for the defensive side of the game. In contrast to the main objective of offense to score a goal, the defense aims to regain the ball from the attacking team. In this context, several studies showed that it is highly important to cover close pass options to regain the ball (Forcher et al. 2024b,c).

To analyze pass options, two key performance indicators were used. First, defensive pressure (DP) was analyzed to measure the spatial pressure exerted by the defending team on a potential pass receiver. Additionally, the coverage of the passing lane (CPL) towards a potential pass receiver was quantified to analyze the defending team’s ability to intercept a given pass. Both pieces of information were used to quantify the expected pass reception of a given pass option (xPR, see Figure 1).

DP was quantified using the approach of Andrienko et al. (2017) which includes the distance and the orientation of a defender to the pressurized attacker. Thereby, an oval pressure area is formed around the attacker (“pass receiver”) which is oriented towards the attacked goal. Within this pressure area, the pressure increases the closer the defender is positioned to the attacker (see Figure 1). This approach was updated by decreasing the pressure area the closer the attacker is positioned towards the attacked goal following the idea of Herold et al. (2022).

CPL was quantified by shaping a triangular pass coverage area around the expected passing lane of a given pass. Within this coverage area a defender increases the possibility of intercepting a possible pass the closer he is positioned toward the passing lane (see Figure 1).

All detailed quantifications (DP, CPL, & xPR) can be found in the appendix (see Appendix 1).

Overall, the resulting xPR metric was used to quantify the coverage of possible pass options both for the 5 closest pass options to the ball and the most dangerous pass options of the attacking team which were defined by the closest distance to the goal.

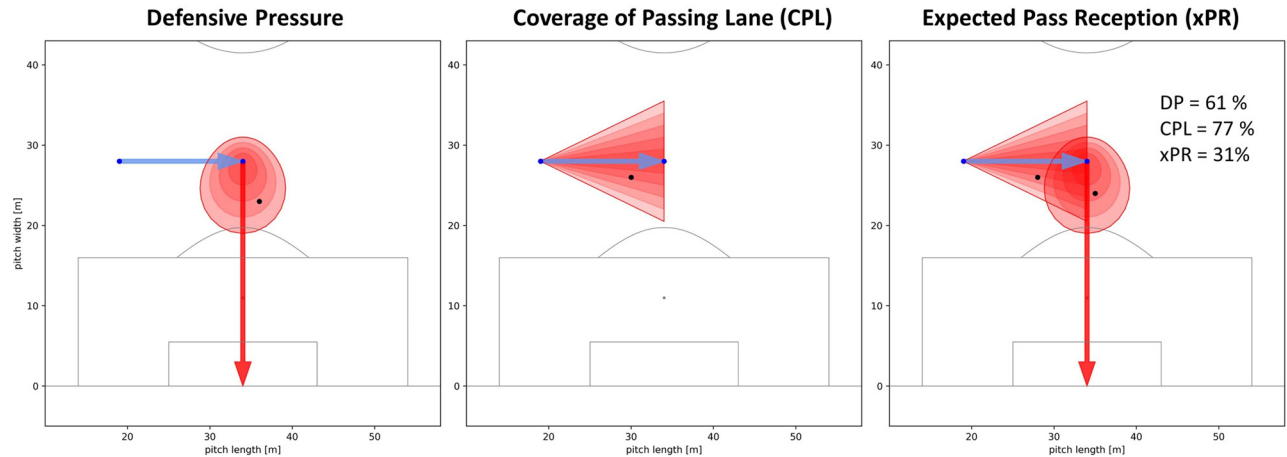


Figure 1: Illustration of the Expected Pass Reception (xPR) metric (right) which is computed based on the two metrics of Defensive Pressure (DP, left) and Coverage of Passing Lane (CPL, middle). For all three illustrations, the attackers (depicted in blue) create a possible pass option from left to right (blue arrow). The defender(s) (depicted in black) put(s) the possible pass-receiving attacker under defensive pressure (61 %) and cover(s) the possible passing lane of the given pass option (77 %) which in combination results in an xPR of 31 %.

2.2.4 Features

The handcrafted features were computed in the moment of an on-ball action of the attacking team using the tracking data of the players and the ball.

Overall, the features were classified into three main categories: (1) contextual information about the match situation (including the absolute & relative ball position and match situation history), (2) offensive tactical features (including the ball control of the ball-leader, space control, pass options, and box occupation of the attacking team), and (3) defensive tactical features (including defensive pressure on the ball-leader and goalkeeper position). The detailed description of all features can be found in Table 1.

All the features were carefully selected by professional match analysts of a first-division club (LeaF & LeoF) and only included if they comprised a technical or tactical idea and thus represented a part of the technical or tactical match performance (da Costa et al. 2009; Forcher et al. 2024c). This selection process enables a practical applicable interpretation of the results to gain detailed knowledge about the match performance in soccer.

2.3 Statistics

To statistically examine the given problem, machine learning Classifiers (Logistic Regression, Random Forest, XGBoost, adaBoost) were used to predict the success of a possession during the playing phase of offensive play. The

success of an offensive playing phase was defined by the outcome of the playing sequence. To specifically define the target variable, multiple time-windows (5, 10, 15, & 20 s time-window) were evaluated, with a situation classified as successful if a goal was scored within the specified time-window of the same possession. In the final model, and consistent with prior EPV studies (Fernández et al. 2019, 2021; Forcher et al. 2025; Rahimian et al. 2024), a 10-s window was selected for EPV prediction, facilitating the identification and attribution of critical actions contributing to goal-scoring opportunities. This approach enables a focus on actions exerting a substantive influence on positive attacking outcomes, while excluding those that do not significantly influence the outcome.

To effectively train the classifiers, all features were normalized, and the dataset was split using a stratified train–test–validation split (60 %, 20 %, 20 %) on a match-by-match basis to prevent events from the same possession appearing in different data sets, thereby avoiding information leakage and overfitting (train: 158 matches, test: 53 matches, validation: 53 matches). Model training included an internal 5-fold cross-validation procedure, also stratified and conducted on a match-by-match basis. For each classifier, hyperparameters were optimized using 80 rounds of randomized grid search. Model selection and tuning prioritized performance on the Brier score, expected calibration error (ECE), and Recall. These measures of predictive performance were emphasized because the main task of the EPV

Table 2: Prediction performance of machine learning classifiers.

Description	Expected possession value (EPV)					Time-window	Target variable	
	Information			Data set distribution (over all 132,851 match situations)	Positives (successful)			
	Number of features	Algorithm	Optimization				Negatives (unsuccessful)	
Logistic regression (all features, 10 s)	17	Logistic regression	Elastic net		10 s	131,820 (99.22 %)	1,031 (0.78 %)	
AdaBoost (all features, 10 s)	17	AdaBoost	ECE, Brier score & recall		10 s	131,820 (99.22 %)	1,031 (0.78 %)	
XGBoost (all features, 20 s)	17	XGBoost	ECE, Brier score & recall		20 s	131,136 (98.71 %)	1,715 (1.29 %)	
XGBoost (all features, 15 s)	17	XGBoost	ECE, Brier score & recall		15 s	131,390 (98.90 %)	1,461 (1.10 %)	
XGBoost (all features, 10 s)	17	XGBoost	ECE, Brier score & recall		10 s	131,820 (99.22 %)	1,031 (0.78 %)	
XGBoost (all features, 5 s)	17	XGBoost	ECE, Brier score & recall		5 s	132,093 (99.43 %)	758 (0.57 %)	
RandomForest (all features, 20 s)	17	RandomForest	ECE, Brier score & recall		20 s	131,136 (98.71 %)	1,715 (1.29 %)	
RandomForest (all features, 15 s)	17	RandomForest	ECE, Brier score & recall		15 s	131,390 (98.90 %)	1,461 (1.10 %)	
RandomForest (all features, 10 s)	17	RandomForest	ECE, Brier score & recall		10 s	131,820 (99.22 %)	1,031 (0.78 %)	
RandomForest (all features, 5 s)	17	RandomForest	ECE, Brier score & recall		5 s	132,093 (99.43 %)	758 (0.57 %)	

