Large Scale Analysis of Offensive Performance in Football

Using Synchronized Positional and Event Data to Quantify Offensive Actions, Tactics, and Strategies

Doctoral Thesis

in order to obtain the title of Doctor from the Faculty of Economics and Social Sciences at the University of Tübingen

presented by

M. Sc. Gabriel Anzer

München

Tübingen, 2021

1st supervisor: Prof. Dr. Oliver Höner
2nd supervisor: Prof. Dr. Augustin Kelava
3rd supervisor: Prof. Dr. Ulf Brefeld

Date of the oral defense: 13.01.2022
Dean: Prof. Dr. Josef Schmid
1st supervisor: Prof. Dr. Oliver Höner
2nd supervisor: Prof. Dr. Augustin Kelava
3rd supervisor: Prof. Dr. Ulf Brefeld

Acknowledgements

This dissertation is a part of a broader research program conducted by Eberhard Karls University Tübingen, DFB-Akademie, Deutsche Fußball-Liga (DFL) and Sportec Solutions AG. Another major pillar of this dissertation is the thesis of Pascal Bauer, which has been conducted in close collaboration.

First, I would like to thank Prof. Dr. Oliver Höner as the main supervisor of both theses guiding our research with critical and valuable feedback and for supporting us whenever necessary. Second, I would like to thank all co-authors involved, especially Pascal Bauer, Joshua Wyatt Smith (PhD) and Prof. Dr. Ulf Brefeld, for informative collaborations and discussions that build a central component of the work presented. Last but not least, I want to thank Prof. Dr. Augustin Kelava for guiding our research in various discussions.

Additionally, I would like to thank DFL, DFB-Akademie and Sportec Solutions AG for providing the positional and event data for the studies and supporting the dissertations.

This work would also not have been possible without the perspective of professional match-analysts and coaches from world class teams who helped us to define relevant features and spent much time evaluating (intermediate) results. Also, I would cordially like to thank Dr. Stephan Nopp, Christofer Clemens (head match-analysts of the German men's National team), Jannis Scheibe (head match-analyst of the German U21 men's national team), Leonard Höhn (head match-analyst for the women's national team) as well as Sebastian Geißler (former match-analyst of Borussia Mönchengladbach).

Contents

Abstract

Offensive performances in football have always been of great focus for fans and clubs alike as evidenced by the fact that nearly all Ballon d'Or winners have been forwards or midfielders. With the increase in availability of granular data, evaluating these performances on a deeper level than just goals scored or gut instinct has become possible. The domain of sports analytics has recently emerged, exploring how applying data science techniques or other statistical methods to sports data can improve decision making within sporting organizations. This thesis follows the footsteps of other sports like baseball or basketball where, at first, offensive performances were analyzed. It consists of four studies exploring various levels of offensive performance, ranging from basic actions to team-level strategy. For that, it uses a dataset part of larger research program that also explores the automatic detection of tactical patterns. This dataset mainly consists of positional and event data from eight seasons of the German Bundesliga and German Bundesliga 2 between the seasons 2013/2014 and 2020/2021. In total this amounts to 4,896 matches, with highly accurate player and ball positions for every moment of the match and detailed logs of every action that occurred, thus making it one of the largest football datasets to be analyzed at this level of granularity. In a first step, this thesis shows how the two different data sources can be synchronized. With this synchronized data it is possible to better quantify individual basic actions like shots or passes. For both actions new metrics (Expected Goals and Expected Passes) were developed, that use the contextual information to quantify the chance quality and passing difficulty. Using this improved quantification of individual actions, the subsequent studies evaluate offensive per-

formance on a tactical pattern level (how goals are scored) and on a strategy level (what team formations are particular effective offensively). Besides their usage on the performance side, these metrics have also been adapted from broadcasters to enhance their data story telling: Expected goals and expected passes are shown during every Bundesliga match to a worldwide audience, thus bringing the field of sports analytics to millions of fans.

1 Introduction

Extracting insights from data has become an integral part of everyday life in nearly all industries ranging from recommendation systems to sports. However, even though football is by far the most popular sport in the world, in other sports, namely baseball and basketball, data analytics was able to change the nature of the game much sooner (Lewis, 2003; Oliver, 2004). In baseball and basketball, relatively simply collected *play-by-play* data about offensive performances was used as a basis of analysis and gave indications of inefficiencies, such as over-valuing home runs or two point attempts. Only about ten years later, the first main-stream work describing advanced analytics in football appeared (Anderson & Sally, 2013). While there were studies exploring the use of data in football before that (Reep & Benjamin, 1968; Gould & Gatrell, 1979; Borrie, Jonsson, & Magnusson, 2002), the quality and granularity of the available data is largely to blame for the slow acceptance in the football industry.

In football, the initial available data only described what was happening with or near the ball. This so-called *event data* (equivalent to play-by-play data in basketball or American football) does not capture what is happening off-the-ball, such as the positions of the remaining players. At first, this data was manually collected for individual studies, but due to the growing interest, several companies started collecting this event data across multiple professional leagues (Lucey, Oliver, Carr, Roth, & Matthews, 2013). In the past several years, a new data type, the so-called *tracking data*, often also referred to as movement data, positional data, or trajectory data has become increasingly available (Seidl, 2019). This tracking data captures the positions of all players (and typically also of the ball) at any moment of the game. This is

either done through local or global positioning systems (LPS/GPS), or through computer vision algorithms (Manafifard, Ebadi, & Moghaddam, 2017; Stein et al., 2017).[1] While LPS/GPS-data is often cheaper to collect and includes additional data like heart rate, its practicality in the tactical sense is somewhat limited, because it would require the opponent to wear the same gear and would still be missing the ball (Goes, Meerhoff, et al., 2020; Buchheit et al., 2014).

Similar to baseball and basketball, the majority of analytics research in football is focused on the offensive side (Reep & Benjamin, 1968; Gould & Gatrell, 1979; Borrie et al., 2002; Anderson & Sally, 2013). This has several reasons: (1) For the longest time, the only widely available data was event data, which consists almost entirely of offensive actions. (2) Media and club interest is biased towards offensive players, as evidenced by the fact that the past 13 Ballon d'Or winners have been midfielders or forwards,[2] as well as the 20 most expensive players are offensive players.[3] (3) Defensive performance is simply harder to evaluate conceptually: while positive offensive performance often corresponds to concrete actions or results, good defensive performance leads to the absence of opposing ones. For these reasons, this dissertation follows the footsteps of other sports and focuses on measuring offensive performance.

There have been several attempts at categorizing performance in football into different levels (Rein & Memmert, 2016; Gréhaigne, Godbout, & Bouthier, 1999; Q. Wang, Zhu, Hu, Shen, & Yao, 2015). Rein and Memmert (2016) differentiates between individ-

[1] Most providers use a set of up high definition cameras installed on-site to deliver highly accurate data, but recent developments, also allow for lower budget options, which work purely based on broadcast videos.

[2] https://www.francefootball.fr/ballon-d-or/palmares/

[3] https://www.transfermarkt.com/marktwertetop/wertvollstespieler (accessed Oktober 8, 2020)

ual tactics (tactical actions conducted by a single player), group tactics (collective tactical actions conducted by a subgroup of players), team tactics (describing the formation of a team), and game tactics (the team's playing philosophy). Moreover, they claim that a clear distinction between tactics and strategy is challenging, since any real-time interaction will be prone by the a priori strategy. Gréhaigne et al. (1999) on the other hand simply differentiate between strategy, defined as the a priori plan of a team, and tactics, defined as the decisions made during a game. Since there is no universal categorization, for the purpose of this thesis we burrow concepts from the literature and define three different levels of offensive performance: *basic offensive actions*, *offensive tactical patterns*, and *offensive team strategy* (see Figure 1). On the highest level there is the overarching *offensive team strategy*, which is typically set before each match, e.g. in which formation a team plans to attack. This strategy influences a team's *offensive tactics* or *offensive tactical patterns* defined as a repeatable and coordinate set of basic offensive actions during a match. This definition includes for example goal scoring patterns or build-up play. And finally, the *basic offensive actions* are single actions performed by an individual with the intent to increase the likelihood of a team to score and mostly consist of the event data (e.g. passes, shots), but can also include actions not involving the ball like offensive off-ball runs. The proposed definitions are not meant as a precise categorization, but should rather serve as guideline to frame our work and to highlight at which levels offensive performance can be measured.

