

Which indicators matter? Using performance indicators to predict in-game success-related events in association football

Steffen Lang^{1*}, Thomas Wimmer², Alexander Erben², Daniel Link^{1,3}

¹ Technical University of Munich, Germany; TUM School of Medicine and Health, Chair of Performance Analysis and Sports Informatics

² Technical University of Munich, Germany; TUM School of Computation, Information and Technology, Chair for Decentralized Information Systems and Data Management

³ Technical University of Munich, Germany; Munich Data Science Institute

Abstract

This study evaluates the predictive power of common performance indicators (PIs) in soccer for success- or scoring-related events (SREs) such as shots, corner kicks, and box entries. Using data from 102 Bundesliga matches, we applied five machine learning methods to assess how well 28 widely used PIs (e.g., passes, ball possession time, opponents outplayed) within a past time span (up to 15 minutes) predict an SRE in a future window (up to 15 minutes). We ranked PIs based on the mean Matthews Correlation Coefficient. Results show $PI_{Dangerousity}$ best predicts SRE_{Goal} and $SRE_{ShotTaken}$, while $PI_{EntriesAttaThird}$ is strongest for $SRE_{Cornerkick}$, $SRE_{EntryAttaThird}$, and $SRE_{EntryOppBox}$. $PI_{Dangerousity}$ and $PI_{SuccPassAttThird}$ consistently rank in the Top 9, highlighting their predictive strength. Combining $PI_{OutplayedOpp}$ and $PI_{TacklingsWon}$ over a five-minute input window improves goal prediction within three minutes, outperforming random guessing by 6%. PIs based on rare events, such as goals and corner kicks, are less effective for SRE prediction, whereas those capturing frequent actions (e.g., final-third possession, Dangerousity, outplayed opponents) perform better. These findings highlight the value of in-game data for short-term event prediction and its potential applications in quantifying match momentum, optimizing live betting odds, and improving performance analysis.

KEYWORDS: FOOTBALL, MACHINE LEARNING, IN-PLAY PREDICTION, MATCH MOMENTUM, PERFORMANCE INDICATORS

Introduction

Predicting the winner of a match is a well-established research topic in sports (see reviews Horvat & Job, 2020; Wunderlich & Memmert, 2021), i.e., in tennis (Kovalchik, 2016), American football (Purucker, 1996), horse racing (Davoodi & Khanteymoori, 2010), basketball (Cao, 2012) or golf (Chae, Park, & So, 2021). Predicting match outcomes is crucial for setting betting odds, researching the nature of a particular sport and for conducting performance analysis (Wunderlich & Memmert, 2021). Particularly in association football (soccer), extensive research exists seeking to predict the exact score (Angelini & Angelis, 2017), the spread (Karlis & Ntzoufras, 2009) or the outcome in terms of a win, draw, and lose (Baboota & Kaur, 2019; Beal, Middleton, Norman, & Ramchurn, 2021; Goddard & Asimakopoulos, 2004). Methodologies employed range from Poisson distribution probabilistic classifiers (Goddard, 2005), Bayesian networks (Owramipur, Eskandarian, & Mozneb, 2013) and Random Forests (Schauberger & Groll, 2018) to Artificial Neural Networks (Igiri & Nwachukwu, 2014).

While existing studies predict match outcomes, researchers could also address the question of what will happen next *during* a match – whether a team will make the next goal, shot attempt, or corner kick. These types of predictions have not yet been fully explored in research but hold substantial relevance for the betting market, where a wide range of in-play betting options generates significant financial turnover (Jackson, 2015; Killick & Griffiths, 2019; Parke & Parke, 2019). In-play predictions can complement traditional outcome predictions, offering finer granularity and thereby improving forecasting accuracy and scope. Existing studies in in-play forecasting focus on various prediction goals. e.g., some aim to predict the outcome of the match while it is ongoing (Arntzen & Hvattum, 2021; Dobson & Goddard, 2017; Robberechts, van Haaren, & Davis, 2019; Yao, Wang, Zhu, Cao, & Zeng, 2022) or forecast whether a specific shot will result in a goal (Lucey, Bialkowski, Monfort, Carr, & Matthews, 2015). Other approaches utilize imitation learning techniques to predict player movement trajectories (Lindström, Jacobsson, Carlsson, & Lambrix, 2020), using sentiment analysis techniques for goal prediction (Wunderlich & Memmert, 2022) or estimate the likelihood of the next event being a goal (Decroos & Davis, 2019). However, to the best of our knowledge, there are no studies that predict in-play match events beyond the immediate next event.

This article investigates to what extent goals and other success- or scoring-related match events (SREs) in the near future, e.g., shots taken, corner kicks, or box entries, can be predicted using prior performance data from the near past. There is good reason to believe that these predictions can be made utilizing the concept of *match momentum* (Veo, 2021; Whitmore, 2021), also known as *game flow* (Ho, McKinley, & Moore, 2018) or *dominance* (Link, Lang, & Seidenschwarz, 2016), representing the actual performance difference between the two teams. Hence, match momentum can reveal a performance streak for one team if it is on its side, which might result in achieving more SREs and therefore a higher probability of this team scoring. Meanwhile, psychological momentum (Iso-Ahola & Dotson, 2016; Iso-Ahola & Mobily, 1980), or often just momentum, is difficult to measure in soccer (Redwood-Brown et al., 2018), although it is somewhat related to match momentum and vice versa. According to Higham et al. (2005, p. 3), momentum in football is described as a “hidden force [sic]” in a match, flowing between the two teams and representing the energy between the competitors. Momentum is also elaborated in other game sports such as American football (Roebber, Burlingame, & deWinter, 2022), basketball (Arkes & Martinez, 2011) and ice hockey (Leard & Doyle, 2011). We argue that match momentum represents the evolving performance difference between teams over a short period, which can be captured through in-game performance indicators (PIs). This concept aims to link past in-game performance trends to near-future outcomes.

In light of the above, the question arises of which in-play PIs or PI-combinations best predict SREs in general and thus represent the actual team performance. Answering this question also contributes to the discussion of the validity of PIs and to the theory of football (Carling, Wright, Nelson, & Bradley, 2014; Mackenzie & Cushion, 2013). A valid PI should be stable in terms of a small random fluctuation (Davis et al., 2024; Lames & McGarry, 2007) and be associated with a component of success (Hughes & Bartlett, 2002). When a PI demonstrates its validity for (*short-term*) success, then it is feasible for (*in-play*) analytical questions (Lames & McGarry, 2007), e.g., to measure the effectiveness of a tactical intervention such as a substitution, change of pressing strategy, or style of play. Beside other techniques (Davis et al., 2024), research suggests using the PIs' predictive power for success or its components to compare and evaluate them (Herold, Kempe, Bauer, & Meyer, 2021; Hughes & Bartlett, 2002). Recent years have seen some studies on this, but usually on seasonal or match-level (Russomanno, Linke, Geromiller, & Lames, 2020) or with a different methodology than ours (Davis et al., 2024; Decroos & Davis, 2019; Herold et al., 2021), but not for the short-term predictive power as we suggest. Regarding PIs and their value, spatiotemporal player tracking data-based PIs are relatively good predictors, e.g., for goal-scoring opportunities (Wagenaar, Okafor, Frencken, & Wiering, 2017), match outcome (Goes, Kempe, & Lemmink, 2019) and the outcome of the second half (Klemp, Wunderlich, & Memmert, 2021). Others used event-related PIs to predict match outcomes (Cintia, Giannotti, Pappalardo, Pedreschi, & Malvaldi, 2015), to identify PIs that discriminate between winning and losing teams (Harrop & Nevill, 2014), rate single actions performed by players (Liu, Luo, Schulte, & Kharrat, 2020) or predict the next event with information from the last actions (Simpson, Beal, Locke, & Norman, 2022; Wang et al., 2024). While some of these studies share similar goals to ours, we focus on short-term event prediction and exclusively use in-play PI data to predict the occurrence of a special event in the next few minutes.

Given this context, this study aims to explore the applicability of common PIs in soccer to predict SREs. To achieve this, we test to what extent our prediction goal (PG), i.e. the occurrence of an SRE, in a defined future time span (prediction window) can be predicted for one team by using the values of in-play PIs (e.g., passes, time of ball possession, opponents outplayed) within a defined time span from the past (input window). For prediction, we employ five different machine learning models (MLM) for multiple window sizes and combinations. The PIs are based on event and spatiotemporal tracking data from 102 German Bundesliga matches. Consequently, our study is organized in four parts: First, (Part I) we evaluate the inter-correlation between PIs and PGs in matches, quantifying their relationship. Second, (Part II) we explore the performance of different MLMs and rank the PIs based on their mean predictive power for our PGs to quantify their general usability. Subsequently, (Part III) we test the prediction performance of different PI-combinations for one selected scenario (predicting a goal in the next three minutes) in order to evaluate the interaction of the PIs. Finally, (Part IV) we propose a real-world application utilizing all findings for a match momentum metric.

Our research holds significance for both the betting industry and performance analysis. It addresses questions regarding the relationship between input window size and prediction performance, as well as identifying which PI or PI-combinations can effectively be utilized as in-game team performance representation. The results can be employed to enhance in-play betting odds and contribute valuable insights to practitioners regarding the validity and usability of PIs for match analysis and real-time decision-making.

Methods

Dataset

Data for our study comprises 102 matches of German professional football leagues, encompassing 82 Bundesliga and 20 Bundesliga 2 matches played across 31 rounds in the 2017/2018 season. Each match record contains three types of data: positional data of players (x/y coordinates), and ball (x/y/z) captured with 25 hertz; event data; and basic match information, e.g., player/team names, and playing positions. Positional data was collected semi-automatically using the TRACAB optical tracking system (ChyronHego, NY), while in-game events were manually recorded by human observers (DFL, 2020). The eight-centimeter accuracy of the tracking system was validated by Linke et al. (2020). All 36 teams are represented with a minimum and maximum of two to twenty matches. Each team had home and away matches, except one team which had only away matches available. For each match and each team, we calculated PIs (see *Parametrization*). We aggregated this data in intervals to use it as input data (see *Motivation and Design*). To ensure sufficient SREs in both the validation and test sets, we randomly split the dataset on match-level into training, validation, and test sets with a 60:20:20 ratio. This split also helps ensure robust evaluation across a diverse range of matches and scenarios. For reproducibility and testing purposes, 102 synthetic datasets are provided in the associated GitHub repository available at https://github.com/SteffenLa/Shortterm_Soccer_Event_Prediction.git.

Experimental Setup

Motivation and Design

Our study aims to predict SREs based on information from the running game (e.g., whether a goal will be scored by a team in the next few minutes using the information given from the previous minutes). Identifying PIs capable of predicting such events can help to validate their applicability as team performance representation in football; however, finding suitable PIs for this type of prediction is challenging. Some PIs have different timescales and characteristics, which makes it difficult for MLMs to utilize them effectively. For instance, the well-known PI *Shots Taken* can be useful to predict goals in a match, but occurs infrequently and is only a good predictor for the immediate future (Mead, O'Hare, & McMenemy, 2023). These characteristics seem to make this PI unsuitable for quantifying team performance for more than a specific moment. Similarly, numerous shots in a shorter timespan could be included in this metric. Therefore, we propose a windowing approach to address the issue of overvaluing a PI's immediate prediction quality by MLMs (see *In-Play Prediction Masking*). Our approach masks data immediately before the prediction window, giving less weight to PIs that have a direct short-term relation to sparse PGs, such as shots leading to goals. This approach reduces the effect of such a PI on the MLM but still enables its utilization on a longer time scale.

Another challenge is that some individual PIs may not have a good prediction quality or are designed for different prediction time spans (e.g., seconds vs. minutes). However, combining those PIs with others may result in a higher prediction accuracy than their individual contributions. Unfortunately, exploring all of them in-depth is computationally infeasible due to the number of PIs, PGs, MLMs and their specific hyperparameters. Even with a large dataset, models can be over-fitted and may not provide a general answer to which PIs are suitable for representing team performance. To tackle this problem, we structured our experiments to focus on gaining insights into the PIs. By evaluating the average performance of an individual PI on predicting various SREs (Part II) employing *default-configured* MLMs, similar to existing research (Decroos & Davis, 2019; Sipper, 2022), we can better understand the PIs utility and

which MLM can most effectively leverage the PI. We do not advocate for or against a particular MLM, as the performance of most MLMs can be improved through experimental tuning and investing time. Instead, we propose PIs that can be easily leveraged to increase performance to benefit a variety of PGs and MLMs.