Training set (60 %)

Performance on original training set (Ø mean & ± standard deviation in 5-fold cross validation after application of oversampling strategy)									
Accuracy	Precision	Recall	F1-score	Log-loss	Brier score	RPS	ECE	AUC	
0.97	0.033	0.783	0.062	0.451	0.145	0.145	0.289	0.869	
0.990	0.367	0.187	0.248	0.157	0.030	0.030	0.126	0.912	
(Ø 0.989 ± 0.000)	(Ø 0.274 ± 0.019)	(Ø 0.127 ± 0.019)	(Ø 0.173 ± 0.019)	(Ø 0.155 ± 0.002)	(Ø 0.029 ± 0.001)	(Ø 0.029 ± 0.001)	(Ø 0.122 ± 0.002)	(Ø 0.862 ± 0.008)	
0.988	0.715	0.099	0.175	0.061	0.011	0.011	0.014	0.786	
(Ø 0.987 ± 0.001)	(Ø 0.337 ± 0.089)	(Ø 0.033 ± 0.009)	(Ø 0.060 ± 0.016)	(Ø 0.066 ± 0.002)	(Ø 0.013 ± 0.000)	(Ø 0.013 ± 0.000)	(Ø 0.013 ± 0.001)	(Ø 0.729 ± 0.014)	
0.990	0.728	0.127	0.216	0.051	0.010	0.010	0.012	0.831	
(Ø 0.989 ± 0.000)	(Ø 0.347 ± 0.045)	(Ø 0.048 ± 0.011)	(Ø 0.084 ± 0.018)	(Ø 0.057 ± 0.002)	(Ø 0.011 ± 0.000)	(Ø 0.011 ± 0.000)	(Ø 0.012 ± 0.001)	(Ø 0.778 ± 0.014)	
0.992	0.675	0.153	0.249	0.040	0.008	0.008	0.011	0.892	
(Ø 0.991 ± 0.001)	(Ø 0.343 ± 0.054)	(Ø 0.062 ± 0.010)	(Ø 0.105 ± 0.016)	(Ø 0.046 ± 0.002)	(Ø 0.009 ± 0.000)	(Ø 0.009 ± 0.000)	(Ø 0.012 ± 0.002)	(Ø 0.849 ± 0.008)	
0.995	0.587	0.271	0.370	0.010	0.026	0.005	0.005	0.961	
(Ø 0.993 ± 0.000)	(Ø 0.323 ± 0.036)	(Ø 0.118 ± 0.009)	(Ø 0.172 ± 0.013)	(Ø 0.009 ± 0.000)	(Ø 0.029 ± 0.001)	(Ø 0.006 ± 0.000)	(Ø 0.006 ± 0.000)	(Ø 0.937 ± 0.005)	
0.987	0.484	0.204	0.287	0.078	0.013	0.013	0.045	0.933	
(Ø 0.985 ± 0.001)	(Ø 0.251 ± 0.030)	(Ø 0.090 ± 0.010)	(Ø 0.133 ± 0.014)	(Ø 0.090 ± 0.002)	(Ø 0.016 ± 0.001)	(Ø 0.016 ± 0.001)	(Ø 0.044 ± 0.001)	(Ø 0.731 ± 0.017)	
0.989	0.469	0.251	0.327	0.068	0.012	0.012	0.041	0.949	
(Ø 0.986 ± 0.001)	(Ø 0.247 ± 0.032)	(Ø 0.117 ± 0.012)	(Ø 0.158 ± 0.017)	(Ø 0.079 ± 0.001)	(Ø 0.015 ± 0.000)	(Ø 0.015 ± 0.000)	(Ø 0.040 ± 0.001)	(Ø 0.781 ± 0.006)	
0.990	0.421	0.350	0.382	0.055	0.011	0.011	0.035	0.972	
(Ø 0.988 ± 0.001)	(Ø 0.219 ± 0.028)	(Ø 0.163 ± 0.022)	(Ø 0.187 ± 0.024)	(Ø 0.066 ± 0.002)	(Ø 0.013 ± 0.000)	(Ø 0.013 ± 0.000)	(Ø 0.036 ± 0.001)	(Ø 0.859 ± 0.010)	
0.990	0.312	0.621	0.416	0.025	0.038	0.009	0.009	0.990	
(Ø 0.987 ± 0.001)	(Ø 0.176 ± 0.003)	(Ø 0.312 ± 0.010)	(Ø 0.225 ± 0.003)	(Ø 0.025 ± 0.001)	(Ø 0.045 ± 0.002)	(Ø 0.011 ± 0.001)	(Ø 0.011 ± 0.001)	(Ø 0.942 ± 0.004)	

Table 2: (continued)

		Test set (20 %)							
		Performance (of best estimator on test set)							
Accuracy	Precision	Recall	F1-score	Log-loss	Brier score	RPS	ECE	AUC	
0.793	0.032	0.743	0.062	0.456	0.147	0.147	0.291	0.856	
0.989	0.312	0.143	0.196	0.160	0.031	0.031	0.126	0.849	
0.986	0.385	0.040	0.073	0.070	0.014	0.014	0.012	0.720	
0.988	0.373	0.046	0.082	0.059	0.011	0.011	0.011	0.767	
0.990	0.322	0.057	0.096	0.046	0.009	0.009	0.011	0.849	
0.994	0.367	0.166	0.228	0.010	0.028	0.006	0.006	0.942	
0.984	0.315	0.105	0.157	0.093	0.017	0.017	0.043	0.725	
0.987	0.298	0.117	0.168	0.081	0.015	0.015	0.041	0.777	
0.988	0.257	0.182	0.213	0.068	0.014	0.014	0.036	0.851	
0.987	0.180	0.374	0.244	0.026	0.045	0.011	0.011	0.941	

model is determined to correctly predict successful possessions leading to goals (Recall), while simultaneously minimizing the Brier score and ECE to ensure well-calibrated probability estimates of scoring probabilities.

Furthermore, the Pearson correlation between the features was analyzed and all features that very largely correlated (Pearson’s $r > 0.75$, very large correlation (Hopkins 2002)) were excluded (see Appendix 2). (excluded features: absolute_pitch_position_third, match_situation_number_actions, offense_onball_ball_control_speed_difference_player_ball, offense_offball_pass_options_close_to_ball_5_mean).

For the final selected model, SHAP values were calculated to explain feature importance and their specific contributions to the predictions to enable a transparent interpretation of the model’s characteristics.

3 Results

Overall, in all 264 matches considered, 18,060 possessions of the playing phase of offensive play were identified. During those possessions 132,851 match situations were analyzed ($\emptyset = 7.36$ on-ball actions per possession).

Increasing the time-window for defining the target variable of possession success resulted in a greater proportion of match situations being classified as successful, ranging from 0.57 % with a 5-s window to 1.29 % with a 20-s window. With the finally selected target variable definition of a 10-s time-window, a total of 1,031 (0.78 %) match situations were identified as successful attacking play ($\hat{=}$ a goal was scored in the following 10 s of the match situation in the same possession). All detailed class distributions and prediction performances of the different models are presented in Table 2 (all feature correlations are displayed in Appendix 2, Figure 1).