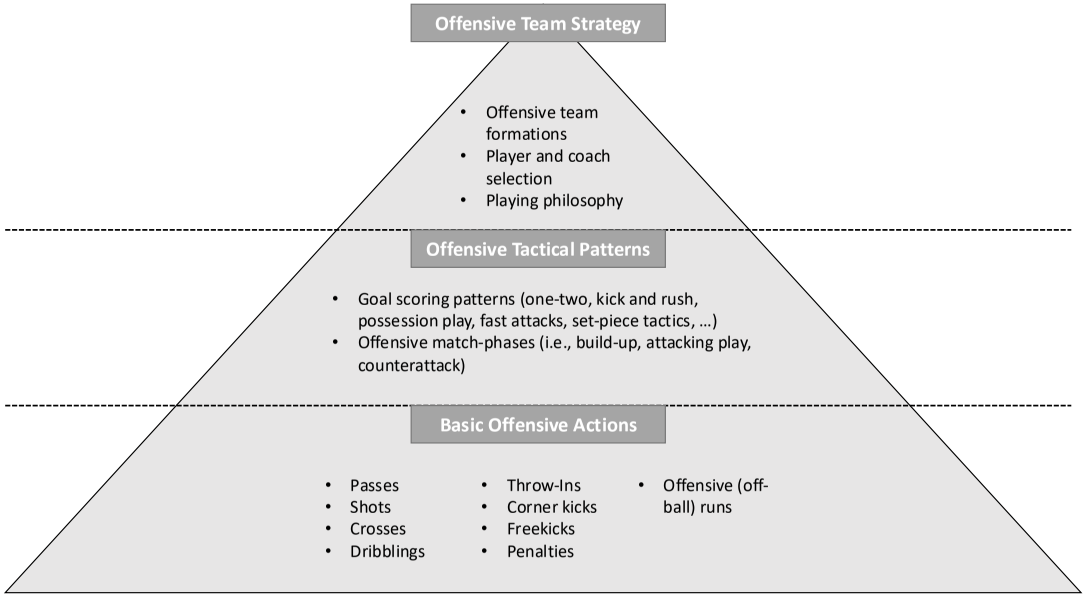

Figure 1: Overview of the different levels of offensive performance with examples for each category.

Early research in football focused especially on goals (Reep & Benjamin, 1968; Pollard & Reep, 1997). Due to low-scoring nature of football, the attention soon shifted towards more frequently occurring events, such as shots or shots on target (Tenga, Holme, Ronglan, & Bahr, 2010). One of the most established advanced metrics in football are the so-called expected goal values (xG's) (Hedar, 2020; Rowlinson, 2020; Robberechts & Davis, 2020; Lucey, Bialkowski, Monfort, Carr, & Matthews, 2014). They estimate the likelihood of a shot being converted to a goal and thus allow for a more granular analysis of the chance quality. Another highly investigated basic offensive action are passes, as they are by far football's most frequently occurring events. Similar to shots, at first research looked at how the number of passes or pass completion rates correlated to wins on a match level (Bradley, Lago-Peñas, Rey, & Diaz, 2013; Król et al., 2017). Later, event-level data was used to analyze passes in more detail (Łukasz Szczepański & Mchale, 2016; McHale & Relton, 2018). Brooks, Kerr, and Guttag (2016), Power, Ruiz, Wei, and Lucey (2017) and Bransen and Haaren (2019) used event-level data to estimate the difficulty of passes, whereas Steiner, Rauh, Rumo, Sonderegger, and Seiler (2019) for example, tried to estimate the reward of a pass. Moreover, several studies aimed to assign value to individual offensive actions either based on event data (Decroos, Haaren, Bransen, & Davis, 2019) or on tracking data (Spearman, Basye, Dick, Hotovy, & Pop, 2017; Power et al., 2017; Fernández, Bornn, & Cervone, 2020; Arbues-Sanguesa, Martin, Fernandez, Ballester, & Haro, 2020; Alguacil, Fernandez, Arce, & Sumpter, 2020; Stöckl, Seidl, Marley, & Power, 2021). One such metric that found wide media coverage is the so called "packing" metric (Steiner et al., 2019).

Because of the increased complexity, offensive tactical patterns

are typically analyzed using tracking data, but there exists some research purely using event data (Gudmundsson & Horton, 2017; Decroos, Haaren, & Davis, 2018). For instance, Gudmundsson and Horton (2017) used event-level data to analyze the effectiveness of long-ball build ups. The studies based on tracking data range from corner kick tactics (Shaw & Gopaladesikan, 2021; Bauer, Anzer, & Smith, 2022) to pitch control patterns (Martens, Dick, & Brefeld, 2021; Brefeld, Lasek, & Mair, 2019; Fernandez & Bornn, 2018). The accompanying work in Bauer (2021) goes into further details regarding the automatic detection of tactical patterns (both offensive ones as well as defensive ones). It first presents an overview of this research area across different sports, then it specifies tactical patterns in football and finally applies machine learning methods to identify them automatically.

Similar to tactical patterns, offensive team strategy has mostly been investigated based on tracking data (Andrienko et al., 2019; Carling, 2011; Müller-Budack, Theiner, Rein, & Ewerth, 2019; Bialkowski et al., 2016; Lucey et al., 2013). Using this positional data Lucey et al. (2013) found that away teams play with a more conservative strategy, in part leading to the home court advantage. Müller-Budack et al. (2019) classified positional data from four matches to predefined formation templates and found that offensive formations are particularly hard to recognize. Before the increased availability of tracking data Pollard and Reep (1997) used event-level data to compare the effectiveness of different offensive strategies.

The interest in this research area is not only of academic nature, but also stems from clubs or federations. Nearly all professional football organizations have incorporated data in some form to their daily processes. This includes the scouting of talented players, opposition analysis, monitoring physical perfor-

mances, preventing injuries, or predicting youth player developments. As a consequence, clubs and federations hire data-scientists and establish dedicated data analytics departments to support decision-making on strategy, tactics, and player recruitment (Andrienko et al., 2019). Additionally, media companies are interested in using analytics to enhance their storytelling to deliver more data-based facts to their customers (Link, 2018b). The German Bundesliga in particular has bought in to this concept. Since May 2020 its broadcast delivers some of the advanced metrics developed within this dissertation to millions of fans world wide.[4] Furthermore, as Tuyls et al. (2021) noted, football data can be considered as the ideal playing ground to develop and test new machine learning methods.

For most of the above listed applications, the offensive performance is especially relevant, as offensive players are generally more valuable than defensive players monetarily, fans tend to celebrate goals more than defensive clearances, and from a practical stand point almost all available event data describes offensive ball actions. Even though we listed several studies evaluating offensive performance on every level using either event or positional data, the research field is missing ones that combine the two sources.

The purpose of this thesis is to address this shortcoming by introducing a novel synchronization algorithm of event and tracking data. It then follows the footsteps of other sports and develops frameworks to objectively measure offensive performances on every level. The synchronization allows for the quantification of the two most frequent basic offensive actions, shots and passes. The thesis then shows how using these enhanced actions

[4] https://www.bundesliga.com/en/bundesliga/news/new-real-time-match-analysis-dfl-and-amazon-web-services-11246

one can better investigate repeating tactical offensive scoring patterns as well as macro level offensive strategies.

Building on this introductory section, which details the history and status quo of sport analytical research investigating offensive performance, Section 2 introduces the two data sources used throughout the thesis in detail with a focus on presenting our novel approach of how to synchronize them. It further presents key concepts in machine learning and describes a few supervised and unsupervised learning methods later used in the empirical studies.

Section 3 builds the core of this thesis, summarizing four empirical studies either establishing or applying novel offensive metrics. The corresponding manuscripts were submitted or have already been published by internationally well renown sport science journals and consist of:

(I) Anzer, G., Bauer, P. (2021). A Goal Scoring Probability Model based on Synchronized Positional and Event Data. *Frontiers in Sports and Active Learning (Special Issue: Using Artificial Intelligence to Enhance Sport Performance)*, 3(0), 1–18. https://doi.org/10.3389/fspor.2021.624475

(II) Anzer, G., Bauer, P. (2022). Expected Passes—Determining the Difficulty of a Pass in Football (Soccer) Using Spatio-Temporal Data. *Data Mining and Knowledge Discovery, Springer US*. https://doi.org/10.1007/s10618-021-00810-3

(III) Anzer, G., Bauer, P., & Brefeld, U. (2021). The Origins of Goals in the German Bundesliga. *Journal of Sport Science*. https://doi.org/10.1080/02640414.2021.1943981

(IV) Bauer, P., Anzer, G., & Shaw, L. (2022). Putting Team Formations in Association Football into Context. *Journal of*

Sports Analytics (submitted).