After this in Part III, we perform an extensive experimental evaluation of PI-combinations for one selected PG, filter them to a sufficiently small set to obtain a complete combinatorial picture, and run cross-validation to verify their stability with respect to the specific dataset distribution. Note, PI-combinations refer to the use of two or more PIs as input features, without creating new PIs.

In-Play Prediction Masking

To address the confounding issues between PIs and PGs, we propose a simple way to mask the data, called In-Play Prediction Masking. Figure 1 illustrates the design of our approach, which defines a single sample based on its input window (IW), hidden window, and prediction window (PW). Given that a game interruption can last several minutes and most PIs are either unreliable or unavailable during this period, we only consider the effective playing time for the input window, usually about 55 minutes in a match (Siegle & Lames, 2012). The IW consists of the PI values sampled every 5 seconds, with varying window lengths, followed by the hidden window, neither belonging to the IW nor to the PW. The trained model predicts whether an SRE has occurred in the PW or not (*binary classification*). We use a rolling window approach, where all the windows move through the entire match with a defined step size. Applying In-Play Prediction Masking, we can assign less importance to very short-term predictors that are not useful in measuring a team's performance over a longer time span. Instead, we concentrate on PIs that represent this performance best and can potentially anticipate future events. Our experiments have shown that this approach is crucial, allowing us to focus on high-quality predictive PIs and to develop MLMs capable of anticipating such events effectively.

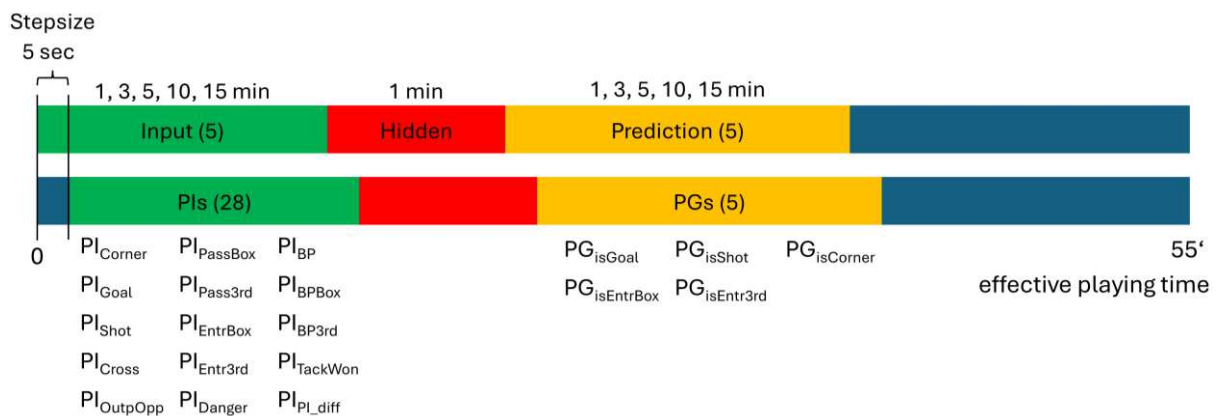


Figure 1. Visualization of the In-Play Prediction Masking approach, the rolling window approach, and the utilized PIs; PGs; IW and PW window sizes. The lower band visualizes the rolling window, whereas all windows moved one step further. In brackets the number of configuration feature levels.

Parametrization

This approach incorporates various parameters, including MLMs; IW, PW, and hidden window sizes; rolling window step sizes; input data resolutions; PIs; and PGs. We selected five commonly used MLMs: Logistic Regression (LR), K-Nearest-Neighbors (KNN), Gaussian Naive Bayes (NB), Support Vector Machine (SVM), and Neural Networks (NN). An overview of the default configurations of our utilized MLMs and minor individual changes to fit the dataset can be found in the appendix (Tab. 2). Further, to provide a solid baseline, we

implemented a random guessing algorithm based on average historical data. The probability used for the Random Guesser is derived from the ratio of all samples in the training set with the event versus those without the event. Statistically, we conducted a Bernoulli experiment for each prediction sample, establishing a minimum threshold that any sophisticated model should surpass to be deemed useful. The IW and PW sizes were each set to 1, 3, 5, 10, and 15 minutes of effective playing time, which covers a variety of possible time spans for measuring the actual team performance during a match (Fig. 1). We opted for a one-minute hidden window size to ensure that no confounding between the PIs and the PGs was present. The step size of the rolling window was set to the input data resolution of five seconds to create as many observations as possible. To measure the in-game performance of a team, various indicators are publicly available and discussed in research. We selected 14 feasible PIs, which can be calculated while the match is running, e.g., time of ball possession, zone entries, opponents outplayed, or Dangerousity (Link et al., 2016), with only Dangerousity being based on spatiotemporal tracking data. As a game of interactions, one team's performance is intricately linked to the opposing team's performance. Consequently, we expanded these 14 PIs for one team by calculating the difference to the opponent in the same interval, resulting in a total of 28 PIs (Tab. 1a). We selected five PGs which are all representative of team performance (Herold et al., 2021; Mackenzie & Cushion, 2013) (Tab. 1b). Features were normalized to a [0,1] range. The scaler was fitted to the training data and subsequently applied to the validation and test sets. To ensure full transparency and reproducibility, all scripts for data processing, model training, and evaluation used in this study are available in the associated GitHub repository.

Table 1: a) Performance indicators (PIs) and b) Prediction goals (PGs) of a team and their definitions utilized in our study. PIs 1-14 as individual PIs and 15-28 as the difference to the opponent team of the individual ones. The definition of a PI is always based on the performance of the respective team in an interval and is either an event performed, or a metric based on actions of the team. The definition of a PG is that the event happens at minimum once in the respective prediction window for the team.

No	Abbreviation	Definition
a) Performance Indicators (PI)		
1	PI _{Corner}	Number of corner kicks
2	PI _{EntrBox}	Number of entries of a player with ball possession into the opponent box
3	PI _{Entr3rd}	Number of entries of a player with ball possession into the attacking third
4	PI _{Goal}	Number of goals scored
5	PI _{Shot}	Number of shot attempts
6	PI _{Cross}	Number of crosses
7	PI _{TackWon}	Number of tacklings won
8	PI _{PassBox}	Number of successful passes in or into the opponent box
9	PI _{Pass3rd}	Number of successful passes in or into the attacking third
10	PI _{BP}	Time of ball possession
11	PI _{BPBox}	Time of ball possession in the box
12	PI _{BP3rd}	Time of ball possession in the attacking third
13	PI _{OutpOpp}	Number of outplayed opponent players by successful passes
14	PI _{Danger}	Goal scoring probability at each moment (Link et al., 2016)
15-28	PI _{PI_diff}	Difference of both PI values, (Team – Opponent)
b) Prediction Goals (PG)		
1	PG _{isGoal}	A goal event for the team occurs
2	PG _{isShot}	A shot event for the team occurs
3	PG _{isCorner}	A corner kick event for the team occurs
4	PG _{isEntrBox}	An entry into the opponent box performed by the team occurs
5	PG _{isEntr3rd}	An entry into the attacking third performed by the team occurs

Evaluation

For Part I, which focuses on gaining deeper insights into the PIs, we assess the inter-correlation between PIs and our PGs during matches. To compute these correlations, we establish pairs of observations as follows: we aggregate the values of one PI across all intervals in a defined window and correspondingly aggregate the PG values in the subsequent window, with each window of the same length. We then employ a rolling window approach across the entire dataset, utilizing a predefined step size of one minute. This process is repeated for each team across all matches. For correlation we use the Pearson correlation coefficient. The results reveal any correlations among all PIs and the PGs, highlighting which PIs may best represent a PG.

For Parts II and III, we employ the Matthews Correlation Coefficient (MCC) also known as Phi coefficient metric (Guilford, 1954) (Eq. 1), recommended for evaluating models with highly unbalanced datasets (Chicco & Jurman, 2020). Our selected PGs represent sparse events in football matches, e.g., 2.43-3.06 goals (Farias, Bergmann, Vaz, & Pinheiro, 2018) or 10 corner kicks per match (Siegle & Lames, 2012), in our dataset 2.99 and 9.54 per match, respectively. With 139,694 five-second intervals as observation units in our study, the dedicated PGs are present in only 0.22% (goals), 0.7% (corner kicks), 1.28% (shots taken), 2.54% (entries into the opponent box) and 6.97% (entries into the attacking third) intervals. The MCC ranges from -1 to 1, where 0 indicates random prediction. Consequently, while a random guessing model would generate MCC results around 0, indicating chance prediction, trained MLMs are expected to yield MCC values higher than 0, signifying learned and predictive capabilities.

In the final phase of our analysis, we conduct an in-depth exploration of one practical scenario to identify the most effective PI-combinations (Part III). Here, we carry out an extensive model training iteration, covering most of the conceivable configurations specifically for PG_{isGoal} within the defined PW_{3min} . Given the large number of possible combinations, we capped the maximum combined PIs at five. To do so, we combined, without repetition, all 28 PIs up to five PIs ($\sum_{k=1}^5 \binom{28}{k} = 122K$). Subsequently, we train the best MLM identified in Part II using all combinations across the five different IW sizes. These results provide insights into the most promising individual PIs and their influence. From these PIs, we selected the Top 10 PIs to conduct the training with all possible combinations ($\sum_{k=1}^{10} \binom{10}{k} = 1023$), to determine the best configuration of $IW \times PI$ -combination for the dedicated scenario. Furthermore, we assess this scenario under 10-fold cross-validation to verify the stability of our results. For evaluation, we aggregate the results of the iterations and compare them using the 95% confidence interval (CI_{95}) as it shows which configurations are, in most cases, superior to others. For the final analysis, we investigate the top ten percent results (Top 10%) to learn more about the PIs with the best predictive power.

$$\text{Equation 1: } MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Results

Inter-correlation between PIs and PGs in matches

Analyzing the relationship between the selected PIs and our PGs in Part I, Figure 2 shows the resulting Pearson correlation coefficient (r) calculated with a window size of 15 minutes for both, PIs and PGs. For PG_{isGoal} , results show that there is only a low correlation between all PIs and the PG ranging from $r = -0.22$ to $r = 0.13$, with the highest results achieved by $PI_{Pass3rd}$,

PI_{Danger} , and their team-differences. However, systematic trends emerge across all PGs. Figure 2 illustrates that higher correlations come in two cases. First, with an increasing frequency of an SRE, i.e. the sparse PG_{isGoal} achieves an $r_{\text{max}} = 0.13$, whereas the most often occurring PG_{Entr3rd} in our dataset has an $r_{\text{max}} = 0.57$. Second, when utilizing the respective PI differences most of the time (e.g., PI_{BP} and $PI_{\text{BP_diff}}$).

Furthermore, PI_{Goal} , PI_{TackWon} , and their differences produce only low correlations with all PGs. Conversely, full pitch time of ball possession (PI_{BP}), time or actions occurring in the final third (PI_{Pass3rd} , PI_{BP3rd} , PI_{Entr3rd}), and their team differences exhibit high correlations. Note as an exception, while PI_{BP} correlates low with PG_{isGoal} , time of ball possession in the final third (PI_{BP3rd}) demonstrates a strong correlation. Besides, there is a noticeable trend indicating that an increase in correlation comes with an increase in total time due to differing window lengths. This is presumably because some PIs and PGs are sparse, and by using a larger time window, they are included more often. The complete correlation results for all examined window sizes are in the supplementary material (Tab. S1).