For the selected 10 s time-window, the XGBoost Classifier achieved the best calibration scores (Brier score = 0.009, ECE = 0.011) with solid discrimination between classes (AUC = 0.849, Recall = 0.057). The Random Forest Classifier showed increased Recall (0.182) but decreased calibration performance (Brier score = 0.015, ECE = 0.041). All detailed predictive performance measures for all models are illustrated in Table 2.

Overall, the XGBoost Classifier using the 10 s time-window was analyzed in more detail using SHAP values and feature importances (see Figure 2). The most important features for the prediction were the distance (1st) and angle (2nd) to the goal, the offensive space control in the final third (3rd), and the relative pitch position (outplayed opposing formation lines, 4th) (Table 3).

Table 3: Hyperparameter of machine learning classifiers.

Hyperparameters												
Selected value of best estimator (tuning range)												
Description	smote sampling_strategy	smote threshold	n_estimators	max_depth	min_sample_split	min_sample_leaf	max_features	bootstrap	class_weight	learning_rate	subsample	max_leaf_nodes
Logistic regression (all features, 10 s)												
AdaBoost (all features, 10 s)	0.5 (0.3/ 0.5/0.8)	0.0001 (0.0/0.0001)		6 (1–6)	2 (2–6 step 2)	2 (2–6 step 2)				0.01 (0.01/0.05/ 0.1/0.2/0.5)		
XGBoost (all features, 20 s)	0.3 (0.3/ 0.5/0.8)	0.0001 (0.0/0.0001)	270 (250–500 step 10)	19 (5–19 step 2)	4 (2–6 step 2)	6 (2–6 step 2)	0.5 (sqrt/0.5)			0.2 (0.05/0.1/0.2)	0.8 (0.6/0.8)	11 (5–11 step 2)
XGBoost (all features, 15 s)	0.3 (0.3/ 0.5/0.8)	0.0001 (0.0/0.0001)	340 (250–500 step 10)	7 (5–19 step 2)	2 (2–6 step 2)	6 (2–6 step 2)	0.5 (sqrt/0.5)			0.2 (0.05/0.1/0.2)	0.8 (0.6/0.8)	11 (5–11 step 2)
XGBoost (all features, 10 s)	0.3 (0.3/ 0.5/0.8)	0.0001 (0.0/0.0001)	350 (250–500 step 10)	17 (5–19 step 2)	4 (2–6 step 2)	2 (2–6 step 2)	sqrt (sqrt/0.5)			0.2 (0.05/0.1/0.2)	0.8 (0.6/0.8)	11 (5–11 step 2)
XGBoost (all features, 5 s)	0.3 (0.3/ 0.5/0.8)	0.0001 (0.0/0.0001)	310 (250–500 step 10)	17 (5–19 step 2)	4 (2–6 step 2)	2 (2–6 step 2)	sqrt (sqrt/0.5)			0.2 (0.05/0.1/0.2)	0.8 (0.6/0.8)	11 (5–11 step 2)
RandomForest (all features, 20 s)	0.3 (0.3/ 0.5/0.8)	0.0001 (0.0/0.0001)	140 (20–200 step 10)	13 (5–15 step 2)	4 (2–10 step 2)	2 (2–10 step 2)	sqrt (sqrt/log2)	1 (1)	None (balanced/ balanced_subsample/ none)			
RandomForest (all features, 15 s)	0.3 (0.3/ 0.5/0.8)	0.0001 (0.0/0.0001)	140 (20–200 step 10)	13 (5–15 step 2)	4 (2–10 step 2)	2 (2–10 step 2)	sqrt (sqrt/log2)	1 (1)	None (balanced/ balanced_subsample/ none)			
RandomForest (all features, 10 s)	0.3 (0.3/ 0.5/0.8)	0.0001 (0.0/0.0001)	140 (20–200 step 10)	13 (5–15 step 2)	4 (2–10 step 2)	2 (2–10 step 2)	sqrt (sqrt/log2)	1 (1)	None (balanced/ balanced_subsample/ none)			
RandomForest (all features, 5 s)	0.3 (0.3/ 0.5/0.8)	0.0001 (0.0/0.0001)	140 (20–200 step 10)	13 (5–15 step 2)	4 (2–10 step 2)	2 (2–10 step 2)	sqrt (sqrt/log2)	1 (1)	None (balanced/ balanced_subsample/ none)			

4 Discussion

This paper aimed to combine the possibilities of machine learning with the substantial information of player tracking data to analyze attacking success in elite soccer. The developed expected possession value (EPV) was modelled using carefully selected KPIs with a specified tactical or technical idea. Consequently, the final selected XGBoost model (all features, 10 s time-window) demonstrated satisfactory predictive performance in identifying successful attacks while providing well-calibrated probability estimates. Together with the implementation of explainable AI methods (see Figures 2 and 3) a detailed interpretation of features and underlying mechanisms of the prediction model ensures a high interpretability of the presented approach. This interpretability of the inner mechanisms of the EPV model is essential to gain deep insights into the dependencies of tactical match performance of soccer players. Thereby, detailed and practice-relevant knowledge about the complexities of achieving attacking success in the Bundesliga is revealed. In detail, the presented results suggest that the most critical factors for scoring are gaining possession near the opponent's goal, breaking through opposing formation lines, and positioning attackers inside the opponent's penalty box or final third.

The analysis of different time-windows for the definition of the target variable of possession success showed that the application of larger time-windows result in more situations labeled as successful (ranging from 0.57 % for the 5 s time-window to 1.27 % of match situations

labeled as successful for the 20 s time-window, see Table 2). However, substantial class imbalance persists across all tested time-windows. For both XGBoost and Random Forest Classifiers, the prediction performance, especially the identification of successful match situations, improves with smaller time-windows (e.g. increased Recall & Accuracy). This indicates that the outcome of match situations are easier to predict the closer they are to the end of the possession illustrating a strong connection between the last actions of an attack and the possession outcome. Balancing predictive performance with the ability to attribute value to preceding actions and in line with prior EPV studies (Fernández et al. 2021; Forcher et al. 2025; Rahimian et al. 2024) the 10-s window was selected for subsequent analyses.

Overall, the presented EPV model shows a powerful prediction performance of successful attacks. In comparison to xG models, the presented EPV model shows comparable discrimination (presented EPV AUC = 0.85; Anzer and Bauer (2021) xG AUC = 0.83) and decreased sensitivity in terms of correctly predicting goals (presented EPV Recall: 0.05; Anzer and Bauer (2021) xG Recall = 0.18; Cavus and Biecek (2022) xG Recall = 0.30). This worse performance in identifying successful match situations may be traced back to the higher imbalance of the dataset in EPV (between 0.57 % & 1.29 % of successful labeled match situation in the presented study) compared to xG (12.52 % of successful shots in the analyzed Bundesliga season 2022/23: 7,758 shots, 971 goals). However, the calibration of the selected EPV model indicates a substantial performance in predicting accurate

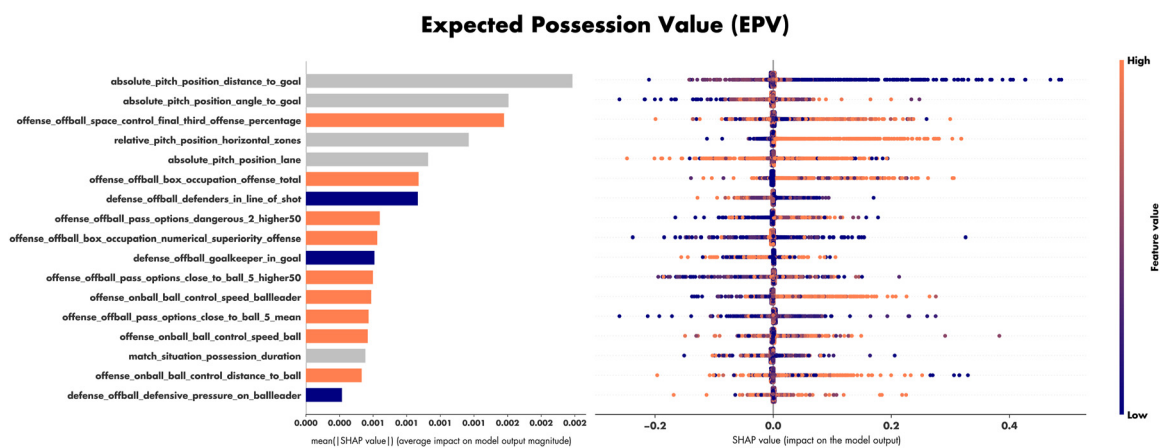


Figure 2: Illustration of the feature importances (left) and SHAP values (right) of the 17 included features within the selected EPV model (XGBoost, all features & 10 second time-window). Feature importances on the left are illustrated according to their category (match situation: grey, offense: orange, defense: blue). The feature expressions are depicted on the right. For the SHAP values on the right, orange indicates high feature values which positively influence the prediction (higher scoring probability) and blue indicates a low feature value with negative influence on the prediction (lower scoring probability).