In Section 4, this dissertation critically assesses the results of the empirical studies and their scientific merit, discusses their limitations, and names potential future improvements or applications.

Of course, defense and group tactical elements are also very relevant aspects of football. These were addressed in various other studies conducted as part of this research program, but not included in the core of this thesis.

(i) Andrienko, G., Andrienko, N., Anzer, G., Bauer, P., Budziak, G., Fuchs, G., Hecker D., Weber H., Wrobel, S. (2019). Constructing Spaces and Times for Tactical Analysis in Football. *IEEE Transactions on Visualization and Computer Graphics*, 27(4), 2280–2297. https://doi.org/10.1109/TVCG.2019.2952129

(ii) Bauer, P., Anzer, G., Smith, J. W. (2022). Individual role classification for players defending corners in football (soccer). *Journal of Quantitative Analysis in Sports (submitted)*.

(iii) Bauer, P., Anzer, G. (2021). Data-Driven Detection of Counterpressing in Professional Football—A Supervised Machine Learning Task based on Synchronized Positional and Event Data with Expert-Based Feature Extraction. *Data Mining and Knowledge Discovery*, 35(5), 2009–2049. https://doi.org/10.1007/s10618-021-00763-7

(iv) Fassmeyer, D., Anzer, G., Bauer, P., Brefeld, U. (2021). Toward Automatically Labeling Situations in Soccer. *Frontiers in Sports and Active Living*, *3(November)*. https://doi.org/10.3389/fspor.2021.725431

(v) Link D., Anzer G. (2021). How the COVID-19 Pandemic has Changed the Game of Soccer. *International Journal of Sports Medicine*. https://doi.org/10.1055/a-1518-7778

(vi) Szymski D., Anzer, G., Alt V., Gärtner B., Krutsch W., Weber H., Meyer T. (2021). Contact times in professional football before and during the SARS-CoV-2 pandemic: Tracking data from the German Bundesliga. *European Journal of Sport Science*. https://doi.org/10.1080/17461391.2022.2032837

(vii) Anzer, G., Bauer, P., Höner, O. (2021). The Identification of Counterpressing in Football. In D. Memmert (Ed.), Match Analysis—How to Use Data in Professional Sport (1st Editio, pp. 228–235). *New York: Routledge*. https://doi.org/ https://doi.org/10.4324/9781003160953

In an exploratory study i (Andrienko et al., 2019) we used visual analytics to find repeating tactical patterns. Studies ii and iii (Bauer, Anzer, & Smith, 2022; Bauer & Anzer, 2021) explore how one can automatically identify the certain defensive tactical patterns, namely corner marking and counterpressing. Study iv (Fassmeyer, Anzer, Bauer, & Brefeld, 2021) uses variational autoencoder to later automatically identify actions or patterns with little labeled data. Studies v and vi (Link & Anzer, 2021; Szymski et al., 2021) used tracking data to evaluate contact times following the COVID-19 pandemic to estimate the risk associated with a restart of a league and explored how much the game has changed afterwards. Additionally, in a book contribution vii (Anzer, Bauer, & Höner, 2021) a general overview of how machine learning is used in football is given.

2 Data and Methods

2.1 Combining Positional and Event Data in Football

Typically, the gathered data during football matches consists of three raw data sources: *meta data*, *event data*, and *tracking data*. Figure 2 shows the three data sources and their derived data types.

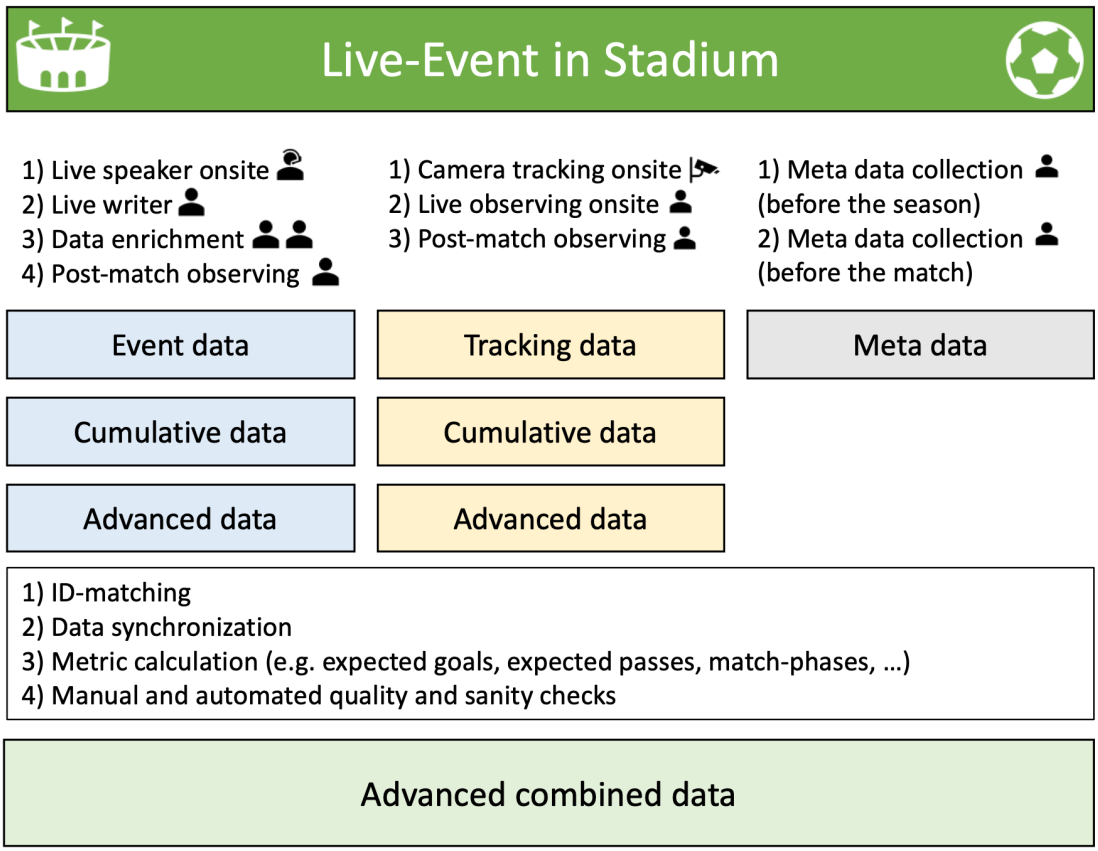

Figure 2: Overview of the different types of raw data sources, their collection process, and their potential for derived statistics. Further the figure lists some of the necessary steps to combine the data sources.

2.1.1 Meta Data

Meta data contains a wide variety of information (mostly) available before the match starts. It includes details about the teams (team names, jersey colors, starting and bench players etc.), the players (birth dates, nationalities, height, weight etc.),

the stadium (address, capacity, attendance etc.), the pitch (dimensions, pitch type etc.), the weather conditions (temperature, precipitation, humidity), the referees (names, age etc.), and the team staff (role, name, age etc.). All this information is gathered manually, either once per season, or once per match. While this data is more often used to give context to analysis of the other two data sources, based on this meta data Link and Weber (2017) analyzed the effect of weather conditions on match results and Brander, Egan, and Yeung (2014) focused on how player age affects their performance (in ice hockey). While not present in the dataset used in this thesis, meta information could also include various other attributes of interest, such as player value estimations, player salaries, or contract lengths.[5]

2.1.2 Event Data

Event data consists of a log of actions happening during the game, often including attributes describing each action in more granularity. This set of actions, attributes, and their definitions vary depending on the collecting company making comparisons between datasets from different providers very difficult.[6] Event data can be divided into three distinct categories: player actions, team actions, and referee actions. Player actions contain all on-ball actions where at least one player touches the ball (e.g. passes, crosses, shots, tackles, ...) or rule violations (e.g. fouls, offsides, ...). Team actions mostly refer to situations where a team is granted a set-piece, e.g. corner kicks, free kicks, penalties, throw-ins. Lastly, referee actions contain all actions where the referee is the primary actor, e.g. cautions or referee balls.