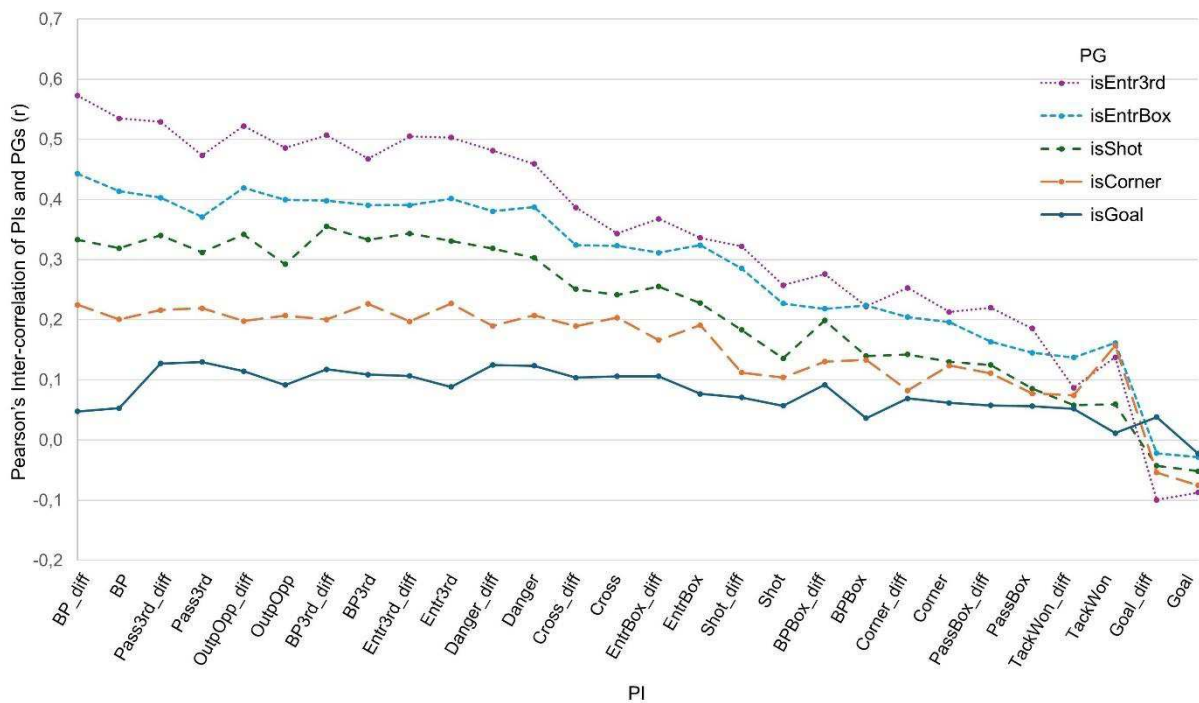


Figure 2. Pearson's inter-correlation of PIs and PGs with a window size of 15 minutes ordered for readability. Complete results are shown in Table 3 in the Appendix.

Performance of machine learning models and performance indicators for different prediction goals

Ranking the examined machine learning models

To classify suitable MLMs and PIs (Part II), we conducted the analysis regarding their predictive performance in all configurations. Given the diverse parametrization generated by each configuration ($MLM \times IW \times PW \times PI \times PG$), wherein certain MLMs or PIs perform differently, there is no singular approach to compare and rank the results of all prediction experiments uniformly. Due to our goal of finding suitable PIs on average, we rank our results by the MCC_{mean} for decision-making. Figure 3 shows the MCC results for each PG and MLM, each boxplot contains 700 results of all examined configurations ($IW(5) \times PW(5) \times PI(28)$), except for $PG_{\text{isEntr3rd}}$ where results of particular configurations ($IW_{10\text{min}}$, $IW_{15\text{min}}$, $PW_{10\text{min}}$,

PW_{15min}) were omitted due to the event occurring in every observational unit. Here, LR outperforms all other MLMs in all PGs by their mean results (31% better MCC_{mean} to 2nd rank). Interestingly, the MCC_{mean} ranking of the examined MLMs is the same for all chosen PGs (except for PG_{isGoal}, where NB and SVM switched places two and three). Our baseline model (random guesser) performs as expected with an MCC_{mean} close to 0 (MCC_{mean_RandomGuesser} = $-2e^{-4}$ to $-5e^{-5}$). The minimum and maximum MCC results fluctuate among each configuration, as it is highly dependent on the data split. Facing this, we perform 10-fold-cross validation in Part III to obtain more robust performance estimates. The complete experimental results are available in the supplementary material (Tab. S2).

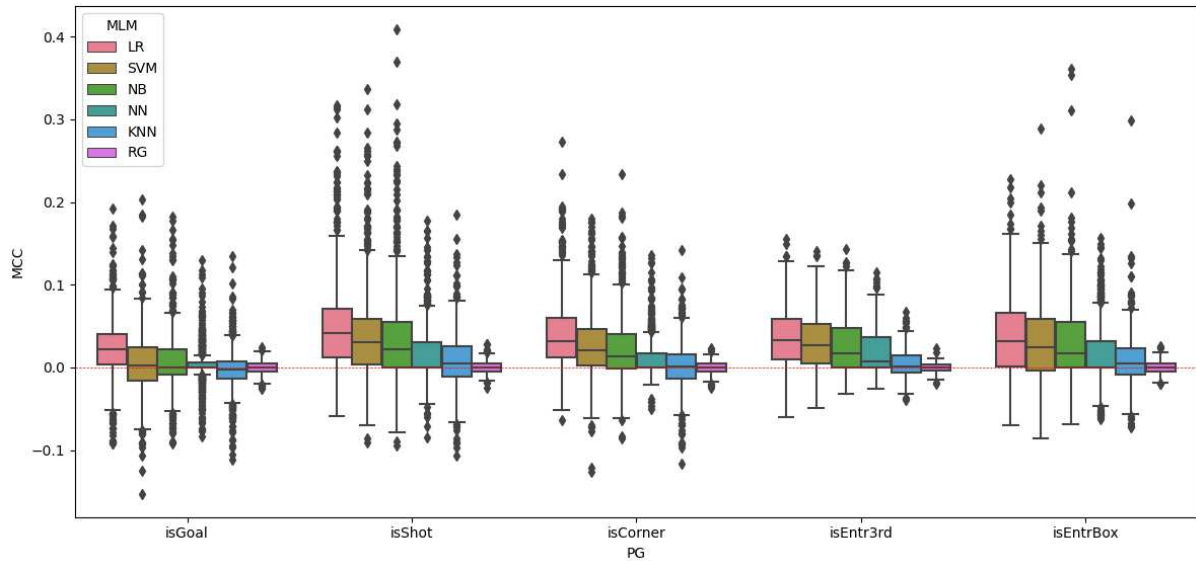


Figure 3. MCC results of all experiments split by MLMs and PGs. Each boxplot contains results of 700 experiments, except for the PG_{isEntr3rd} with 250 experiments. The red dotted line indicates MCC=0.

Ranking individual performance indicators for in-play event prediction

For our second task in Part II, Figure 4 shows the same experimental results from the task before but now split by the individual PIs. Each boxplot contains 545 experiments each (MLM (5) $\times$ IW (5) $\times$ PW (5) $\times$ PG (5)). The results show the overall Top 3 PIs are PI_{Danger} (MCC_{mean} = 0.051 ± 0.053), PI_{Entr3rd} (0.045 ± 0.055) and PI_{Danger_diff} (0.042 ± 0.049). Lowest MCC_{mean} were achieved by PI_{Goal} (-0.006 ± 0.033), PI_{PassBox} (-0.005 ± 0.033), and PI_{TackWon_diff} (-0.001 ± 0.024) on average, indicating random results with an MCC_{mean} around 0.

Going deeper into these results with an additional split by the PGs, Figure 5 shows the Top 3 and Bottom 3 PIs per PG in descending order by their MCC_{mean}. PI_{Danger} is ranked best for PG_{isGoal} (MCC_{mean} = 0.037 ± 0.036) and PG_{isShot} (0.073 ± 0.068), whereas PI_{Entr3rd} is ranked at the top for PG_{isEntr3rd} (0.058 ± 0.041), PG_{isCorner} (0.051 ± 0.047), and PG_{isEntrBox} (0.069 ± 0.056). Additionally, exceptional good results were achieved by PI_{Danger}, PI_{Danger_diff}, and PI_{Pass3rd} having all a Top 9 rank for each PG. Similar to our results in Part I, PI_{Goal}, PI_{TackWon}, and their team-differences seem not to be useful as individual PIs to predict our PGs and are ranked in the bottom third. Complete PI rankings for each PG are available in the appendix (Fig. 8-12).

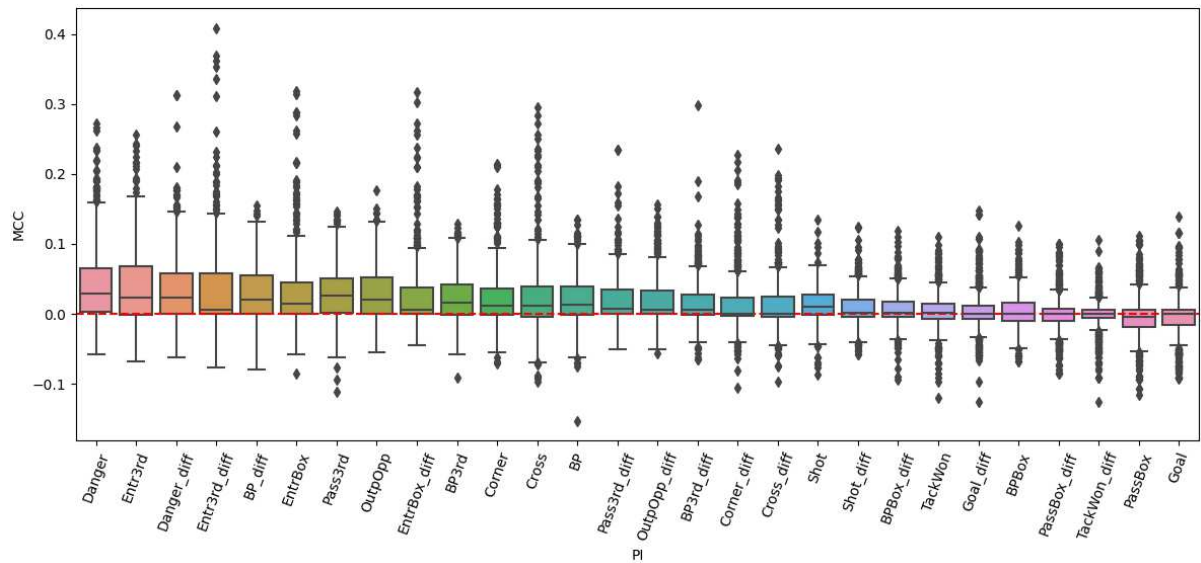


Figure 4. MCC results of all experiments split by PIs. Each boxplot contains the results of 545 experiments. Boxplots are sorted by their MCC_{mean} in descending order. The red dotted line indicates $MCC=0$.

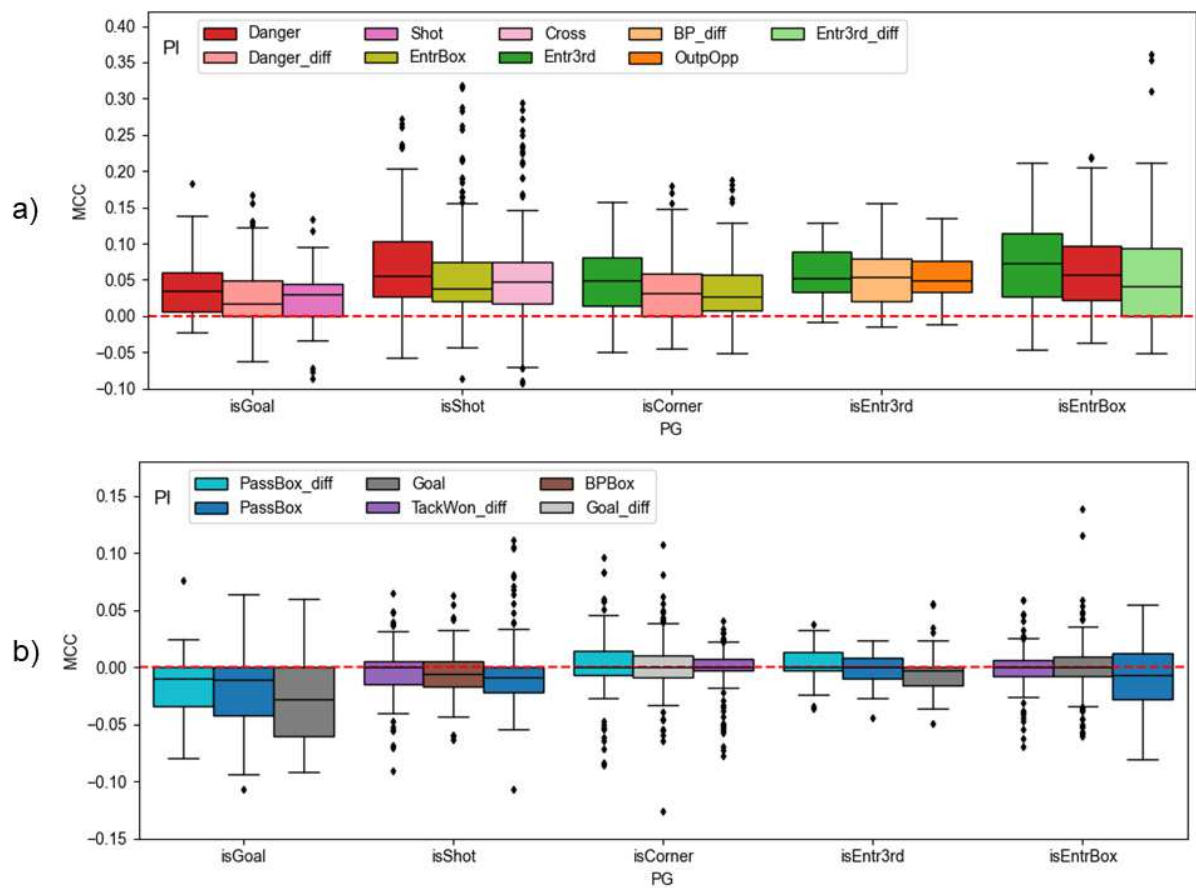


Figure 5. a) Top 3 and b) Bottom 3 PIs for each PG. PIs are sorted by their mean MCC results in descending order. Each boxplot contains the results of 125 experiments, except for the PG_{isEntr3rd} with 45 experiments. The red dotted line indicates $MCC=0$.