MATCH SITUATION

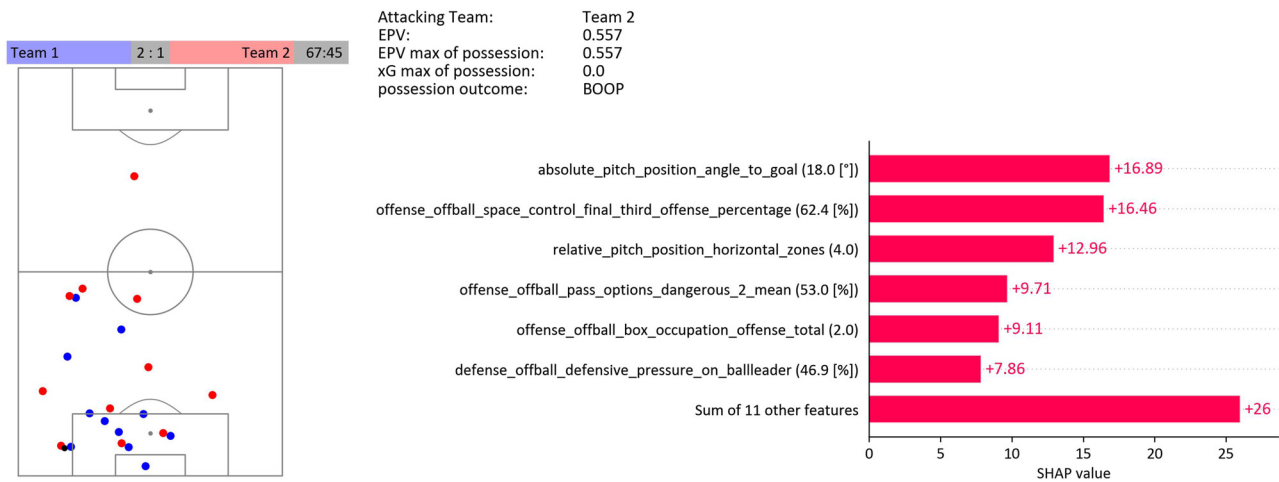


Figure 3: Illustration of an exemplary EPV analysis of a match situation (left) in the 68th minute with the home team (blue) leading with a current score of 2:1. The away team (red) attacking from the right lane (ball depicted in black) which results in an EPV of 56 %. Additionally, the SHAP values of this single prediction are represented (depicted in red on the bottom right), with the outplayed opposing defensive formation lines being the most important variable in this specific match situation.

probability estimates (Brier score = 0.009, Cavus and Biecek (2022) xG Brier score = 0.073). Accordingly, the prediction performance of the presented model is substantial. Still, the highly unbalanced nature of the dataset and the amount of uncertainty in possession outcomes in soccer needs to be considered when implementing this model in practice to analyze match performances and support decision-making.

While the comparison with xG models gives a first insight of the overall performance of our EPV model, comparisons between EPV models and xG models have to be made with caution since the models predict different match situations (xG: success of shots, EPV: success of whole attacks). Nevertheless, they are indispensable, since benchmarking the presented EPV model against other models for attacking success in the literature (e.g. EPV studies) is unfeasible due to the absence of adequate reported prediction performance metrics. To the best of the authors' knowledge, there is no comparable attacking success prediction study (e.g. EPV, xT, VAEP) that analyzed the feature importances or individual feature expressions of an attack success prediction model (Decroos et al. 2019; Fernández et al. 2019, 2021; Rahimian et al. 2024) in a way that would enable such a detailed comparison with the presented approach. Therefore, the presented results need to be validated in future studies.

In comparable EPV studies, log-loss is mainly reported as a prediction performance metric (Fernández et al. 2019, 2021; Rahimian et al. 2024). While this measure indicates the overall accuracy of the prediction approach, it does not

provide important information about the performance in predicting the minority class (in this case: goals). In this special case of predicting goals as successful outcome of an attack, the main objective of an attacking success model (in this case EPV) should be about predicting the minority class (goals or successful attacks) next to the other goal of producing well-calibrated probability estimates (see Section 2 Methods Statistics). In this context, a model that solely predicts the majority class of unsuccessful match situations would result in a high accuracy and almost perfect log-loss in the highly unbalanced data set ($\approx 99\%$ unsuccessful attacks vs. $\approx 1\%$ successful attacks). However, in doing so, it would fail to address the main objective of the given problem, to correctly predict goals, and would consequently not be helpful in assessing offensive match performance. This is also true for the sole reporting of AUC in other attacking success predictions such as VAEP (presented EPV AUC = 0.85; Decroos et al. 2019 VAEP AUC = 0.77).

Accordingly, a valid decision regarding which model is best suited for predicting the problem of successful attacks ultimately depends on the perspective taken. However, due to the lack of important prediction performance metrics in most studies in this context, it is not legitimate to make a definitive judgment on which model is best for predicting successful attacks at this stage. Understanding the detailed prediction performance of a model, particularly in predicting the minority class of goals and producing well-calibrated probability estimates, is essential for evaluating the uncertainty of such an approach accurately.

This exact knowledge of uncertainty, along with high interpretability (e.g. SHAP values, see Figures 2 and 4), is essential to effectively implement these models in performance analysis and ultimately support decision-making in elite soccer. This is the juncture where the presented EPV approach sets itself apart from other models of attacking success in the current literature.

By carefully selecting the features that map a specified aspect of match performance in soccer (see Table 1) and applying explainable AI (see Figure 2) the explainability of the presented approach is ensured. Based on that, the individual features are analyzed in detail below. With it, the inside mechanisms of the prediction can be interpreted and derived to summarized principles of play which contributes to the objective analysis of players' and teams' performances.

First of all, the most important feature of the presented EPV model is the distance of the ball towards the attacked goal (see Figure 2). This finding is in line with xG studies that similarly found this feature to be the most important in predicting goals (Anzer and Bauer 2021; Mead et al. 2023). Accordingly, this confirms the simple basic assumption that bringing the ball closer to the opponent's goal increases the probability of scoring in the respective possession.