For each event various sub-attributes are collected, some present

[5] www.transfermarkt.de
[6] https://dtai.cs.kuleuven.be/sports/blog/how-data-quality-affects-xg

for all actions (e.g. time-stamps, x/y-location on the pitch, ...) and most dependent on the event type. For example, for every pass its receiver, the pass height (low or high), the direction, and several more details are collected. The full catalogue including the definitions of each event type and attribute are propitiatory to the companies collecting the data, but simple versions are described in the literature (Bialkowski et al., 2016; Pappalardo et al., 2019; Stein et al., 2017). Figure 3 shows an excerpt from a former version the German Bundesliga used. This type of data is also systematically collected in many other sports with different catalogues nowadays (Vračar, Štrumbelj, & Kononenko, 2016).

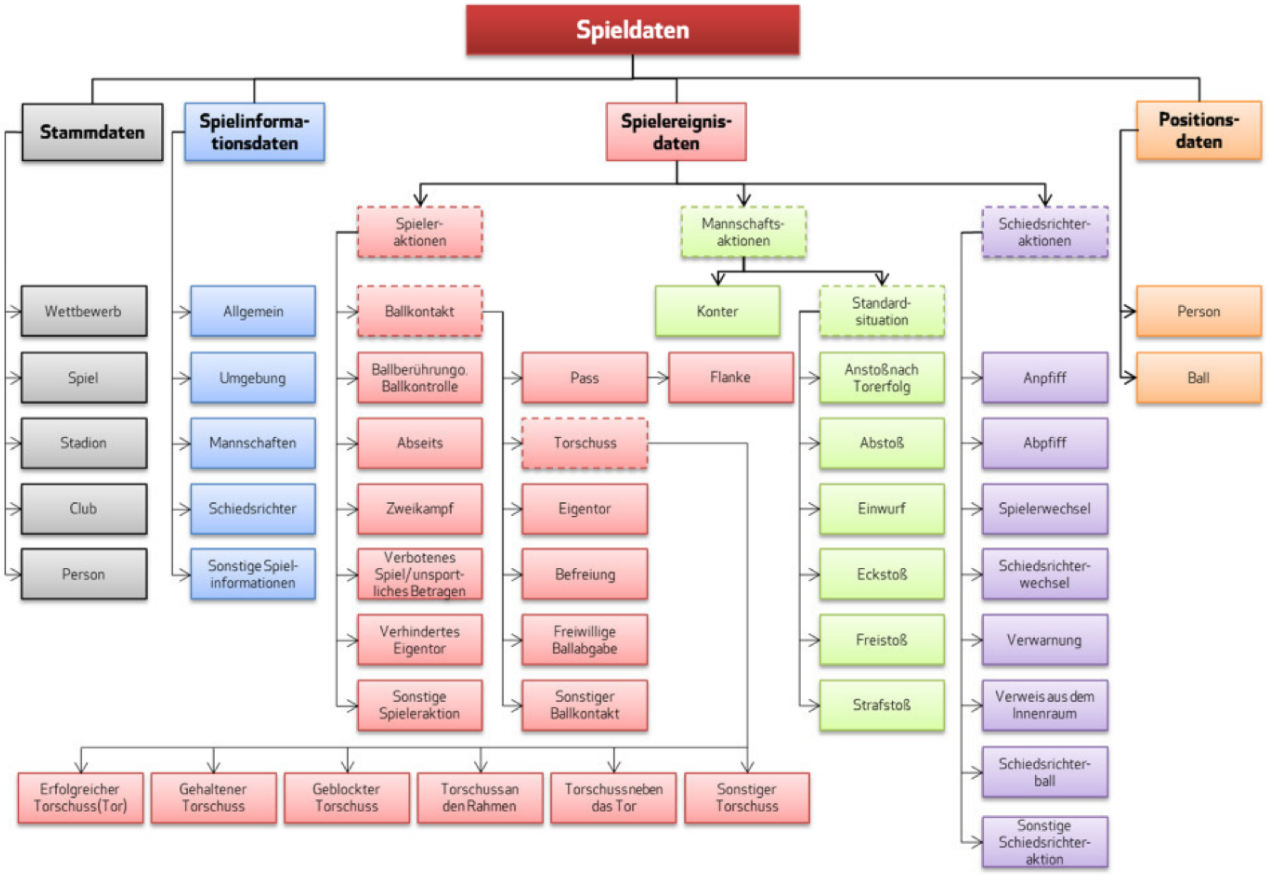

Figure 3: Excerpt (in German) from a former version of the DFL definitions catalogue (https://s.bundesliga.com/assets/doc/10000/2189_original.pdf). It includes all the event categories that were collected.

The Bundesliga requires its event data to arrive with a low latency and high depth and accuracy. To achieve the low latency the current provider uses a setup that consists of live speakers onsite in the stadiums that communicate the most important actions to live writers in a offsite collection center where they record this data in the collection software.[7] In this collection center there are additional observers using a video stream to enrich the data with further information (e.g. body part used during a shot). Finally, after a match has concluded, further details are added (e.g. the location of every pass) and quality control on the live collected data is performed. But different methodologies and tools exist for the annotation of event data across different companies, i.e. other companies encode the match purely from video footage into their own dedicated software systems (Pappalardo et al., 2019).

Sequential events by the same team are often joined together as one possession sequence, but there are varying definitions, if a short touch by the opposing team (e.g. deflection) should interrupt a possession sequence or not (Stein et al., 2017).

The most trivial use of this raw event data is the cumulative aggregation, either on a team or a player level, often also referred to as *frequency analysis* (Chawla, Estephan, Gudmundsson, & Horton, 2017; Borrie et al., 2002; Sarmento, Anguera, Campaniço, & Leitão, 2010). This cumulative data simply shows the frequency of certain events occurring over a given time frame, e.g. how many shots a team had in the first half or how many successful passes a player achieved in a match.

Further, various advanced metrics purely based on event data were developed. Decroos, Bransen, van Haaren, and Davis (2020)

[7]The so-called speaker-writer method is explained here: https://www.dfl.de/en/innovation/how-is-the-official-match-data-collected/.

compute the value of each action, Roy, Robberechts, chi Yang, Raedt, and Davis (2018) measure players' decision-making skills, and Bransen and van Haaren (2020) analyze team chemistry. A limitation of the event data is, that it doesn't capture the position of the remaining players, so individually calculated values may not always be very precise, but over large sample sizes (e.g. for scouting players), these advanced metrics are a helpful source of information (Decroos et al., 2019).

2.1.3 Tracking Data

Tracking data, contrary to event data, captures what is happening on the entire pitch throughout every moment of the match automatically. The tracking data used throughout this thesis was collected using the Chyhron Hego Gen 4/5 system.[8] It computes the center of gravity of all players, referees, and the ball and transforms them into a two-dimensional Cartesian coordinates system. The ball data includes a third dimension, the ball height (in meters). The data is recorded at a frequency of 25 Hz. In other words, one second worth of tracking data consists of 25 frames. Each frame also contains the derived values *distance covered since the preceding frame*, current *speed*, and *acceleration* values for every object (ball or person). These are computed using a 5th-order 1.0- Hz Butterworth filter to smooth the data and remove outliers. The automatically gathered data is enhanced by two manually collected attributes: *ball possession* and *status of play* for every frame. Ball possession indicates which team is in possession of the ball. A team's ball possession starts when one of its players touches the ball for the first time after an opposing ball possession phase and ends as soon as the ball is out of play or an opposing player controls the ball. Status of play describes if the

[8] https://tracab.com/products/tracab-technologies/

ball is in play or not. The latter is case if the referee has halted the game (e.g. due to a foul) or if the ball has left the pitch. This manually gathered data is collected by a live observer within the stadium, who is also responsible for initial player assignments as well as resolving potential player swaps. After the match a post-match observer manually corrects further issues with the tracking data (e.g. unrealistic ball paths).

Based on this tracking data, one can easily compute aggregated cumulative data. The most prominent aggregated statistic being the total distance covered by a player (Andrzejewski, Chmura, Pluta, & Konarski, 2015), but top speed, number of sprints, non-interrupted playing time, or percentage of possession are further examples (Link, 2018a).

Several advanced statistics purely based on tracking data have been developed in the literature (Martens et al., 2021; Fernandez & Bornn, 2018; Andrienko et al., 2017; Link, Lang, & Seidenschwarz, 2016) as well as in media coverage, like average positions,[9] pressure,[10] or attacking directions.[11]

For more details regarding the history and studies validating the positional data, see the accompanying work of Bauer (2021).