Prediction of a goal in the next three minutes

For our third part, we conducted extensive experiments where we first searched for the best ten individual PIs and later combined these in all possible combinations for the specific scenario of predicting a goal in the next three minutes (Part III). Figure 6a shows the Top 10 on average best performing PIs selected. The resulting PI-combinations were 10-fold cross-validated on all five IW sizes with LR, as it showed the best results in Part II. Regarding the overall CI_{95} -results, we can select three interesting insights from the Top 10% configurations. First, the resulting ranking of individual PIs shows that some have a significant effect over others. Here, in the Top 10% PI-combinations, $PI_{TackWon}$, PI_{Danger_diff} , and $PI_{Entr3rd_diff}$ occur the most often, 67%, 52%, and 49%, respectively (Fig. 6a). Second, regarding the number of individual PIs in a combination, four and five show the highest appearances with 25% and 23%, respectively (Fig. 6b). Third, the most occurring IW size is 5 minutes with 65% of all combinations, followed by 15 minutes with 32% (Fig. 6c). Interestingly, in 11 out of 512 cases (2%), only one PI was necessary to achieve a good prediction result. Moreover, $PI_{OutpOpp_diff}$ (rank 5, IW_{5min} , $CI_{95}=0.037$, $MCC_{mean}=0.055 \pm 0.025$), and PI_{Danger} (rank 8, IW_{15min} , $CI_{95}=0.035$, $MCC_{mean}=0.055 \pm 0.016$) yield an overall Top 10 rank, which shows the power of some individual PIs in their applicability for predicting an SRE. By combining $PI_{OutpOpp_diff}$ and $PI_{TackWon}$, the best result was achieved with $CI_{95}=0.047$ (rank 1, IW_{5min} , $MCC_{mean}=0.061 \pm 0.02$). Respective to the large fluctuation and dependencies on the data split, the best result with the lowest standard deviation was achieved by combining PI_{Danger} and $PI_{Entr3rd}$ (rank 3, IW_{15min} , $CI_{95}=0.039$, $MCC_{mean}=0.047 \pm 0.012$). An analysis of the two PIs that most often occur together revealed that $PI_{TackWon}$ and PI_{Danger_diff} are present in 37% of the cases where at minimum two PIs are combined. Indeed, this combination has also the second-best result overall (rank 2, IW_{5min} , $CI_{95}=0.04$, $MCC_{mean}=0.061 \pm 0.029$). The complete results of our selected scenario (predicting a goal in the next three minutes) are presented in the supplementary material (Tab. S3).

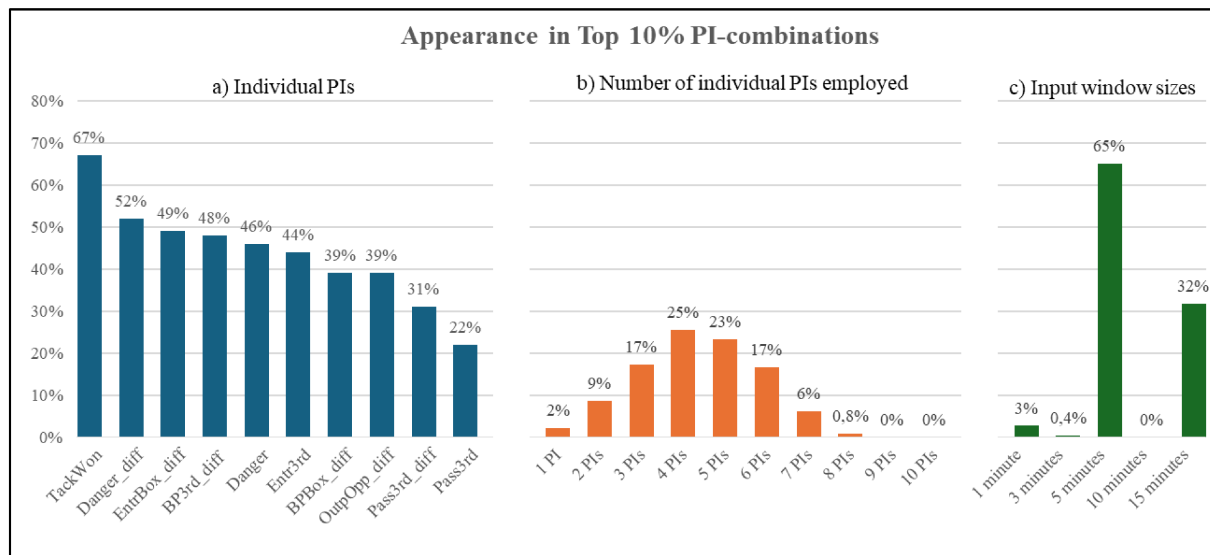


Figure 6. The percentual appearance in the Top 10% PI-combinations in the application scenario (Part III) of a) the selected Top 10 individual PIs, b) the number of individual PIs combined, and c) the input window length.

Application

In the last part of our study (Part IV), we show the applicability of our results, whereas we train a new model with our complete dataset and apply it to an unseen match. For this, we use the

configuration of the third-ranked model in Part III ($PI = PI_{\text{Danger}} \& PI_{\text{Entr3rd}}$, $IW_{15\text{min}}$, $PW_{3\text{min}}$, MLM_{LR} , PG_{isGoal}), as it showed only slightly fewer CI_{95} -result (rank 1: $CI_{95}=0.047$; rank 3: $CI_{95}=0.039$) but the lowest standard deviation overall (rank 1: $MCC_{\text{mean}}=0.061 \pm 0.02$ vs. rank 3: $MCC_{\text{mean}}=0.047 \pm 0.012$), indicating the most stable configuration. Then, we predict over the course of the match whether a goal will be scored in the next three minutes for both teams taking part in the match FC Bayern Munich (FCB) vs. SC Paderborn (SCP) of the season 19/20. This match was not part of our study so far and can be seen as an application test for our trained model. Subsequently, in Figure 7, we plot for this match the predicted probabilities, the relevant PIs (PI_{Danger} and PI_{Entr3rd}), and important match events. Additionally, eight important phases are highlighted (P1-8) for explanation (see below). In the appendix, we offer figures for the first and second ranked configurations as well (Fig. 13 and 14). The top three trained models from our analysis are publicly available in the accompanying GitHub repository, facilitating reproducibility and further research.

The psychological momentum, frequently discussed in sports, is difficult to measure directly in football due to the low number of scoring events (Redwood-Brown, Sunderland, Minniti, & O'Donoghue, 2018). As a result, match momentum was introduced to represent the actual performance difference between the competing teams and streaks of one team. If one team has match momentum on its side, it could lead to a psychological advantage and create a “momentum”. Currently, there is no gold standard described for measuring match momentum. However, higher team performance typically leads to more SREs, suggesting that team performance should, in some manner, predict the occurrence of these events.

Therefore, we propose a possible application by utilizing all findings of our study for defining the match momentum metric. We define match momentum as the difference between the predicted goal probabilities of the two teams (Fig. 7). This empirically grounded metric accounts for both teams' performances and is stable with a high update frequency. For our application scenario, we selected one of the best configurations with the lowest variation, while other configurations also showed strong results and could be used similarly. Our match momentum definition provides a valuable metric for performance analysis since it might help to identify optimal moments for tactical interventions, e.g., when the opponent's likelihood of scoring is increasing, and it can be applied in setting live betting odds as it enables real-time prediction.

P1 indicates that prediction for the initial minutes of the game is impossible due to insufficient information. To address this, we add zeros for each PI and as soon as information is available, we include it in the prediction. **P2** shows that no prediction is made during interruptions in the match, as predictions are only calculated when the match is running. **P3** marks a dominant phase with an increasing predicted probability for FCB, where a goal was scored, and an additional shot was taken. **P4** involves a goal scored by SCP, with media reports stating, “Out of nowhere, a blunder by [FCB goalkeeper] Neuer brings SCP back into the game!” (Sport.de, 2020). SCP performed poorly in the period leading up to the goal, as indicated by the low number of entries and the missing shots. Predictions based on these PIs cannot foresee such goals. **P5** and **P6** demonstrate that both teams can simultaneously increase the probability of scoring in the near future by performing well. The prediction for one team is independent of the other team's performance. At the end of P6, FCB scored their second goal. Similarly to P4, **P7** features another unforeseeable goal, with reports noting that FCB was “sniffing around for the decisive third goal” (Sport.de, 2020) just before SCP scored their second goal. In **P8**, FCB increased the pressure and their performance and so increased their probability of scoring. At the end of this phase, FCB scored the decisive goal and won the match.

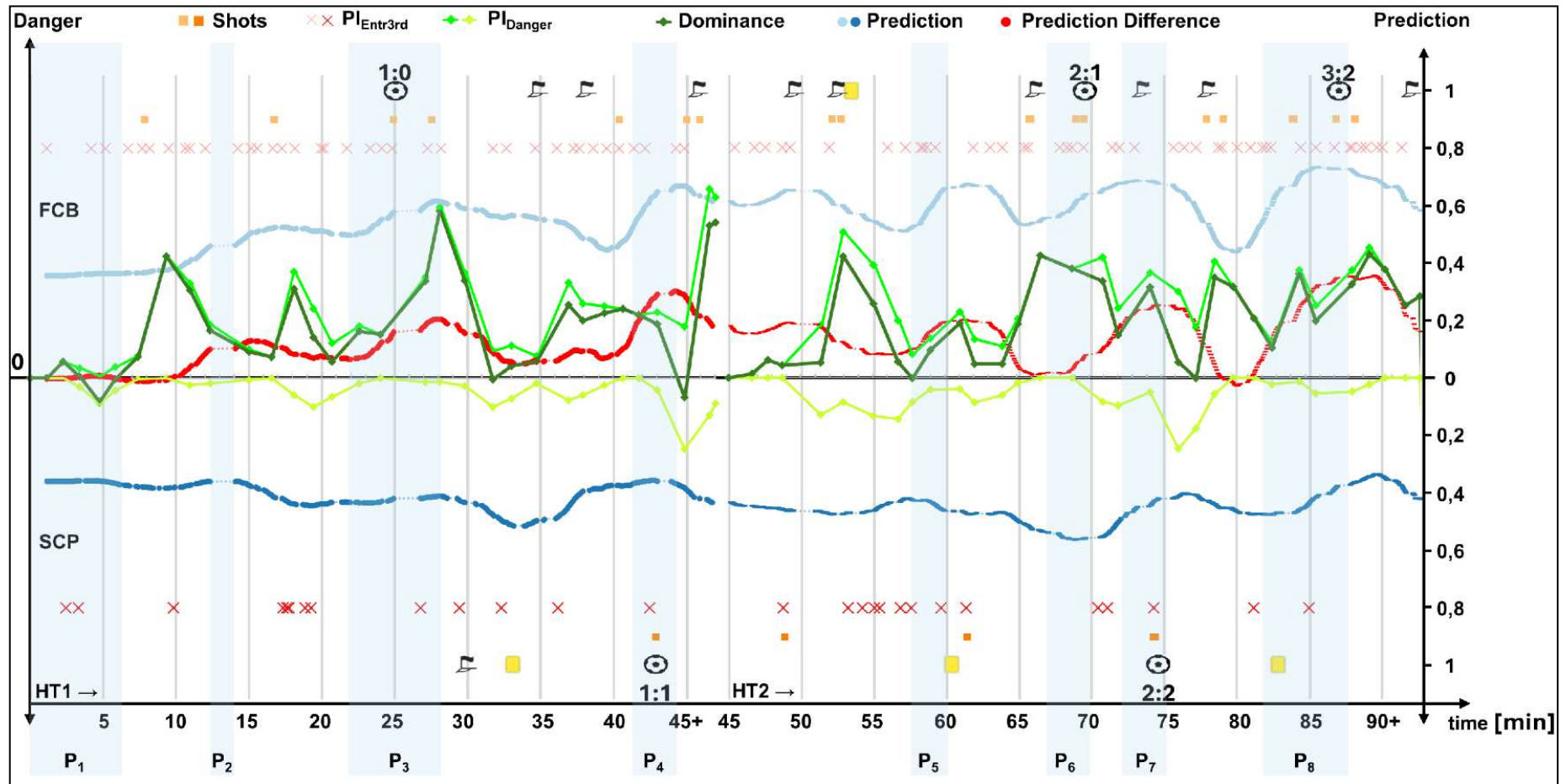


Figure 7. Application of the trained model (rank 3) for an unseen match between FC Bayern Munich (FCB) vs SC Paderborn (SCP) in Season 19/20 which resulted in a 3:2. In the upper half (FCB) and the lower half (SCP) for each team important events (corner kicks, given cards, goals scored, shots taken), PI_{Danger} and $PI_{Enter3rd}$, and prediction values are illustrated over the course of the match. Also, Dominance by Link et al. (2016) and the goal prediction difference between both teams, as our proposed match momentum metric, are shown. Additionally, eight important sequences (P1-8) are highlighted.