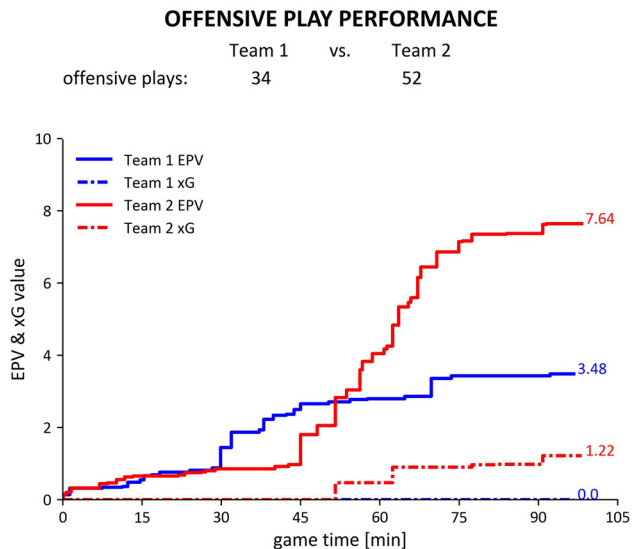


Figure 4: Illustration of an exemplary EPV analysis of a single match. The home team (blue) had 34 offensive plays in the respective match, in which they created an EPV sum of 3.48 (blue solid line). They especially dominated the away team (red) during the first half (EPV in solid lines). In contrast, the away team outperformed the home team in their offensive play performance especially during the second half while increasing their danger to an overall sum EPV of 7.64 (red solid line). Furthermore, the sum xG values are presented in dotted lines in which the away team also outperformed the home team during the second half.

Complementing this simple principle, the 4th most important feature, the relative pitch position (with respect to the opposing defensive formation lines), for the prediction provides associated information. This feature can complement the information of the distance to the goal by indicating how far the attack has progressed. For instance, in match situations with a high-pressing opponent, this feature indicates that the danger of an attack increases when the defensive formation lines are overlapped even though the ball is still far away from the opposing goal. Accordingly, outplaying opposing defensive formation increases the scoring probability (e.g. by bringing the ball behind the defensive line, see Figure 3). This may be due to a disruption of the opponent's defensive organization following a line-breaking pass which can then open up space for possible passing options (Forcher et al. 2024a; Goes et al. 2018).

In line with previous xG studies, the angle to the goal is also important for the prediction (2nd important feature in the presented EPV model, 5th important variable in Anzer and Bauer (2021)). In detail, the probability of a successful attack is increased if the ball is situated in the middle of the pitch with greater angle to the goal ($0^\circ \triangleq$ ball is directly beneath goal, $90^\circ \triangleq$ ball is in the middle of the pitch directly in front of goal) (see Figure 2).

The 3rd most important variable for the prediction of attacking success is the space control of the attacking team in the final third. By placing potentially dangerous attackers close to the opposing goal the space control of the attacking team in the final third may be increased. Therefore, creating space for possible dangerous pass options close to the goal (8th important feature of the model) is a key to create scoring opportunities in areas close to the goal (Fernández et al. 2018).

Furthermore, the offensive box occupation is important for the EPV model (5th important feature). In detail, the positioning of a higher number of attackers in the opposing box increases the probability of scoring a goal. In this context, it was shown that almost 90 % of all goals are scored within the penalty box (Fischer and Toetz 2024). Therefore, positioning possible pass receivers inside the box increases the chance of scoring in the following. For instance, in the match situation of Figure 3, two attackers are positioned inside the penalty box which increases the numerical ratio of the offense (11th most important feature). Both players may become a pass option of a cross from the ball-leading player in the outer lane of the pitch which overall increases the probability of scoring.

The features discussed are in line with recent findings in the literature and the formulated hypotheses (see Table 1, Technical & Tactical Idea).

4.1 Practical application

Next to the aim of gaining knowledge about the match performance in soccer, one of the main aims of sports scientific analyses is to apply the findings in practice. The outcomes of the EPV model can be used in two main ways.

First, the conducted interpretation of features and their expressions (see above) can be used to derive principles of offensive play. Those principles of play are based on the findings of the developed EPV model and therefore hold important objective information on the success factors of the playing phase of offensive play. They are a suitable tool to translate the presented findings to applicable guidelines for the practical use of coaches or analysts. For instance, the principles can be used to design training regimes to potentially increase the match performance of players.

Objective principles of offensive play that increase the scoring probability of an attack:

- Bring the ball close to the opponent's goal!
- Overplay opposing formation lines!
- Create space in the opposing final third!
 - e.g. by threatening deep spaces behind the opposing defending line with deep runs
 - e.g. by occupying the opposing final third
- Occupy the opposing box!
- Create dangerous pass options close to the opposing goal!

Second, next to the formulation of principles of play, the quantitative results of the EPV model can be used to analyze individual player and collective team performances. For instance, individual actions can be evaluated by analyzing the amount of EPV gained by a player's action (Fernández et al. 2019). With it, the contribution of each individual player to the attacking success of a team (e.g. by advancing the ball into dangerous areas by dribbling or finding a teammate in a dangerous scoring position by playing a key pass) can be assessed in seconds. In performance analysis of offensive play, this information is key to objectively evaluate players and their actions. In this context, EPV added from individual actions (e.g., dribbles or passes) can be aggregated across matches or seasons to quantify a player's overall performance and to identify the most impactful players in a given competition. An applied example of this individual player analysis is provided in (Forcher et al. 2025).

Furthermore, the offensive performance of teams can be analyzed in individual match situations, whole matches, or entire seasons by analyzing the danger that was created during the offensive plays (quantified by EPV). An application of the EPV model to analyze a specific match situation

is illustrated in Figure 3. In this context, the analysis of the SHAP values for this particular situation (shown in the bottom right of Figure 3) assists soccer analysts and coaches in understanding the underlying mechanisms of the prediction, highlighting the features that had the most significant impact on the prediction outcome in this match situation. This information about how teams create dangerous match situations can be used by the experts in a targeted manner to improve strategy and decision-making in opponent analysis (pre-game) or post-match analysis of the own team. Overall, this explainability is an absolute key to understand machine learning models to effectively apply them in performance analysis.

In addition to the detailed information given by the analysis of single match situations, the EPV performance throughout an entire match can provide the coaching staff with valuable insight into their overall attacking performance. An example of such a match analysis is shown in Figure 4. In this context, the analysis of the xG match performance has become standard in soccer analysis, providing a first indication of the offensive performance of a team resulting from shots at goal. However, such analyses lack insights into the overall attacking performance missing important information about dangerous attacks that did not entail a shot at goal (e.g. near miss from a flat cross in front of an empty net). Therefore, this xG match analysis can be enriched with insights from the presented EPV model, which can highlight attacking performance without having to shoot at goal (see Figure 4 in red dominated the first half despite not taking any shots ($xG = 0$)).

In summary, EPV can provide valuable information on team and individual performances and thereby complement established metrics such as expected goals (xG) (Anzer and Bauer 2021), particularly when aggregated at the team or season level. In our analyses, season-level EPV showed a slightly stronger correlation with the total number of goals scored ($r = 0.92$ in offensive play) than xG ($r = 0.91$ in offensive play), indicating its suitability for longer-term performance assessment. All detailed correlations can be found in Appendix 3. On a match-by-match basis, however, xG exhibited higher correlations with immediate goal outcomes (xG $r = 0.59$ in offensive play; EPV $r = 0.27$ in offensive play), suggesting greater sensitivity to short-term fluctuations. Furthermore, EPV and xG show a very high correlation at the season level ($r = 0.97$ in offensive play), while this correlation notably decreases at the match level ($r = 0.55$ in offensive play). In this context, a recent comparative analysis between EPV and xG has shown that EPV is more useful for pre-match predictions, whereas xG provides greater utility in post-match analysis scenarios

(Forcher et al. 2025c). Overall, EPV represents a meaningful KPI for both team-level aggregation and fine-grained analysis of individual actions and match situations, thereby offering a holistic perspective on offensive performance.

4.2 Limitations and future research

This paper features limitations that should be taken into account when interpreting the results. First, for the analysis of a match situation solely the current positions of the players were taken into account. The players' body orientation, running direction, and speed (except for the ball-leading player and the ball) were not considered. This complementary information may increase the accuracy of prediction.