2.1.4 General Description of the Data

This thesis is built on data owned and collected by the Deutsche Fußball Liga (DFL). For consistency purposes it has developed its own catalogue of definitions[12] and requires all data providers to record data according to these definitions. Event and meta data have been collected by Sportec Solutions AG[13] and the tracking

[9] https://www.bundesliga.com/en/bundesliga/news/match-facts-dfl-aws-revamp-average-positions-trends-14706

[10] https://aws.amazon.com/de/sports/bundesliga/most-pressed-player/

[11] https://aws.amazon.com/de/sports/bundesliga/attacking-zones/

[12] https://s.bundesliga.com/assets/doc/10000/2189_original.pdf

[13] https://www.sportec-solutions.de/

data stems from Chyron Hego's TRACAB system.[14] The dataset used throughout the research program consists of all German Bundesliga and German Bundesliga 2 matches, spanning across eight seasons between the 2013/2014 and 2020/2021 seasons. This totals to 4,896 matches, making it one of the largest collections of event and tracking data in the literature. The event definitions catalogue of the DFL contains 34 different event types with in total 123 unique attribute categories. On average, for each event 20 attributes are gathered. Over the entire dataset there were 2.85 goals, 26.19 shots, and 894.25 passes (78.46% of them successful) per match. Even though this is one of the largest sets of event data, a single game of tracking data contains more information than an entire season worth of event data. As noted above, the tracking data of a single game consists of at least 135,000 frames (90 minutes, recorded at a frequency of 25Hz) plus a varying number of frames collected during added time, typically with over 130 attributes per frame. This means that one season of tracking data requires about 2.3 Terabyte of storage, which also leads to computational challenges. Over the data set the gross playing time is on average 94:39min, while the net playing time (defined as the total time when the tracking data status of play is set to in play) is only 55:10min. All this data is collected live during the matches and underlies extensive quality control loops to ensure a high data quality. In a standardized process called *observing*, a human operator manually checks and corrects suspicious sequences of player trajectories using a dedicated software. The accuracy of the used tracking data has been evaluated in the literature (Linke, Link, & Lames, 2020). While this proprietary dataset cannot be shared, there exist small open-source datasets containing either event (Pappalardo et al., 2019) or posi-

[14] https://chyronhego.com/wp-content/uploads/2019/01/TRACAB-PI-sheet.pdf

tional data (Pettersen et al., 2014) that can be used to reproduce approaches presented throughout this thesis on a smaller scale. Furthermore, there exist open-source sample datasets directly released by event[15] or tracking[16] data providers.

Apart from the sheer size, the dataset comes with a few notable other challenges. While the definitions of the events remained mostly constant throughout the seasons, several new attributes were added in between and not retroactively collected. Moreover, as advances in computer vision and optical tracking system hardware led to constant improvements of the positional data quality, performing longitudinal analysis became more complicated. Additionally, both data sources are susceptible to occasional quality issues: manually collected event data is prone to human errors (even though the extensive quality control systems limit these) and the automatic optical tracking data collection can be affected by occlusions or strong weather effects (e.g. heavy snow, fog, ...). But the largest challenge when combining both datasets is that the manual collected time-stamps depend on human reaction time and therefore may deviate a lot from when it actually happened in the tracking data.

2.2 Synchronization of Event and Positional Data

Reconstructing a match based on event data is like looking at a completely dark pitch for 90 minutes, where a spotlight flashes at the ball every four seconds. In tracking data, this pitch is always light up, but it doesn't say when the relevant moments of the games are, which should be looked at more closely. Combining the two data sources is essential for unlocking the full

[15]Statsbomb (https://github.com/statsbomb/open-data)

[16]Skillcorner (https://github.com/SkillCorner/opendata) and Metrica Sports (https://github.com/metrica-sports/sample-data)

advantages of both datasets: when the relevant actions occurred and what they were (recorded at a greater level of detail than would be possible with tracking data alone), and spatio-temporal context of what happened before, at, and after the action. This information gain is shown in Figure 4a): Only using event data, all one would know is, that there was a successful pass starting in the middle of the pitch in the direction of the opposing goal. However, when we add the positional data in Figure 4b) of all the players at the time of the pass and the following ball trajectory, we can see that this short pass played under pressure, by-passed nearly the entire defending team and the receiving teammate (#23) is in a position where there is only the goalkeeper to beat. Contrast that to a different pass depicted in Figure 4c): in the event data this pass would look almost identical to the previous one based on starting and end location, but the combination with tracking data shows that it was merely a pass into the half space between the two defensive lines.

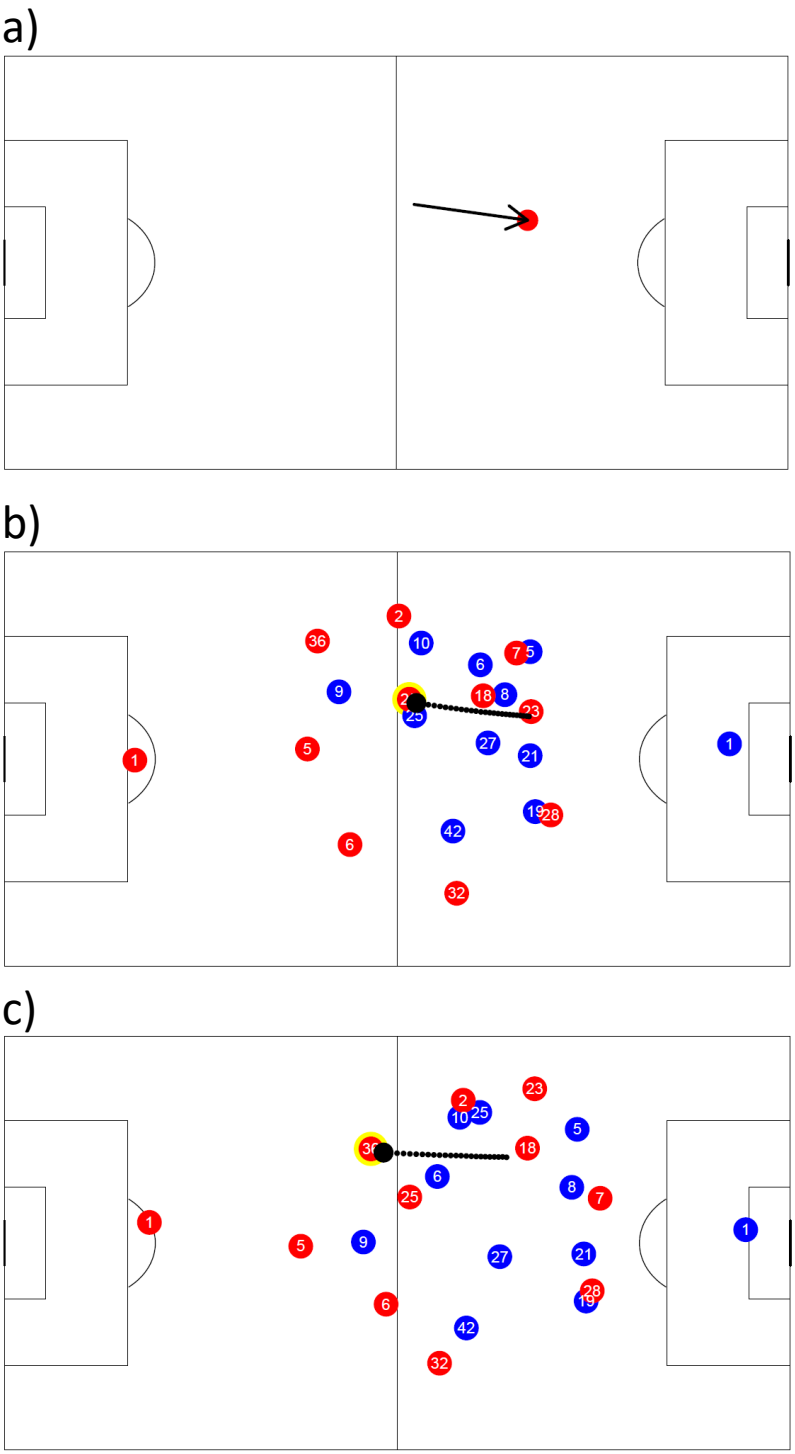

Figure 4: Visualization of two different passes: Sub-figures a) and b) show the same forward pass, once purely based on information contained in the event data and once including all the tracking data. Sub-figure c) depicts a different forward pass in a similar location, but the tracking data information shows a very different context than in b).

The biggest challenge when combining the data sources is that they are generally not aligned. This is mainly caused by the following two factors:

(a) The two data types are typically collected by different companies each using their own internal clock causing a systematic offset.