Discussion

Discussion of results

Our results reveal several interesting points. The correlation analysis indicates that in 11 out of 14 PIs, the PI difference between the two teams leads to significantly higher correlations than the individual PIs themselves. This is reasonable, since a greater PI difference should correspond to more SREs for the better-performing team. Similarly, increasing the correlation window length (from 1 minute up to 15 minutes in our experiment) increases the correlation. This is expected, as most of the examined PIs and PGs are sparse (i.e., goals, shots taken, corner kicks, entries). In summary, high correlations can be achieved by having more PI information available, the more frequently a PG occurs, and by utilizing the interaction with the opponent (PI differences).

Ranking the PIs by their mean prediction performance across all PGs provides valuable insights into their general usability. The PI ranking varies across the utilized PGs, indicating that there is no universally optimal PI. However, certain PIs consistently appear in the Top 10 results across all PGs, demonstrating their general usability. Conversely, some PIs consistently perform poorly (e.g., PI_{Goal} , $PI_{TackWon}$, $PI_{PassBox}$, PI_{Corner} , and their differences). Additionally, the analysis shows that some PIs ranked in the middle (e.g. PI_{Cross} or PI_{BP3rd_diff}) achieve high MCC results for particular configurations, which indicates either an overfitting for a specific data split or that in some cases the PI can be used to predict the dedicated PG quite well. As mentioned, our goal was not to find the best configuration for a particular PI, instead, we wanted to find those PIs overall performing on average good with default configurations across different PGs and MLMs. In the future, it is recommended to have a deeper analysis of these PIs, yet they did not perform best on average here. Summarizing, PIs based on spare events, e.g., goals, corner kicks and box actions, seem not to be useful for SRE prediction, whereas PIs based on a range of actions and with high frequency are beneficial. As far as the MCC results of all experiments are concerned, all configurations provide only low predictive performance, which was to be expected since soccer is strongly influenced by chance (Lames, 2018) and our PGs are very sparse. Nevertheless, MCC results indicate that prediction is better than random guessing and can be increased by further hyperparameter tuning. Specifically, for our application scenario, the best configuration achieved an MCC_{mean} approximately 6% better than random guessing. Considering that goals occur only about three times per match on average, this level of improvement represents a meaningful predictive gain in the context of such a sparse and chance influenced event.

Further insights are gained from our application scenario analysis. $PI_{TackWon}$, as a standalone indicator, has a low correlation with our selected PGs and demonstrates poor prediction quality. Conversely, combining $PI_{TackWon}$ with other PIs results in the highest performance and most frequent occurrence of a PI in our application scenario, underscoring the potential of combining PIs to enhance predictive power. Similarly, other PIs, such as PI_{Shot} and PI_{Cross} , could be valuable in combination. However, they were not included in our application analysis as they did not rank in the Top 10 during pre-selection, despite achieving Top 3 results in specific PGs. Surprisingly, results in Figure 6b indicate that combining more PIs does not necessarily lead to better prediction results, contrary to the expectation that more information would enhance prediction performance. This discrepancy might be due to the limitations of LR utilizing the full scope of available information, whereas a more sophisticated MLM could potentially make better use of it.

Additionally, PI_{Danger} , PI_{Danger_diff} , and $PI_{Entr3rd}$ consistently show strong results across all selected PGs. This consistency highlights their great value for predicting SREs and representing team performance. They demonstrate higher correlation and prediction quality

than others and are more frequently selected among the Top 10% configurations in our application scenario. Moreover, including $PI_{\text{Danger_diff}}$ in a PI-combination yields more stable results in the 10-fold cross-validation, with significantly lower standard deviations than without ($SD_{\text{Group_Danger_diff}} = 0.0216$ and $SD_{\text{Group_Not_Danger_diff}} = 0.0239$; U-statistic = 42766.5, $p < 0.001$). For future work, it is recommended to use these well-performing PIs and additionally utilize hyperparameter optimization to achieve even better prediction performance.

Discussion of methods and limitations

Regarding our employed methods, we encountered several issues. Utilizing MLMs with default configurations, without hyperparameter engineering, tuning, and searching (Decroos & Davis, 2019), penalizes deep learning models and those with numerous hyperparameters. However, we propose a method that enables a fair comparison across thousands of model configurations simultaneously. As a result, it highlights which PIs perform reliably across different model architectures. This is particularly useful for researchers and practitioners seeking initial insights without significant resource investment in hyperparameter tuning. Hence, LR was selected as the best-performing MLM presumably due to its fewer hyperparameters, whereas NN, for instance, were not chosen. Nevertheless, all utilized MLMs exhibited high MCC values in particular configurations, indicating their potential usability with hyperparameter optimization in the future. Using the MCC_{mean} for ranking the PIs can lead to a misunderstanding of their predictive power. This is because the mean includes both the best-performing MLMs (e.g., LR) and the worst-performing ones (e.g., KNN). Consequently, the maximum achieved MCC_{mean} for a PI per PG is quite low, e.g., between 0.037 to 0.073 meaning around 4% to 7% better than guessing, which does not accurately reflect the possibly predictive power of each PI. Instead, it highlights their general usability, which was our primary goal.

Furthermore, our examined PIs are restricted to those publicly available and can be calculated in-game. It would be from large interest and value to include and test other metrics, i.e., expected goals (Rudd, 2011), expected thread (StatsBomb, 2024), or VAEP (Decroos, Bransen, van Haaren, & Davis, 2020). These PIs are generally more sophisticated as they integrate multiple dimensions of play and therefore represent a richer description of the current game state, similar to Dangerousity. Although integrating match context variables, such as current league table position, results of recent matches, market value of the teams, or other inter-match metrics, would increase complexity, it could further enhance the accuracy of the predictions, as it has been shown valuable in other studies (Horvat & Job, 2020). While PI_{Danger} requires fine-grained positional tracking data, the majority of performance indicators analyzed in our study are based on event data alone. This ensures that our framework can be applied broadly, including in leagues and competitions where only event-level data is available. Notably, open-source event data of professional matches is accessible to researchers through resources such as StatsBomb's Open Data (StatsBomb, 2025), offering a valuable foundation for further replication and extension of our work. While the SREs investigated in this study are inherently sparse, this sparsity reflects their relevance in the context of soccer performance, as such events typically occur infrequently within or across possessions. Importantly, the proposed prediction framework is adaptable and can be applied to alternative event types (e.g., turnovers, freekicks) depending on the analytical focus and use case.

Our proposed In-Play Prediction Masking method addresses a known issue where specific PIs predict specific events very accurately within a short time window. For example, entries into the box can predict events, such as shots taken or goals scored, within a short time window of around one minute very good, as most shots and goals occur from inside the box. This confounding effect is problematic for MLMs and their analysis, as our models are intended to predict SREs over a longer time window in the future. To tackle this issue and examine the predictive ability of all PIs equally, we implemented a one-minute hidden window before the

PW. This method is not yet described in the literature and could penalize PIs that are better suited for very short-term prediction tasks.

Discussion of application

Our study examined different PIs by their predictive ability for SREs. The achieved MCC results are low for events like shots taken and goals due to their sparsity and the difficulty of prediction. Nonetheless, the results indicate that some PIs are in general more useful than others and warrant more intensive examination in the future. Furthermore, the combination of PIs appears to achieve even better results, particularly for PIs that have not been emphasized as important in literature anymore (e.g., Tacklings Won). Additionally, our analyses demonstrated that a 5-minute input window length achieves better results than a 15-minute window. This could be attributed to a psychological advantage based on a performance streak that may not last longer than 5 minutes.

As a possible application we use the difference in both teams' in-game goal predictions as an empirical based and stable representation of match momentum. This proposed match momentum metric updates more frequently and smoothly than other sparse PIs, and without the high fluctuation seen, for instance, in the Dangerousity metric (Link et al., 2016), making it a valuable tool for performance analysis and live betting markets. However, our proposed methodology for measuring match momentum has limitations, such as imprecise predictions during the initial minutes when the input data is mostly zero-filled, and the absence of predictions during match interruptions. To address these issues, we suggest adapting our methodology by including pre-game information as mentioned and by examining the interruption length influence on different PIs, as a performance advantage can be stopped by a longer interruption. For future research, we suggest utilizing different SRE predictions in combination, which would even better reflect both teams' actual performance. Overall, our results enhance the understanding of valid PIs by identifying those strongly associated with SREs and demonstrating their predictive stability.

Conclusion

In this paper, we examined the applicability of common in-game PIs for predicting success- or scoring-related match events (SREs) in soccer. To achieve this, we investigated which PIs or PI-combinations (e.g., time of ball possession, opponents outplayed, Dangerousity, shots taken) offer the highest predictive performance, as these should best reflect a team's current performance. For prediction, we employed data from 102 matches in German professional football leagues, we generated 28 well-known in-game PIs at every moment of effective playing time and tested thousands of machine learning model configurations. Our results demonstrate that while predicting in-game SREs is inherently challenging due to soccer's chaotic nature and the high influence of chance, in-game performance data contains valuable information that can be leveraged for short-term event prediction. However, PIs vary in their predictive power and general usability across different target events, indicating that some PIs are more applicable than others. Furthermore, combining PIs or using the difference between the competing teams' PI values generally leads to better predictive performance. In summary, PIs based on rare events, such as goals, corner kicks, and box actions, are less effective for SRE prediction. In contrast, PIs derived from a wider range of actions and collected at high frequency (e.g., ball possession in the last third, Dangerousity, outplayed opponents) are more beneficial. To the best of our knowledge, we are the first to predict in-play events beyond the immediate next event. Building on these findings, we propose a match momentum metric based on the difference in both teams' goal predictions. This empirically grounded tool can help identify optimal moments for tactical decisions, such as reacting to an opponent's increasing

likelihood of scoring. These insights lay the foundation for developing improved PIs and enhancing performance analysis in both coaching and live betting.

Additional Information

Author contribution statement

Conceptualization: SL, AE.
Data curation: SL, TW.
Formal analysis: SL, TW.
Funding acquisition: DL.
Investigation: SL.
Methodology: SL, TW, AE.
Project administration: SL.
Resources: AE.
Implementation: SL, TW.
Supervision: AE, DL.
Validation: SL.
Visualization: SL.
Writing ± original draft: SL.
Writing ± review & editing: SL, DL, AE.

Declaration of interest statement

The author(s) declare no competing interests.

Funding information

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

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to proprietary restrictions. Specifically, the dataset comprising 102 professional soccer matches was provided by the Deutsche Fußball Liga (DFL) under a non-disclosure agreement, which prohibits public sharing or redistribution of the data. Researchers interested in accessing similar data may contact the DFL directly. The original data may be made available from the corresponding author upon reasonable request, subject to DFL approval.

Supplementary Material

Supplementary material for this article is available online under https://figshare.com/projects/Lang_et_al_2025_Which_indicators_matter_Using_performance_indicators_to_predict_in-game_success-related_events_in_association_football/223491

This includes:

Supplementary Table S1: Complete experimental results of Part I in the study.
Supplementary Table S2: Complete experimental results of Part II in the study.
Supplementary Table S3: Complete experimental results of Part III in the study.

To ensure full transparency and reproducibility, all scripts for data processing, model training, and evaluation used in this study are available at https://github.com/SteffenLa/Shortterm_Soccer_Event_Prediction.git.