Additionally, in contrast to other attacking success predictions (Decroos et al. 2019) no contextual factors, such as match status, venue, or quality of opposition were considered in the presented analysis. This procedure was selected to reveal general insights into offensive match performance, valid across various teams with diverse playing styles and under different match circumstances (venue, scoreline). From a match analysis point of view, those contextual factors do not change the outcome of an objective tactical analysis in a match situation scenario. For instance, an objective analysis of similar match situations should provide the same output no matter what the current score line is to effectively evaluate performances. Following this idea, the EPV can subsequently be used to analyze the influence of contextual factors on offensive match performance in different contextual situations (EPV performance playing at home against a top team when leading).

While this paper analyzed a substantial dataset covering an entire season of one of Europe's elite soccer leagues, it only focused on one league. An application of the current approach to other leagues (e.g. Premier League) could make the approach even more universal.

Further, fixed time-windows (e.g. 10-s time-window) were used to define the success of an offensive playing phase. Events in possessions ending earlier without a scored goal were labeled as negative outcomes, which may slightly influence the estimated probabilities. Future work could consider possession-length adjusted time-windows.

In the end, one may argue that this paper solely analyzed the playing phase of offensive play. However, this is one of the main strengths of this approach since most attacking success predictions did not differ between playing phases (Fernández et al. 2019, 2021). Match situations during the different playing phases (offensive play, defensive play, offensive transition, defensive transition, set plays) differ significantly and demand different tactical behavior from the players. For instance, during an offensive corner kick,

the attacking players need to perform other actions compared to an offensive transition. Accordingly, the analysis of the tactical behavior needs to account for those match situations to be able to effectively analyze and evaluate the tactical match performance. Still, there is a need for similar attacking success metrics for offensive play or offensive set pieces. Therefore, future studies should focus on the creation of adequate features that display a part of technical or tactical match performance in those specific match situations to eventually increase the knowledge about tactical match performance in soccer.

5 Conclusions

This paper presented one of the first EPV approaches in soccer that focussed on the explainability and interpretability of the presented model. Together with the careful selection of features using expert knowledge of professional match analysts and the reflection of the defensive side of the game, detailed insights were presented on the offensive match performance of players and teams in the Bundesliga. The most important features to predict the scoring probability of possessions were the distance to the goal, the number of outplayed opposing formation lines, the offensive box occupation, and the space control in the final third. Applying the developed EPV model practice-oriented case studies provided insights into the dependencies of attacking success which includes the insightful analysis of match situations and matches (see Figures 3 and 4) and the derivation of principles of offensive play (see Section 4 Discussion Practical application). With those findings and the EPV model itself, various decision-making and analysis processes in professional soccer can be supported.

Overall, by analyzing a whole season of Bundesliga tracking data with cutting-edge machine learning approaches this paper gave valuable insights into the dependencies of attacking success in soccer.

Acknowledgments: The authors thank the German Football League (Deutsche Fußball Liga, DFL) for providing the match data used in this study.

Research ethics: This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the local ethics committee (Human and Business Sciences Institute, Saarland University, Germany, identification number: 22–02, 10 January 2022).

Informed consent: Not applicable.

Author contributions: Dr. Leander Forcher: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Writing – Original Draft.

Dr. Leon Forcher: Conceptualization, Methodology, Validation, Resources, Writing – Review & Editing. **Dr. Stefan Altmann:** Methodology, Validation, Resources, Writing – Review & Editing, Supervision. **Prof. Dr. Alexander Woll:** Conceptualization, Resources, Writing – Review & Editing, Supervision.

Use of Large Language Models, AI and Machine Learning Tools: To develop an EPV model, machine learning Classifiers (Logistic Regression, Random Forest, XGBoost, adaBoost) were used based on python 3.11.8 and scikit-learn libraries.

Conflict of interest: The authors report there are no competing interests to declare.

Research funding: Not applicable.

Data availability: The used data is property of the German Football League (Deutsche Fußball Liga, DFL) and is not publicly available. The authors do not have permission to share the data publicly. This work can be reproduced using similar data from professional soccer (e.g. tracking and event data of other soccer leagues).

Appendix 1: Expected pass reception metric

Defensive pressure (DP)

To quantify DP an improved version of the approach of Anrienko et al. (2017) was used which was adapted by Herold et al. (2022).

- (1) $L = D_{\text{back}} + (D_{\text{front}} - D_{\text{back}})(z^3 + 0.3z)/1.3$
- (2) $z = (\cos\alpha + 1)/2$
- (3) $D_{\text{front}} = (\text{GoalDis} * 0.02) + 6.9$
- (4) $D_{\text{back}} = D_{\text{front}} / 3$
- (5) $q = 1.75$
- (6) $\text{Pr} = (1 - d/L)^q \times 100 \%$

$L \triangleq$ Shape of the pressure area

$\alpha \triangleq$ Angle between defender and attacker

$D_{\text{front}} \triangleq$ Length of pressure area in front of attacker (in threat direction)

$D_{\text{back}} \triangleq$ Length of pressure area behind attacker (in opposite direction to threat direction)

$\text{GoalDis} \triangleq$ Distance of the player in ball possession to the goal, if attacker enters penalty box ($\text{GoalDis} = \text{GoalDis}/2$) to further decrease the pressure area

$q \triangleq$ Weight of pressure increase from 0 to 100 [%] inside the pressure area

$d \triangleq$ Distance from defender to attacker

$\text{Pr} \triangleq$ Pressure value in [%]

Coverage of passing lane (CPL)

To quantify CPL a novel approach was developed. This approach quantifies the extent to which a pass line can be covered by a defender.

$$(1) s = 0.5 * (a + b + c)$$

$$(2) h = 2/s * \sqrt{(s * (s - a) * (s - b) * (s - c))}$$

$$(3) d = \sqrt{(a^2 - h^2)}$$

$$(4) h_{\text{max}} = d * 0.5$$

$$(5) q = 1.75$$

$$(6) \text{CPL} = (1 - h/h_{\text{max}})^q * 100$$

$h \triangleq$ Shortest distance of defender to passing lane (interception point)

$d \triangleq$ Distance of interception point to passer

$h_{\text{max}} \triangleq$ Size of CPL area

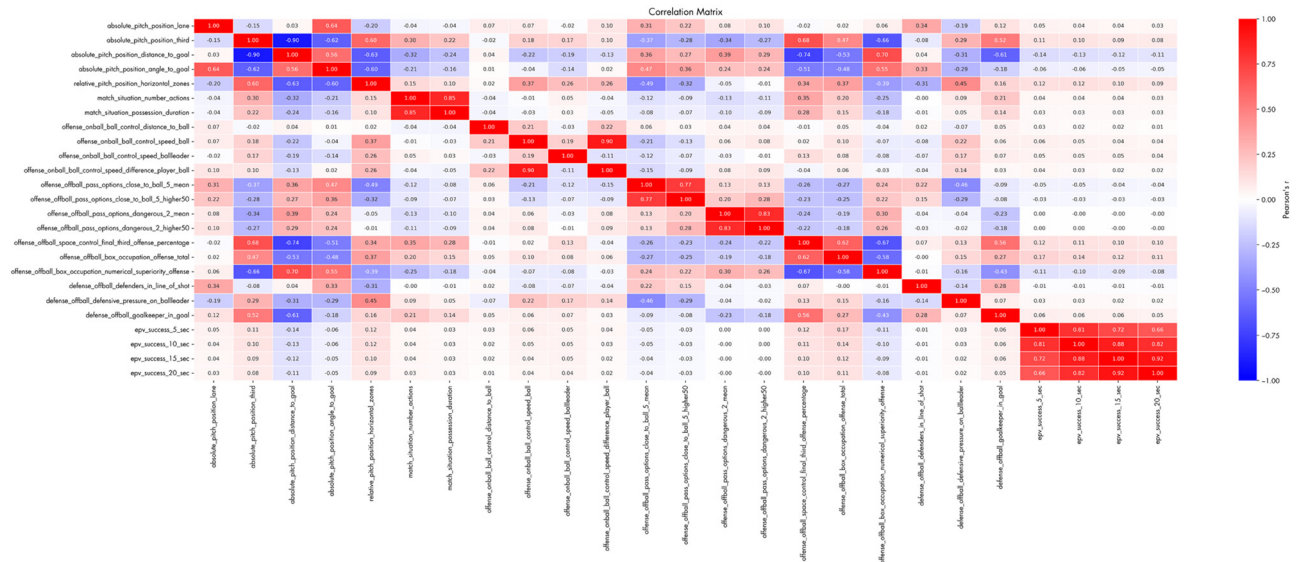
$q \triangleq$ Weight of CPL increase from 0 to 100 [%] inside the CPL area