(b) Manually collected event time stamps are affected by reaction times, distractions, and decision times of the human operators.

For these reasons, a "naive" synchronization—using the time stamp from the event data—to identify player positions at the time of an event leads to large inaccuracies. Figure 5 shows the differences between time stamps included in the event data and calculated ones.[17] As can be seen, there are large differences (up to 20 seconds) between the two time points and hence the need for an accurate synchronization algorithm. To the author's best knowledge, this dissertation and the contained studies are the first to introduce a methodology that reliably solves this problem.

The general idea of our synchronization algorithm is to take all relevant information from the event data of an action and find the moment in the tracking data that most closely resembles it. Since there is no large set of ground truth data (with highly accurate time stamps), we chose to approach this problem with a rule-based solution and optimized the parameters used in an iterative process instead of using a machine learning based approach. The algorithm uses a general framework and adapts it slightly dependent on the event type. The exact details of the algorithm for shots and passes can be found in Anzer and Bauer (2021) and Anzer and Bauer (2022). This algorithm is slightly altered for other event types, e.g. when matching interceptions/ball receptions, the algorithm looks for the beginning

[17] As shown in the manual validation study within (Anzer & Bauer, 2022), the calculated timestamps generally capture the ground truth very well.

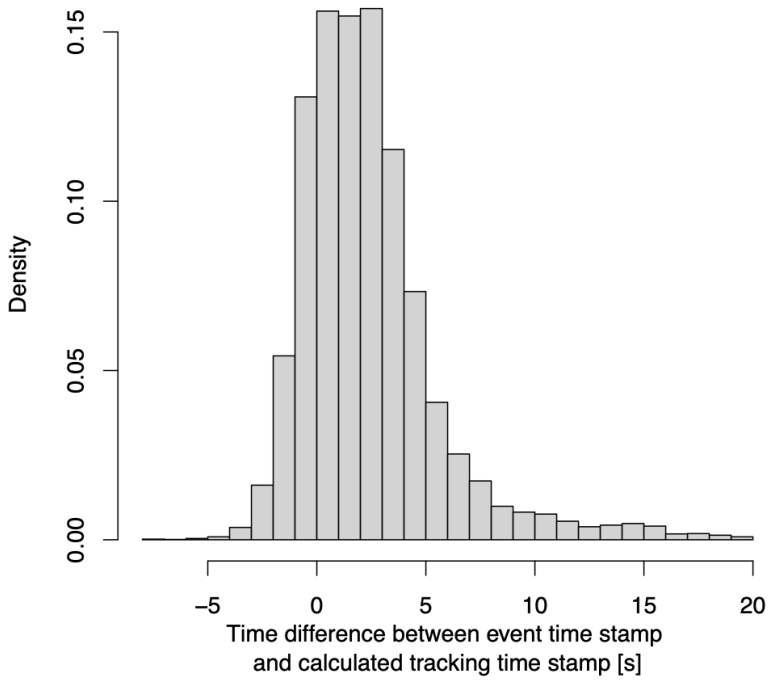

Figure 5: Time difference between event and tracking timestamps as presented in Anzer and Bauer (2021) (p. 6).

of an individual ball possession phase instead of its end and for duels, it also takes the distance between both involved players and the ball into consideration. With this synchronization we create a completely new dataset as shown in the bottom of Figure 2.

2.3 Machine Learning Basics

Due to the growing amount of data available in virtually any domain paired with the increasing affordability of computing power, the research area of machine learning experienced tremendous advances in the past decades. Following the general objective of machine learning, early approaches aimed to teach algorithms playing board games (Samuel, 1959). Nowadays, machine learning applications are ubiquitous across various domains sup-

porting human processes in image recognition, natural language processing, recommendation systems, autonomous driving, and many more.

The three major problems in machine learning are clustering, classification, and (logistic) regression tasks (Goodfellow, Bengio, & Courville, 2016). Classification and regression tasks aim to predict a pre-defined label (dependent variable or set of classes) based on input data. Clustering aims to group similar objects into different clusters and thus to explore patterns in the data. In Anzer, Bauer, and Höner (2021) we present an introduction to supervised and unsupervised machine learning applications. Within the scope of this thesis, we address logistic regression tasks to predict the probability of a pass or shot success (Sections 3.1 and 3.2), clustering tasks in Sections 3.3 and 3.4, as well as a classification task in Section 3.4.

Tree-based algorithms present a substantial part of machine learning algorithms solving classification, as well as regression tasks. Following the basic idea of Friedman (2002), who introduced gradient boosting as an additive regression model, Chen and Guestrin (2016) presented a sophisticated optimization of the method, called extreme gradient boosting or short XGBoost gaining significant results in many classification or regression tasks. A major advantage of the XGBoost algorithms is, that they can handle imbalanced data. Whereas the basic idea of supervised machine learning methods (e.g. regressions) is to optimize a set of free parameters in an algorithm (e.g. XGBoost) in order to minimize the prediction error on the training data, another crucial design choice is the hyperparameter space. Hyperparameters describe a set of variables in an algorithm that are not optimized during the training process, but rather must be chosen before the training. With XGBoost (just as

with many other algorithms), this step can be performed automatically using Bayesian optimization (Bergstra, Bardenet, Bengio, & Kégl, 2011; Dewancker, McCourt, & Clark, 2016). By using a game theoretical approach to visually interpret trained XGBoost models, Thomson and Roth (1991), Rodríguez-Pérez and Bajorath (2020), and Lundberg and Lee (2017) addressed one of the biggest limitations of XGBoost, namely the difficult interpretability. As demonstrated in other domains (Antipov & Pokryshevskaya, 2020; Meng, Yang, Qian, & Zhang, 2020; Ibrahim, Mesinovic, Yang, & Eid, 2020) we show that SHAP values[18] can help to understand predictions of our expected goals model (Anzer & Bauer, 2021). In Anzer and Bauer (2021) as well as in Anzer and Bauer (2022) we compared XGBoost to various other algorithms (e.g. logistic regression, ADA boost, random forest or gradient boosting) and found that XGBoost yielded the best results.

A shortcoming of XGBoost models is that they require feature crafting to be effective. In contrast, artificial neural networks, another family of machine learning algorithms that can be used inter alia for classification tasks, are able to better handle less structured raw data. The basic concept of artificial neural networks was introduced in 1958 (Rosenblatt, 1958), however, only improvements by Werbos (1994) as well as advancements of the back-propagation algorithm in 2006 (Hinton, Osindero, & Teh, 2006) paired with available data and computing power enabled its recent success. Especially for image and video processing, the introduction of convolutional neural networks—a group of neural networks that are optimized to handle data structured as images using convolutional layers (Zhang et al., 2019)—helped

[18]SHAP is the abbreviation for SHapley Additive exPlanation. SHAP values originate from game theoretical concepts and describe the impact certain features have on machine learning predictions (Lundberg & Lee, 2017).

to outperform humans in various image classification tasks. By taking the typically 105×68 m seized pitch as the frame of an image and setting the trajectories of the player of each team as well as of the ball as shaded pixels in a different colours, convolutional neural networks have been used to perform classification tasks on spatio-temporal positional data in many sports (Mehrasa, Zhong, Tung, Bornn, & Mori, 2018). We use this approach applying convolutional neural networks to tracking data in Bauer, Anzer, and Shaw (2022).

For the clustering of goal origins (Anzer, Bauer, & Brefeld, 2021) and team formations (Bauer, Anzer, & Shaw, 2022), we rely on the most traditional method of agglomerative hierarchical clustering. It works bottoms up, in the sense that it starts with every observation as a single cluster and keeps merging two clusters until there is only one left. The structure describing at what stage two clusters were merged is generally referred to as a *dendrogram* (Murtagh & Contreras, 2012). There exist several ways to decide which two clusters to merge, ranging from single-linkage (Sibson, 1973) to max-linkage (Defays, 1977). While these describe the two extremes, we use more balanced approaches. In Anzer, Bauer, and Brefeld (2021) we opted for average-linkage (Sokal, 1958) and in Bauer, Anzer, and Shaw (2022) we used Ward's method (Ward & Joe, 1963). They all also require a distance metric to compute the similarity between observations. Again, there is a variety of different metrics to choose from, the most common one being the Euclidean distance. We selected ones better suited for our problems: the cosine distance (Qian, Sural, Gu, & Pramanik, 2004) and the Wasserstein distance (Olkin & Pukelsheim, 1982). Finally, while objective metrics like Silhouette values exist to decide on a number of clusters, this choice portrays an ill-posed problem (Rousseeuw, 1987). We use

these objective metrics in combination with a substantial amount of expert knowledge to both align on this number, but also to contextualize the results.