References

- Angelini, G., & Angelis, L. de (2017). Parx model for football match predictions. *Journal of Forecasting*, 36(7), 795–807. <https://doi.org/10.1002/for.2471>
- Arkes, J., & Martinez, J. (2011). Finally, Evidence for a Momentum Effect in the NBA. *Journal of Quantitative Analysis in Sports*, 7(3). <https://doi.org/10.2202/1559-0410.1304>
- Arntzen, H., & Hvattum, L. M. (2021). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, 21(5), 449–470. <https://doi.org/10.1177/1471082X20929881>
- Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, 35(2), 741–755. <https://doi.org/10.1016/j.ijforecast.2018.01.003>
- Beal, R., Middleton, S. E., Norman, T. J., & Ramchurn, S. D. (2021). Combining Machine Learning and Human Experts to Predict Match Outcomes in Football: A Baseline Model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17), 15447–15451. <https://doi.org/10.1609/aaai.v35i17.17815>
- Cao, C. (2012). *Sports data mining technology used in basketball outcome prediction*. Dissertation. Retrieved from <https://arrow.tudublin.ie/scschcomdis/39/>
- Carling, C., Wright, C., Nelson, L. J., & Bradley, P. S. (2014). Comment on 'Performance analysis in football: A critical review and implications for future research'. *Journal of Sports Sciences*, 32(1), 2–7. <https://doi.org/10.1080/02640414.2013.807352>
- Chae, J., Park, J., & So, W.-Y. (2021). Victory Prediction of Ladies Professional Golf Association Players: Influential Factors and Comparison of Prediction Models. *Journal of Human Kinetics*, 77, 245–259. <https://doi.org/10.2478/hukin-2021-0023>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 1–13. <https://doi.org/10.1186/s12864-019-6413-7>
- Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D., & Malvaldi, M. (2015). The harsh rule of the goals: Data-driven performance indicators for football teams. In E. Gaussier (Ed.), *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics: Ieee DSAA'2015 : 2015 IEEE International Conference on Data Science and Advanced Analytics, 19-21 October 2015, Paris* (pp. 1–10). Piscataway, NJ, USA: IEEE. <https://doi.org/10.1109/DSAA.2015.7344823>
- Davis, J., Bransen, L., Devos, L., Jaspers, A., Meert, W., Robberechts, P., . . . van Roy, M. (2024). Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned. *Machine Learning*, 113(9), 6977–7010. <https://doi.org/10.1007/s10994-024-06585-0>
- Davoodi, E., & Khanteymoori, A. (2010). Horse racing prediction using artificial neural networks. In V. Munteanu, R. Raducanu, G. Dutica, A. Croitoru, V. E. Balas, & A. Gavriliuț (Eds.), *Recent advances in neural networks, fuzzy systems & evolutionary computing: Proceedings of the 11th WSEAS International Conference on Neural Networks (NN'10) : proceedings of the 11th WSEAS International Conference on Evolutionary Computing (EC'10) : proceedings of the 11th WSEAS International Conference on Fuzzy Systems (FS'10) : " G. Enescu " University, Iasi, Romania, June 13-15, 2010* (pp. 155–160). [Wisconsin, USA]: WSEAS Press.
- Decroos, T., Bransen, L., van Haaren, J. , & Davis, J. (2020). VAEP: An Objective Approach to Valuing On-the-Ball Actions in Soccer (Extended Abstract). In C. Bessiere (Ed.),

- Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)* (pp. 4696–4700). International Joint Conferences on Artificial Intelligence. <https://doi.org/10.24963/ijcai.2020/648>
- Decroos, T., & Davis, J. (2019). Interpretable prediction of goals in soccer. Retrieved from https://tomdecroos.github.io/reports/interpret_vaep.pdf
- DFL (2020). How is the official match data collected? Retrieved from <https://www.dfl.de/en/innovation/how-is-the-official-match-data-collected/>
- Dobson, S., & Goddard, J. (2017). Evaluating probabilities for a football in-play betting market. In P. Rodríguez, B. R. Humphreys, & R. Simmons (Eds.), *The Economics of Sports Betting* (pp. 52–70). Edward Elgar Publishing.
- Farias, V. M., Bergmann, G. G., Vaz, L. M. T., & Pinheiro, E. d. S. (2018). Goal scoring frequency in soccer in different age groups. *Rev. Bras. Ciênc. Mov*, 47–52. Retrieved from <https://pesquisa.bvsalud.org/portal/resource/pt/biblio-965586>
- 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>
- Goddard, J., & Asimakopulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1), 51–66. <https://doi.org/10.1002/for.877>
- Goes, F., Kempe, M., & Lemmink, K. (2019). *Predicting match outcome in professional Dutch football using tactical performance metrics computed from position tracking data*. EasyChair. <https://doi.org/10.29007/4jjb>
- Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). McGraw-Hill. Retrieved from <https://psycnet.apa.org/record/1955-00102-000>
- Harrop, K., & Nevill, A. (2014). Performance indicators that predict success in an English professional League One soccer team. *International Journal of Performance Analysis in Sport*, 14(3), 907–920. <https://doi.org/10.1080/24748668.2014.11868767>
- Herold, M., Kempe, M., Bauer, P., & Meyer, T. (2021). Attacking Key Performance Indicators in Soccer: Current Practice and Perceptions from the Elite to Youth Academy Level. *Journal of Sports Science & Medicine*, 20(1), 158–169. <https://doi.org/10.52082/jssm.2021.158>
- Higham, A., Harwood, C., & Cale, A. (2005). *Momentum in soccer: Controlling the game*. Armley, Leeds: Coachwise.
- Ho, C. H., McKinley, E., & Moore, J. (2018, September 18). Ebb-and-flow: Using xPG to capture soccer's momentum swings. *American Soccer Analysis*. Retrieved from <https://www.americansocceranalysis.com/home/2018/9/10/xpg-pt-2>
- Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5), e1380. <https://doi.org/10.1002/widm.1380>
- Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of Sports Sciences*, 20(10), 739–754. <https://doi.org/10.1080/026404102320675602>
- Igiri, C. P., & Nwachukwu, E. O. (2014). An Improved Prediction System for Football a Match Result. 2278-8719. Retrieved from <http://ir.mtu.edu.ng/xmlui/handle/123456789/113>
- Jackson, L. A. (2015). Bet365 Reveal 80% Of Sports Betting Revenue Comes From Live In-play Betting. Retrieved from <https://www.online-betting.me.uk/news/bet365-reveal-80-of-sports-betting-revenue-comes-from-live-in-play-betting>

- Karlis, D., & Ntzoufras, I. (2009). Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, 20(2), 133–145. <https://doi.org/10.1093/imaman/dpn026>
- Killick, E. A., & Griffiths, M. D. (2019). In-Play Sports Betting: A Scoping Study. *International Journal of Mental Health and Addiction*, 17(6), 1456–1495. <https://doi.org/10.1007/s11469-018-9896-6>
- Klemp, M., Wunderlich, F., & Memmert, D. (2021). In-play forecasting in football using event and positional data. *Scientific Reports*, 11(1), 24139. <https://doi.org/10.1038/s41598-021-03157-3>
- Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, 12(3). <https://doi.org/10.1515/jqas-2015-0059>
- Lames, M. (2018). Chance involvement in goal scoring in football – an empirical approach. *German Journal of Exercise and Sport Research*, 48(2), 278–286. <https://doi.org/10.1007/s12662-018-0518-z>
- Lames, M., & McGarry, T. (2007). On the search for reliable performance indicators in game sports. *International Journal of Performance Analysis in Sport*, 7(1), 62–79. <https://doi.org/10.1080/24748668.2007.11868388>
- Leard, B., & Doyle, J. M. (2011). The Effect of Home Advantage, Momentum, and Fighting on Winning in the National Hockey League. *Journal of Sports Economics*, 12(5), 538–560. <https://doi.org/10.1177/1527002510389869>
- Lindström, P., Jacobsson, L., Carlsson, N., & Lambrix, P. (2020). Predicting Player Trajectories in Shot Situations in Soccer. In U. Brefeld (Ed.), *Communications in Computer and Information Science Ser: v.1324. Machine Learning and Data Mining for Sports Analytics: 7th International Workshop, MLSA 2020, Co-Located with ECML/PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings* (Vol. 1324, pp. 62–75). Cham: Springer International Publishing AG. https://doi.org/10.1007/978-3-030-64912-8_6
- Link, D., Lang, S., & Seidenschwarz, P. (2016). Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data. *PloS One*, 11(12), e0168768. <https://doi.org/10.1371/journal.pone.0168768>
- Linke, D., Link, D., & Lames, M. (2020). Football-specific validity of TRACAB's optical video tracking systems. *PloS One*, 15(3), e0230179. <https://doi.org/10.1371/journal.pone.0230179>
- Liu, G., Luo, Y., Schulte, O., & Kharrat, T. (2020). Deep soccer analytics: learning an action-value function for evaluating soccer players. *Data Mining and Knowledge Discovery*, 34(5), 1531–1559. <https://doi.org/10.1007/s10618-020-00705-9>
- 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*. Retrieved from <http://disneyresearch.s3.amazonaws.com/wp-content/uploads/20150308192147/quality-vs-quantity%e2%80%9d-improved-shot-prediction-in-soccer-using-strategic-features-from-spatiotemporal-data-paper.pdf>
- Mackenzie, R., & Cushion, C. (2013). Performance analysis in football: A critical review and implications for future research. *Journal of Sports Sciences*, 31(6), 639–676. <https://doi.org/10.1080/02640414.2012.746720>
- Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE*, 18(4), e0282295. <https://doi.org/10.1371/journal.pone.0282295>

- Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team. *International Journal of Computer Theory and Engineering*, 812–815. <https://doi.org/10.7763/IJCTE.2013.V5.802>
- Parke, A., & Parke, J. (2019). Transformation of Sports Betting into a Rapid and Continuous Gambling Activity: A Grounded Theoretical Investigation of Problem Sports Betting in Online Settings. *International Journal of Mental Health and Addiction*, 17(6), 1340–1359. <https://doi.org/10.1007/s11469-018-0049-8>
- Purucker, M. C. (1996). Neural network quarterbacking. *IEEE Potentials*, 15(3), 9–15. <https://doi.org/10.1109/45.535226>
- Redwood-Brown, A. J., Sunderland, C. A., Minniti, A. M., & O'Donoghue, P. G. (2018). Perceptions of psychological momentum of elite soccer players. *International Journal of Sport and Exercise Psychology*, 16(6), 590–606. <https://doi.org/10.1080/1612197X.2017.1313295>
- Robberechts, P., van Haaren, J., & Davis, J. (2019). *Who will win it? An in-game win probability model for football*. Retrieved from https://dtai.cs.kuleuven.be/events/mlsa19/papers/robberechts_mlsa19.pdf
- Roebber, P. J., Burlingame, B. M., & deWinter, A. (2022). On the existence of momentum in professional football. *PloS One*, 17(6), e0269604. <https://doi.org/10.1371/journal.pone.0269604>
- 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 2011*. Cambridge.
- Russomanno, T., Linke, D., Geromiller, M., & Lames, M. (2020). Performance of Performance Indicators in Football. In M. Lames, A. Danilov, E. Timme, & Y. Vassilevski (Eds.), *Advances in Intelligent Systems and Computing: Vol. 1028. Proceedings of the 12th International Symposium on Computer Science in Sport (IACSS 2019)* (1st ed., Vol. 1028, pp. 36–44). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-35048-2_5
- Schauberger, G., & Groll, A. (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling*, 18(5-6), 460–482. <https://doi.org/10.1177/1471082X18799934>
- Siegle, M., & Lames, M. (2012). Game interruptions in elite soccer. *Journal of Sports Sciences*, 30(7), 619–624. <https://doi.org/10.1080/02640414.2012.667877>
- Simpson, I., Beal, R. J., Locke, D., & Norman, T. J. (2022). Seq2event: Learning the Language of Soccer Using Transformer-based Match Event Prediction. In A. Zhang (Ed.), *ACM Digital Library, Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 3898–3908). New York, NY, United States: Association for Computing Machinery. <https://doi.org/10.1145/3534678.3539138>
- Sipper, M. (2022). High Per Parameter: A Large-Scale Study of Hyperparameter Tuning for Machine Learning Algorithms. *Algorithms*, 15(9), 315. <https://doi.org/10.3390/a15090315>
- Sport.de (2020). Liveticker FC Bayern München - SC Paderborn 07 3:2 (Bundesliga 2019/2020, 23. Spieltag). Retrieved from https://www.sport.de/fussball/deutschland-bundesliga/ma8936099/fc-bayern-muenchen_sc-paderborn-07/liveticker/
- StatsBomb (2024). What is Expected Threat (xT)? Possession Value Models Explained. Retrieved from <https://statsbomb.com/soccer-metrics/possession-value-models-explained/>
- StatsBomb (2025). StatsBomb Open Data. Retrieved from <https://github.com/statsbomb/open-data>