$\text{CPL} \triangleq$ CPL value in [%]

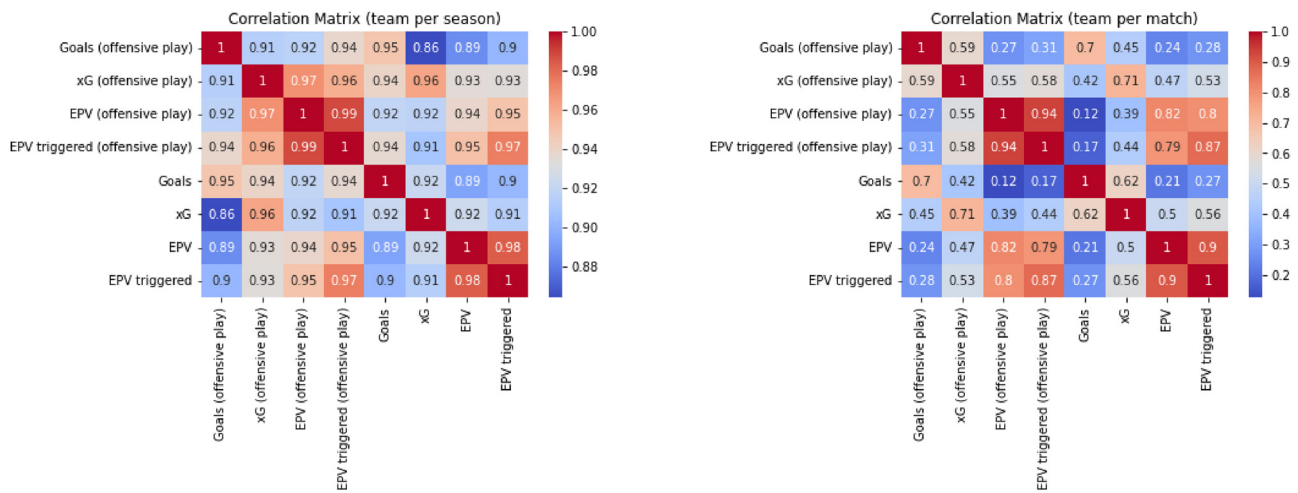
Expected pass reception metric (xPR)

$$(1) \text{xPR} = 100 - (\text{DP}/2 + \text{CPL}/2)$$

Appendix 2: Correlation matrix of features and different target variables (5, 10, 15, 20 s time-window)



Appendix 3: Correlation matrix of EPV (expected possession value) and xG (expected goals)



Appendix 1 Figure 1: Correlation matrices illustrating the associations between the expected possession value (EPV), EPV triggered (defined as number of possessions with EPV greater than 0.5), expected goals (xG), and goals scored in the 2022/23 Bundesliga season. The KPIs are reported for overall match performance and for performance specifically during the offensive playing phase. The left panel presents correlations based on season-level aggregates, while the right panel presents correlations based on match-level aggregates.

References

- Andrienko, G., Andrienko, N., Budziak, G., Dykes, J., Fuchs, G., von Landesberger, T., and Weber, H. (2017). Visual analysis of pressure in football. *Data Min. Knowl. Discov.* 31: 1793–1839.
- Anzer, G. (2022). *Large scale analysis of offensive performance in football – using synchronized positional and event data to quantify offensive actions, tactics, and strategies*, Ph.D. thesis. Universität Tübingen, Available at: <https://publikationen.uni-tuebingen.de/xmlui/handle/10900/124678>.
- Anzer, G. and Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Front. Sports Act. Living* 3: 1–15.
- Bauer, P. (2022). *Automated detection of complex tactical patterns in football – using machine learning techniques to identify tactical behavior*, Ph.D. thesis. Universität Tübingen, Available at: <https://publikationen.uni-tuebingen.de/xmlui/handle/10900/124679>.
- Bauer, P., Anzer, G., and Shaw, L. (2023). Putting team formations in association football into context. *J. Sports Anal.* 9: 39–59.
- Cavus, M. and Biecek, P. (2022). Explainable expected goal models for performance analysis in football analytics. In: *2022 IEEE 9th international conference on data science and advanced analytics (DSAA)*. IEEE, pp. 1–9, Available at: <https://ieeexplore.ieee.org/abstract/document/10032440/> (Accessed 2 April 2024).
- da Costa, I.T., da Silva, J.M.G., Greco, P.J., and Mesquita, I. (2009). Tactical principles of soccer: concepts and application. *Motriz: Rev. Educ. Fis.* 15: 657–668.
- Decroos, T., Bransen, L., Van Haaren, J., and Davis, J. (2019). Actions speak louder than goals: valuing player actions in soccer. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '19: the 25th ACM SIGKDD conference on knowledge discovery and data mining*. ACM, Anchorage, AK, USA, pp. 1851–1861.
- DFL (2014). *Definitionskatalog Offizielle Spieldaten (Definitions for Official Gama Data)*, DFL: *Definitionskatalog Offizielle Spieldaten*, Available at: https://www.bundesliga.com/de/bundesliga/news/dfl-definitionskatalog-offizielle-spieldaten_0000265687.jsp (Accessed 5 December 2022).
- DFL (2024). *Bundesliga match plan season 23/24*, Available at: <https://www.bundesliga.com/de/bundesliga/spieltag/2023-2024/1>.
- Escher, T. (2020). *Der Schlüssel zum Spiel: wie moderner Fußball funktioniert*. Rowohlt Taschenbuch Verlag, Hamburg, Available at: <https://rds-blb.ibs-bw.de/link?kid=1665734698>.
- Fernández, J., Bornn, L., and Fraser, S. (2018). Wide open spaces: a statistical technique for measuring space creation in professional soccer. *Proceedings of the 12th Annual MIT Sloan Sports Analytics Conference*, Boston, MA.
- Fernández, J., Bornn, L., and Cervone, D. (2019). Decomposing the immeasurable sport: a deep learning expected possession value framework for soccer. In: *13th MIT sloan sports analytics conference*, Available at: http://www.lukebornn.com/papers/fernandez_sloan_2019.pdf.
- Fernández, J., Bornn, L., and Cervone, D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Mach. Learn.* 110: 1389–1427.
- Fischer, M. and Toetz, C. (2024). Die Tore der Champions League 2018/2019 – Statistische Analyse und Auswertung aller 366 Turniertreffer, Available at: <https://www.dfb-akademie.de/die-tore-der-champions-league-2018-2019/-/id-11009193/>.
- Forcher, L. (2024a). *Success factors in soccer defense – match analysis in soccer based on positional tracking data*, Ph.D. thesis. Karlsruhe Institute of Technology (KIT).
- Forcher, L. (2024b). *The beautiful game unveiled – the influence of tactical factors on soccer match performance*, Ph.D. thesis. Karlsruhe Institute of Technology (KIT).
- Forcher, L., Kempe, M., Altmann, S., Forcher, L., and Woll, A. (2021). The “hockey” assist makes the difference – validation of a defensive disruptiveness model to evaluate passing sequences in elite soccer. *Entropy* 23: 1607–1619.
- Forcher, L., Altmann, S., Forcher, L., Jekauc, D., and Kempe, M. (2022a). The use of player tracking data to analyze defensive play in professional soccer – a scoping review. *Int. J. Sports Sci. Coach.* 17: 1567–1592.
- Forcher, L., Forcher, L., Jekauc, D., Wäsche, H., Woll, A., Gross, T., and Altmann, S. (2022b). How coaches can improve their teams’ match performance – the influence of in-game changes of tactical formation in professional soccer. *Front. Psychol.* 13: 1–11.
- Forcher, L., Preine, L., Forcher, L., Wäsche, H., Jekauc, D., Woll, A., Gross, T., and Altmann, S. (2023). Shedding some light on in-game formation changes in the German Bundesliga: frequency, contextual factors, and differences between offensive and defensive formations. *Int. J. Sports Sci. Coach.* 18: 2051–2060.
- Forcher, L., Forcher, L., Altmann, S., Jekauc, D., and Kempe, M. (2024a). Is a compact organization important for defensive success in elite soccer? Analysis based on player tracking data. *Int. J. Sports Sci. Coach.* 19: 757–768.
- Forcher, L., Forcher, L., Altmann, S., Jekauc, D., and Kempe, M. (2024b). The keys of pressing to gain the ball – characteristics of defensive pressure in elite soccer using tracking data. *Sci. Med. Footb.* 8: 161–169.
- Forcher, L., Beckmann, T., Wohak, O., Romeike, C., Graf, F., and Altmann, S. (2024c). Prediction of defensive success in elite soccer using machine learning – tactical analysis of defensive play using tracking data and explainable AI. *Sci. Med. Footb.* 8: 317–332.
- Forcher, L., Forcher, L., and Altmann, S. (2024d). How soccer coaches can use data to better develop their players and be more successful. In: Dükling, P. and Sperlich, B. (Eds.). *Individualizing training procedures with wearable technology*. Springer International Publishing, Cham, pp. 99–123.
- Forcher, L., Forcher, L., Woll, A., and Altmann, S. (2025). AI in Bundesliga match analysis – expected possession value (EPV) vs. expected goals (xG) to predict match outcomes in soccer. *Front. Sports Act. Living* 7, 1–15.
- Gibney, E. (2022). Could machine learning fuel a reproducibility crisis in science? *Nature* 608: 250–251.
- Goes, F. (2023). *Tactical behaviour in professional soccer – the secret of successful attacks*, Ph.D. thesis. University of Groningen.
- Goes, F., Kempe, M., Meerhoff, L.A., and Lemmink, K.A. (2018). Not every pass can be an assist: a data-driven model to measure