3 Empirical Studies

The main research question of this thesis is how one can use synchronized tracking data to quantify offensive performances on the three levels Figure 1 introduced. The goal of this work is neither to find one overarching statistic that summarizes all offensive performance to one number, nor to analyze every single aspect of offensive play. Instead, it aims to show how one can quantify some of the most important aspects of offensive performance on each level using the synchronized data (see Section 2.2). Studies I and II (Sections 3.1 and 3.2) analyze the offensive actions shots and passes in detail. These are then also used in study III (Section 3.3) to quantify offensive tactical patterns leading to goals using an unsupervised clustering technique. Lastly, study IV (Section 3.4) focuses on the offensive team strategy level, namely the question which build-up formation is most effective against various opposing formations.

Studies I, II, and III (Sections 3.1, 3.2, and 3.3) were published as a first author and the work done for study IV (Section 3.4) was done as a co-author. The following sections merely summarize, discuss and put the studies into the context of this thesis, while their full manuscripts can be found in the Appendix.

3.1 Study I: A Goal Scoring Probability Model for Shots based on Synchronized Positional and Event Data in Football (Soccer) (Anzer & Bauer 2021)

Goals are always the deciding factor of a football match, yet they occur very rarely—only about 1% of all possessions and about 10% of all shots end up as a goal (Pollard & Reep, 1997; Tenga et al., 2010; Lucey et al., 2014). Hence, measuring the offensive performance of a player or team purely based on goals

is subject to large variations and noise, especially when considering short time windows. Therefore, shots are often used as a proxy instead, even though the quality of shooting situations can vary considerably. The aim of this study (Anzer & Bauer, 2021) is to quantify the basic offensive action shot by computing the expected goals (xG) metric that estimates its quality. This is done by computing the probability of a shot being converted to a goal using a machine learning model. It follows the footsteps of other sports, like baseball or basketball, where more process-based metrics were established to measure on-base percentage (James, 1988) and shot locations (Chang et al., 2014) rather than home runs or points. While there already existed work on shot probabilities in football in the "grey literature" like master theses (Hedar, 2020; Rowlinson, 2020) and conference proceedings (Lucey et al., 2014), this study is the first to introduce a positional data-driven xG model in a peer-reviewed journal. Previously, only in a conference proceeding Lucey et al. (2014) used event and positional data from 10,000 shots from the English Premier League to develop an xG model.

For our approach we extracted nine hand-crafted features from the synchronized positional and event data of 105,627 shots and fed them into various supervised machine learning models. The best performing model consists of the XGBoost method (see Section 2.3). It achieves a ranked probability score (RPS) of 0.197, making it more accurate than any previously published expected goals model. This increased accuracy is largely due to important features only contained in the synchronized data (see Section 2.2), like the position of the goalkeeper, that would otherwise be missing.

The model enables various applications to analyze the underlying performance of teams and players. One can aggregate the

expected goal values on a team level per match to get an estimate of how many goals a team would have been expected to score, given their chance qualities during a match. We showed that these aggregated expected goal values pick up more information about the underlying performance than shots and are less susceptible to noise than goals, especially on short- to mid-term and thus are ideally suited to measure a team's current form. By nature, over longer periods xG's and goals should converge eventually, baring some systematic reasons for over/under performance (e.g. a really gifted striker). Figure 6 shows how one can measure a team's current performance level, by computing a rolling average of the aggregated match results over the last four matches (both offensively, and defensively, i.e. xG allowed to the opponents). It indicates that RB Leipzig had a weak spell

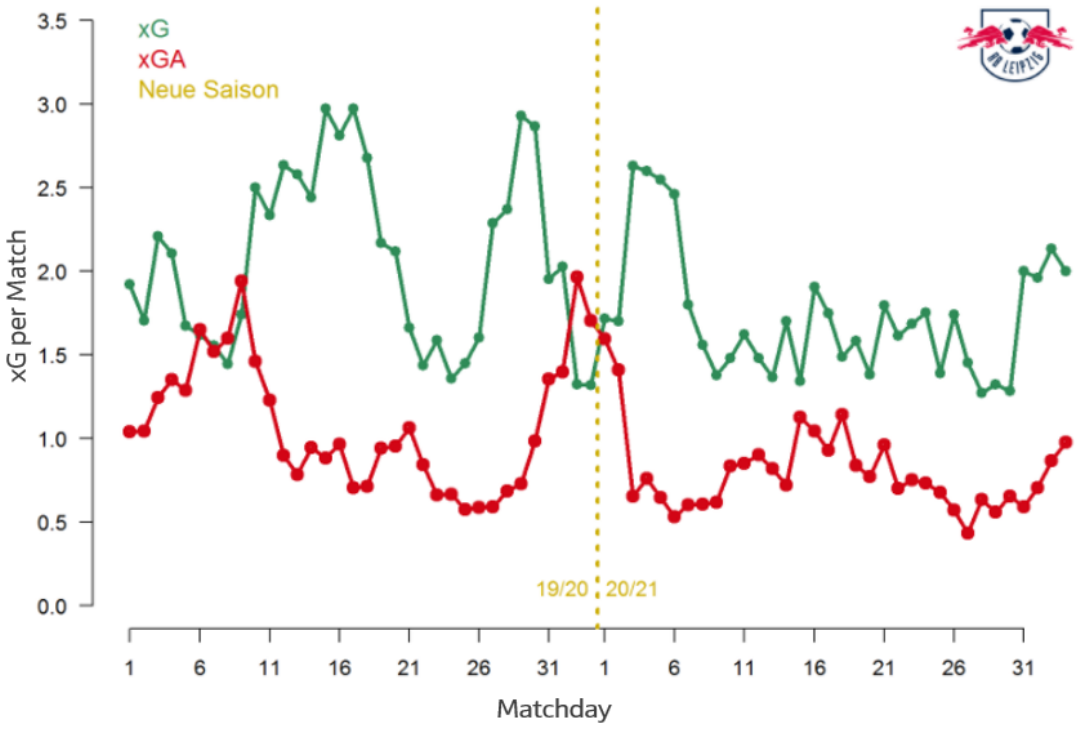

Figure 6: Rolling average of team aggregated xG values The figure shows for each matchday a rolling average of the team aggregated xG values (green xG created and red xG against) over the past four matches.

towards the end of their 2019/2020 campaign, where they on av-

erage allowed higher quality chances than they created. Another useful set of applications exists on a player level: with a player's aggregated xG values one can better measure how often a player finds himself in a promising scoring situation than would be possible by looking at his goals or shots tally. Further, over a large sample size, when comparing xG values to actual goals scored, one can get an indication of a player's finishing skill. It also allows us to address a shortcoming of a popular metric: Assists. The number of assists a player achieves depends to a high degree on their teammates' efficiency. By assigning the resulting xG value to the player that assisted the shot, we can measure chance creation capabilities more accurately and independent of the shooter's ability to score. This value is typically referred to as expected assists (xA).

A limitation of xG values in general is that they do not assign any value to dangerous situations where no shot was attempted. While these situations are rare (occurring only 0.93 times per match across our dataset), there exists ample research to extend this approach to all situations and not just shots (Link et al., 2016; Spearman, 2018; Fernández, Bornn, & Cervone, 2019; Decroos et al., 2019). Moreover, as present throughout this thesis, data quality and the results of the synchronization play a vital role in this model. If in actuality the goalkeeper was between the shooter and the goal, but in the synchronized data (for example due to erroneous tracking data), he is not, the xG model will substantially overestimate the chance of scoring. Even for purely event based xG models, the importance of accurate input data was highlighted in Robberechts and Davis (2020). Another limitation is that some potentially very relevant features are not recorded in current event and tracking datasets. For instance, future work could evaluate if adding the body orientation or the

level of ball control could improve the accuracy of the model.

3.2 Study II: Expected Passes: Determining the Difficulty of a Pass in Football (Soccer) Using Spatio-Temporal Data (Anzer & Bauer 2022)

Unlike goals or shots, the basic offensive action of a pass happens far more frequently during a match, but only few lead to shots and even fewer lead to goals directly (Goes, Kempe, Meerhoff, & Lemmink, 2019). Hence, there is a need to quantify them and a player's passing ability independent of the following action. Often a player's passing skill is measured by his pass completion rate (Król et al., 2017), i.e. the percentage of passes played, that successfully arrive at a team mate. This approach neglects that passes are of varying difficulty: a pass between two central defenders in open space is a lot simpler to complete than a chip pass played under high spatio-temporal pressure behind the last defending line.