- Veo (2021). Momentum Graph. Retrieved from <https://www.youtube.com/watch?v=0SRF5OeX8Bc>
- Wagenaar, M., Okafor, E., Frencken, W., & Wiering, M. (2017). Using Deep Convolutional Neural Networks to Predict Goal-Scoring Opportunities in Soccer. *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. Retrieved from <https://research.rug.nl/en/publications/using-deep-convolutional-neural-networks-to-predict-goal-scoring->
- Wang, Z., Veličković, P., Hennes, D., Tomašev, N., Prince, L., Kaisers, M., . . . Tuyls, K. (2024). Tacticalai: An AI assistant for football tactics. *Nature Communications*, 15(1), 1906. <https://doi.org/10.1038/s41467-024-45965-x>
- Whitmore, J. (2021, November 18). What is Match Momentum? *The Analyst*. Retrieved from <https://theanalyst.com/eu/2021/11/what-is-match-momentum/>
- Wunderlich, F., & Memmert, D. (2021). Forecasting the outcomes of sports events: A review. *European Journal of Sport Science*, 21(7), 944–957. <https://doi.org/10.1080/17461391.2020.1793002>
- Wunderlich, F., & Memmert, D. (2022). A big data analysis of Twitter data during premier league matches: Do tweets contain information valuable for in-play forecasting of goals in football? *Social Network Analysis and Mining*, 12(1), 1–15. <https://doi.org/10.1007/s13278-021-00842-z>
- Yao, W., Wang, Y., Zhu, M., Cao, Y., & Zeng, D. (2022). Goal or Miss? A Bernoulli Distribution for In-Game Outcome Prediction in Soccer. *Entropy*, 24(7), 971. <https://doi.org/10.3390/e24070971>

Appendix

Table 2: An overview of the five utilized machine learning models applied with default configurations in the study and the minor individual changes in configuration that were determined as optimal in hyperparameter tuning experiments on subsets of the data.

Machine Learning Model	Python library	Configuration	URL to documentation
Logistic Regression (LR)	Scikit-learn 1.4.0	class weights in loss: balanced (inversely proportional weighting according to class frequencies in train set) training metric: binary cross entropy	https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression
Gaussian Naive Bayes (NB)	Scikit-learn 1.4.0		https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
Support Vector Machine (SVM)	Scikit-learn 1.4.0	class weights in loss: balanced training metric: hinge loss	https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC
K-Nearest-Neighbors (KNN)	Scikit-learn 1.4.0	n_neighbors: 2 weights: Euclidean distance	https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier
Neural Network (NN)	PyTorch 2.3.0	class weights in loss: balanced 4 hidden layers with 256, 512, 128, 16 channels, respectively learning rate: 0.001 with exponential decay optimizer: Adam training until convergence of validation loss (early stopping) dropout probability: 0.4 activation function: LeakyReLU (Sigmoid in the last layer) training metric: binary cross entropy	https://pytorch.org/docs/stable/nn.html

Table 3: Pearson's inter-correlation results of PIs and PGs for the window length of 15 minutes. Top 3 (green) and Bottom 3 (red) PIs per PG are highlighted. The best and worst results per PG are bold.

	PG _{isGoal}	PG _{isCorner}	PG _{isShot}	PG _{isEntrBox}	PG _{isEntr3rd}
PI _{BP_diff}	.048	.225	.333	.443	.573
PI _{BP}	.053	.201	.319	.414	.535
PI _{Pass3rd_diff}	.127	.216	.340	.403	.529
PI _{Pass3rd}	.130	.219	.312	.371	.473
PI _{OutpOpp_diff}	.114	.198	.342	.419	.522
PI _{OutpOpp}	.092	.207	.292	.400	.486
PI _{BP3rd_diff}	.118	.201	.355	.398	.507
PI _{BP3rd}	.109	.227	.333	.391	.468
PI _{Entr3rd_diff}	.107	.197	.343	.391	.505
PI _{Entr3rd}	.089	.227	.331	.401	.503
PI _{Danger_diff}	.125	.190	.319	.380	.481
PI _{Danger}	.124	.207	.303	.387	.459
PI _{Cross_diff}	.104	.190	.251	.324	.387
PI _{Cross}	.106	.204	.242	.323	.343
PI _{EntrBox_diff}	.106	.166	.255	.312	.368
PI _{EntrBox}	.077	.191	.228	.324	.336
PI _{Shot_diff}	.071	.112	.183	.286	.322
PI _{Shot}	.057	.104	.136	.227	.258
PI _{BPBox_diff}	.092	.130	.199	.218	.276
PI _{BPBox}	.036	.133	.140	.224	.222
PI _{Corner_diff}	.069	.083	.142	.205	.253
PI _{Corner}	.062	.124	.130	.196	.213
PI _{PassBox_diff}	.058	.111	.125	.164	.220
PI _{PassBox}	.057	.078	.086	.145	.186
PI _{TackWon_diff}	.052	.074	.058	.138	.087
PI _{TackWon}	.012	.157	.060	.162	.138
PI _{Goal_diff}	.038	-.054	-.043	-.022	-.099
PI _{Goal}	-.022	-.075	-.052	-.028	-.087

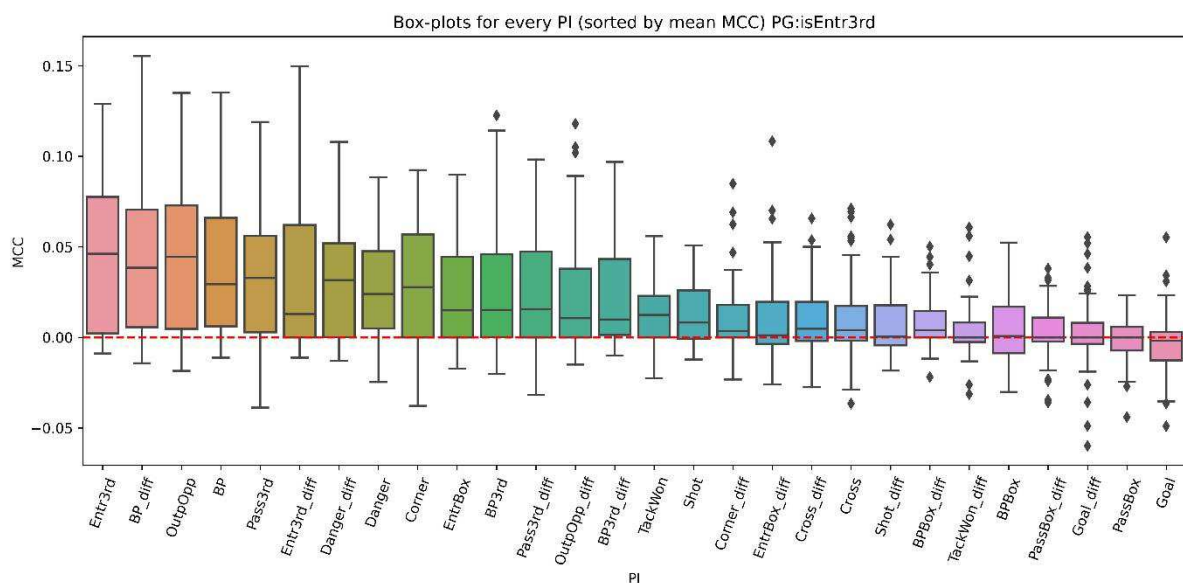


Figure 8. Ranking of PIs for PG_{isEntr3rd}. PIs are sorted by their mean MCC results in descending order. The red dotted line indicates MCC=0.

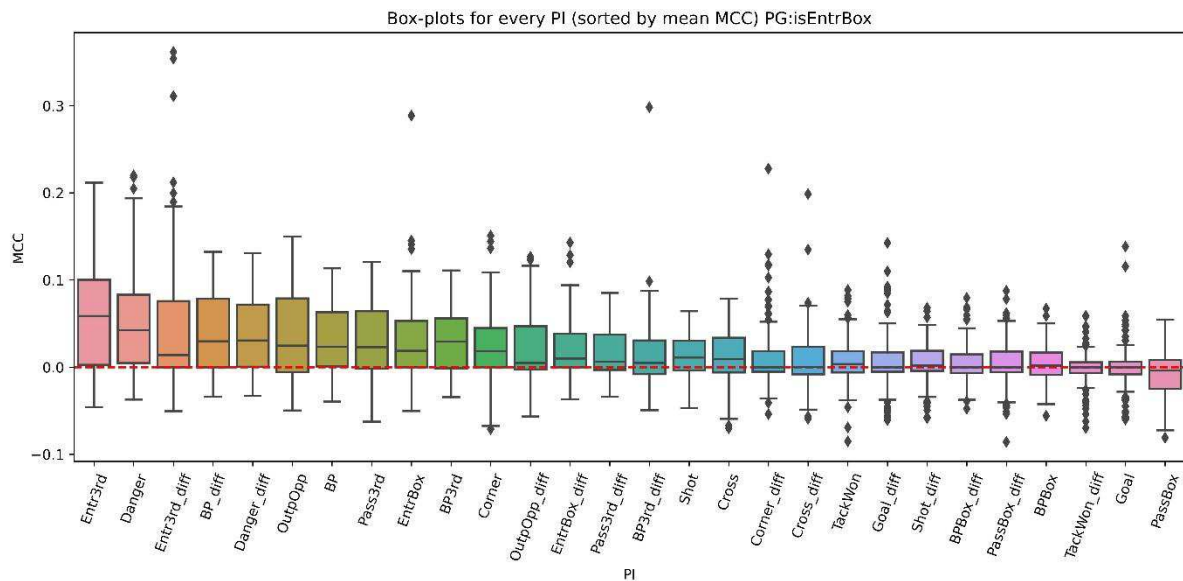


Figure 9. Ranking of PIs for $PG_{isEntrBox}$. PIs are sorted by their mean MCC results in descending order. The red dotted line indicates $MCC=0$.

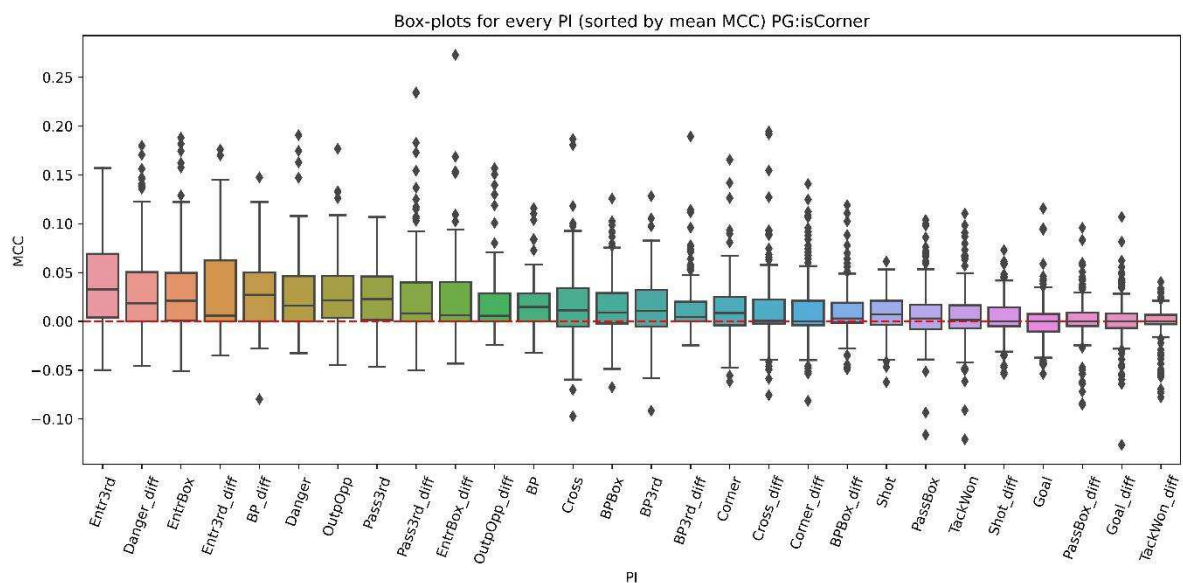


Figure 10. Ranking of PIs for $PG_{isCorner}$. PIs are sorted by their mean MCC results in descending order. The red dotted line indicates $MCC=0$.