- pass effectiveness in professional soccer matches. *Big Data* 7: 1–14.
- Goes, F., Schwarz, E., Elferink-Gemser, M., Lemmink, K., and Brink, M. (2021a). A risk-reward assessment of passing decisions: comparison between positional roles using tracking data from professional men's soccer. *Sci. Med. Footb.* 3: 1–9.
- Goes, F., Brink, M.S., Elferink-Gemser, M.T., Kempe, M., and Lemmink, K.A. (2021b). The tactics of successful attacks in professional association football: large-scale spatiotemporal analysis of dynamic subgroups using position tracking data. *J. Sports Sci.* 39: 523–532.
- Gonzalez-Rodenas, J., López-Bondia, I., Aranda-Malavés, R., Tudela Desantes, A., Sanz-Ramírez, E., and Aranda Malaves, R. (2019). Technical, tactical and spatial indicators related to goal scoring in European elite soccer. *J. Hum. Sport Exerc.* 15: 186–201.
- Headrick, J., Davids, K., Renshaw, I., Araújo, D., Passos, P., and Fernandes, O. (2012). Proximity-to-goal as a constraint on patterns of behaviour in attacker–defender dyads in team games. *J. Sports Sci.* 30: 247–253.
- Herold, M., Hecksteden, A., Radke, D., Goes, F., Nopp, S., Meyer, T., and Kempe, M. (2022). Off-ball behavior in association football: a data-driven model to measure changes in defensive pressure. *J. Sports Sci.* 40: 1412–1425.
- Hewitt, J. and Karakuş, O. (2023). A machine learning approach for player and position adjusted expected goals in football (soccer). *Frankl. Open* 4, <https://doi.org/10.1016/j.fraope.2023.100034>.
- Hewitt, A., Greenham, G., and Norton, K. (2016). Game style in soccer: what is it and can we quantify it? *Int. J. Perform. Anal. Sport* 16: 355–372.
- Hopkins, W.G. (2002). Dimensions of research. *Sportscience* 6 [Preprint].
- Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4: 1–12.
- Laakso, T., Travassos, B., Liukkonen, J., and Davids, K. (2017). Field location and player roles as constraints on emergent 1-vs-1 interpersonal patterns of play in football. *Hum. Mov. Sci.* 54: 347–353.
- Linke, D., Link, D., and Lames, M. (2020). Football-specific validity of TRACAB's optical video tracking systems. *PLoS One* 15: 1–17.
- Lucey, P., Bialkowski, A., Monfort, M., Carr, P. and Matthews, I., (2015). “Quality vs quantity”: improved shot prediction in soccer using strategic features from spatiotemporal data. In: *MIT sloan sports analytics conference*, Boston, United States of America, pp. 1–9, Available at: https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/5fee09c092fcd51ecf_1034_rppaper_SoccerPaper5.pdf.
- Mead, J., O'Hare, A., and McMenemy, P. (2023). Expected goals in football: improving model performance and demonstrating value. *PLoS One* 18: 1–29.
- Memmert, D., Lemmink, K.A.P.M., and Sampaio, J. (2017). Current approaches to tactical performance analyses in soccer using position data. *Sports Med.* 47: 1–10.
- Olthof, S. and Davis, J. (2025). Perspectives on data analytics for gaining a competitive advantage in football: computational approaches to tactics. *Sci. Med. Footb.*: 1–13, <https://doi.org/10.1080/24733938.2025.2533784>.
- Power, P., Ruiz, H., Wei, X., and Lucey, P. (2017). Not all passes are created equal: objectively measuring the risk and reward of passes in soccer from tracking data. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. Association for Computing Machinery (KDD '17), New York, NY, USA, pp. 1605–1613.
- Rahimian, P., Van Haaren, J., and Toka, L. (2024). Towards maximizing expected possession outcome in soccer. *Int. J. Sports Sci. Coach.* 19: 230–244.
- Rudd, S. (2011). A framework for tactical analysis and individual offensive production assessment in soccer using Markov chains. In: *New England symposium on statistics in sports*.
- Sarmiento, H., Clemente, F.M., Araújo, D., Davids, K., McRobert, A., and Figueiredo, A. (2018). What performance analysts need to know about research trends in association football (2012–2016): a systematic review. *Sports Med.* 48: 799–836.
- Singh, K. (2018). Introducing expected threat (xT), Available at: <https://karun.in/blog/expected-threat.html>.
- Sumpter, D. and Andrzejewski, A. (2022). Soccermetrics, Available at: <https://soccermetrics.readthedocs.io/en/latest/>.
- Welch, M., Schaerf, T.M., and Murphy, A. (2021). Collective states and their transitions in football. *PLoS One* 16: 1–20.