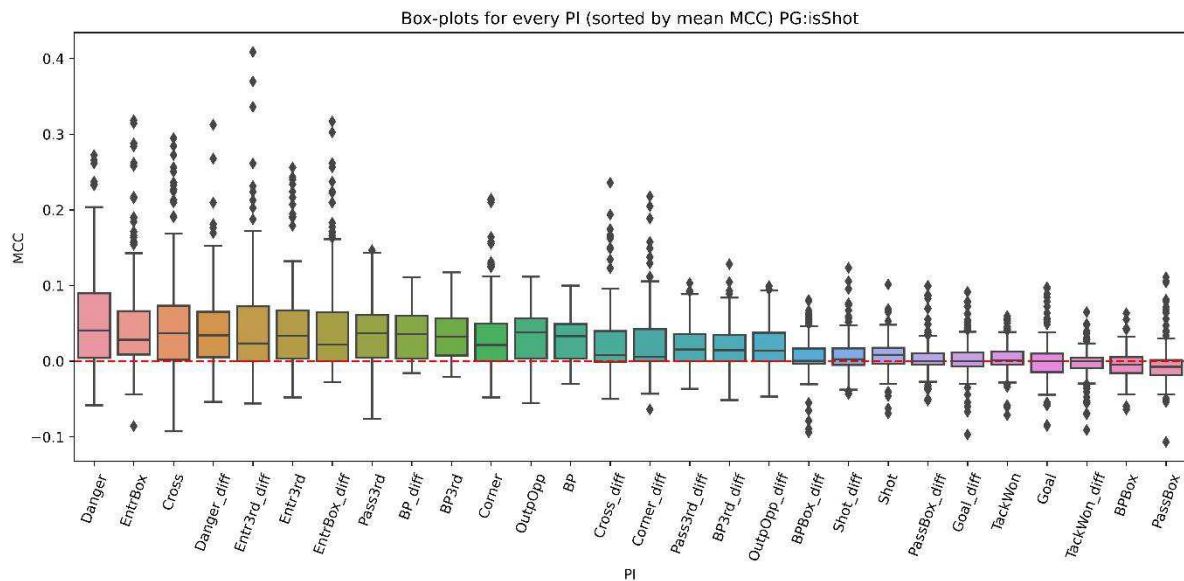


Figure 11. Ranking of PIs for PG_{isShot} . PIs are sorted by their mean MCC results in descending order. The red dotted line indicates $MCC=0$.

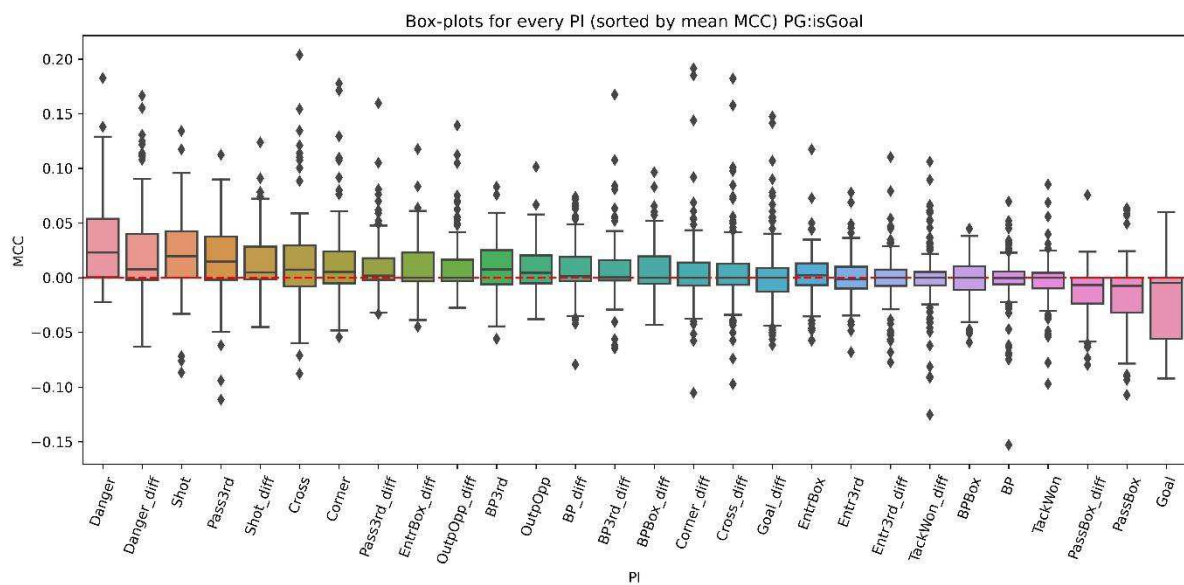


Figure 12. Ranking of PIs for PG_{isGoal} . PIs are sorted by their mean MCC results in descending order. The red dotted line indicates $MCC=0$.

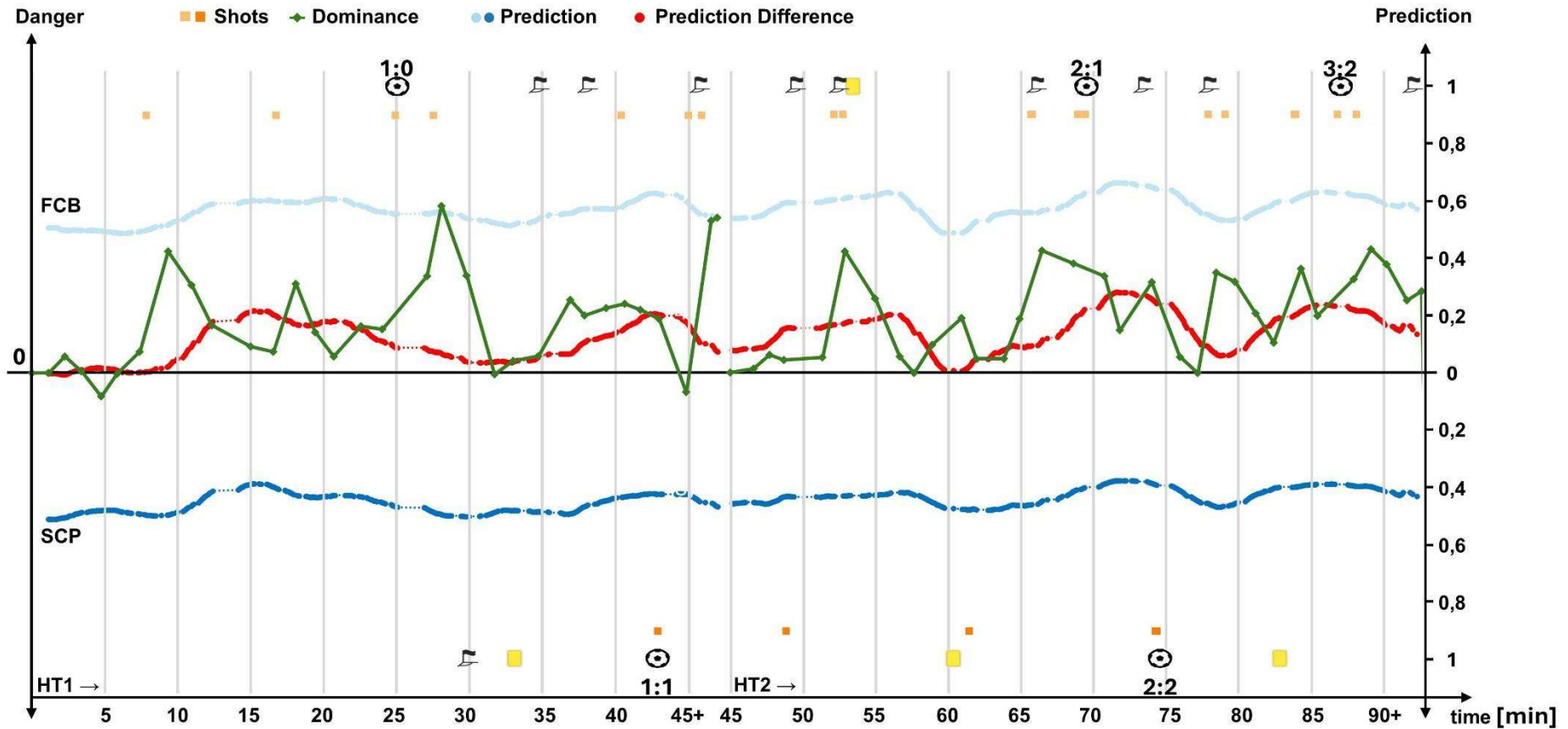


Figure 13. Application of the 1st ranked model for an unseen match between FC Bayern Munich (FCB) vs SC Paderborn (SCP) in Season 19/20 which resulted in a 3:2. In the upper half (FCB) and the lower half (SCP) for each team important events (corner kicks, given cards, goals scored, shots taken), Dominance by Link et al. (2016) and the prediction difference, as our proposed match momentum metric, are shown.

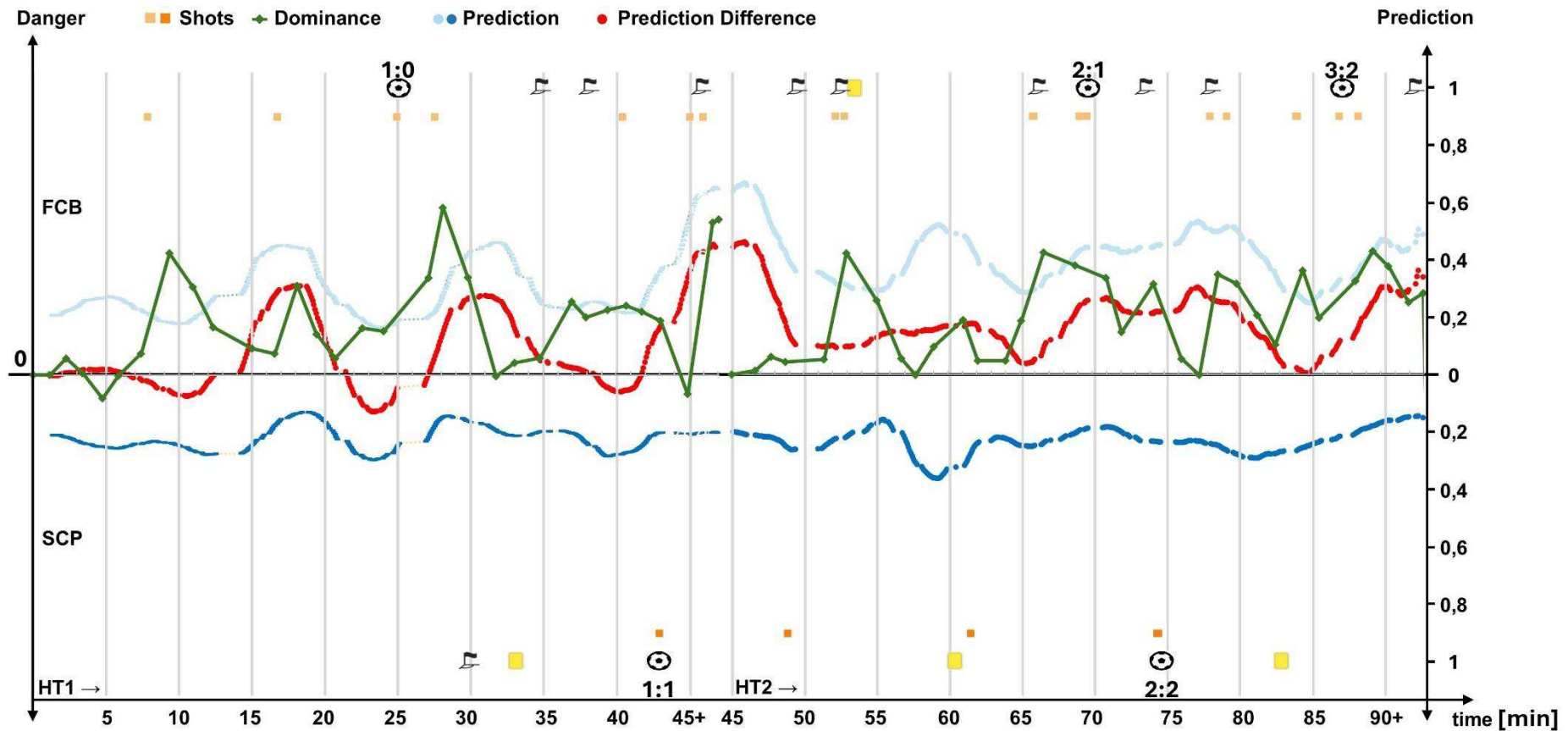


Figure 14. Application of the 2nd ranked model for an unseen match between FC Bayern Munich (FCB) vs SC Paderborn (SCP) in Season 19/20 which resulted in a 3:2. In the upper half (FCB) and the lower half (SCP) for each team important events (corner kicks, given cards, goals scored, shots taken), Dominance by Link et al. (2016) and the prediction difference, as our proposed match momentum metric, are shown.