Thus, study II (Anzer & Bauer, 2022) presents a model to quantify this basic action by measuring the pass difficulty. Similar as described in study I (Section 3.1), this is done using a supervised machine learning approach that in this case calculates the probability of any given pass being completed and is aptly named expected pass (xPass) (Spearman et al., 2017; Power et al., 2017; Fernández et al., 2020; Arbues-Sanguesa et al., 2020; Alguacil et al., 2020; Stöckl et al., 2021). Spearman et al. (2017) were the first to use tracking data from 10,875 passes to predict the probability of a pass being completed using ball and player trajectories. While it is possible to construct such a model purely on event data (Łukasz Szczepański & Mchale, 2016), we showed that synchronized positional data (see Section 2.2) is es-

sential to achieve a high accuracy. But even within this data, one quintessential attribute is missing: who the intended receiver was is not recorded in the event data for unsuccessful passes. Since this is crucially important information to any supervised machine learning model classifying the pass results, a major part of this study shows how one can determine the targeted player for unsuccessful passes. For that purpose, we use a state-of-the-art movement model (Brefeld et al., 2019; Fernandez & Bornn, 2018) to derive the potential positions of all players within a certain time window and combine this with physics-based ballistic ball trajectory model (Spearman et al., 2017). From this combination we can determine the teammate most likely to reach the ball first.

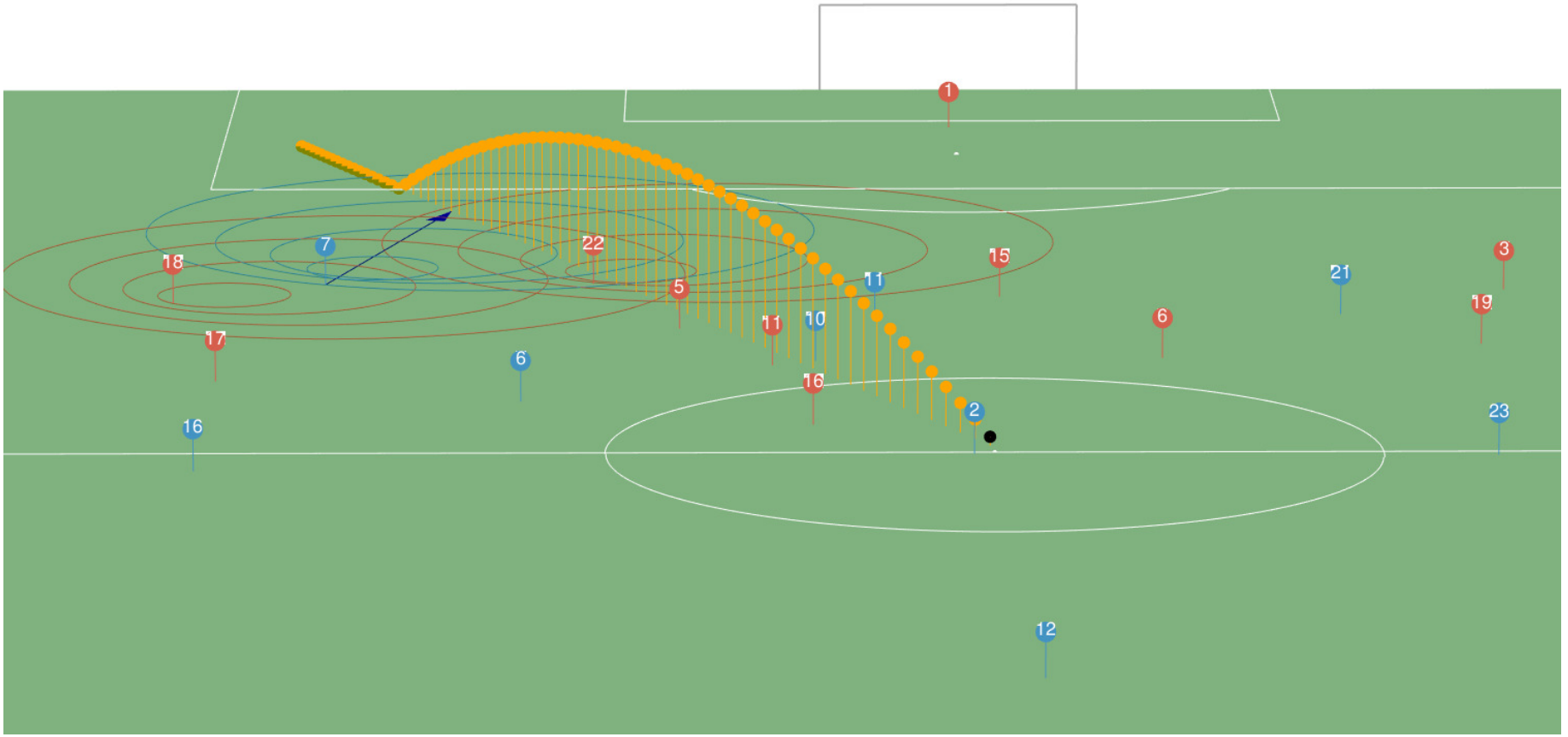

Figure 7: Estimated target of a pass with ball-trajectory and movement models. This figure is copied from Anzer and Bauer (2022) (p.6), where more details about the pass (including a video of the situation) are given.

Figure 7 depicts a pass where the estimated ball trajectory is shown in orange and the movement areas of players in vicinity of the pass destination are shown in circles. The accuracy of the intended receiver estimation of 93% surpasses the best previously published work of 81% (Spearman et al., 2017). The intended target estimation has the added benefit, that it yields some useful features (e.g. possible angle or exit velocity) for the machine learning models estimating the pass success probability. These physics based features are combined with a wide set of hand-crafted ones that were developed in close collaboration with practitioners. Various subsets of these features are then fed into an XGBoost algorithm (see Section 2.3). The model on the full feature set achieves an area under the curve (AUC) of 93.4%, and significantly outperforms traditional models using either only event data or tracking data without information about the intended receiver. Since we can only estimate the target, if the ball trajectory is available (at least for the first 10 frames), the success probabilities can only be calculated for non-blocked passes. To remove this bias from the results, we additionally trained a model to compute the likelihood that a pass is blocked. This novel blocking model is used to discount the xPass values. Apart from the high accuracy, a strength of this study are the three exhaustive manual validation studies of its sub-elements. We evaluated:

(1) the synchronization algorithm for passes,

(2) the accuracy of the intended target identification,

(3) and the final xPass values.

With this metric, we can quantify a very important offensive skill a player's passing ability, by comparing his actual comple-

tion rate to his expected rate. The more he outperforms his expectation, the better his passing performance. Furthermore, by computing xPass values for alternative hypothetical passes, we can also assess a player's risk profile.

But this application also highlights a limitation of our approach: we can only evaluate the risk of pass but not its reward. In the future, this risk model could be combined with a reward model (Steiner et al., 2019; Goes et al., 2019; Fernandez & Bornn, 2018) to also give a full picture of the value of the action and to evaluate a player's decision-making skill. As in study I, data quality has a strong impact on the resulting probability estimations. In this case the target estimations (only using 10 frames) is susceptible to occasional errors in the ball data. Moreover, as stated above with this approach we cannot reasonably estimate the success probability of blocked passes, since we are unable to identify the target. This could be addressed in the future by using a different method for identifying intended targets for blocked passes (e.g. simply the closest team mate).

3.3 Study III: The Origins of Goals in the German Bundesliga (Anzer, Bauer, & Brefeld 2021)

Similar to study I, study III (Anzer, Bauer, & Brefeld, 2021) investigates goal scoring. But instead of focusing on the basic final action (i.e. the shot), this study analyzes the tactical patterns leading to the goal. Although every goal in football is sui generis due to the complexity of the game, most of them do not occur randomly, but originate from certain underlying patterns. Due to their rareness, finding these goal patterns is a difficult task that has motivated researchers for decades (Reep & Benjamin, 1968; Szwarc, 2007; Mitrotasios & Armatas, 2012; Plummer, 2013;

González-Ródenas et al., 2019). For instance, González-Ródenas et al. (2019) manually classified 380 goals into two categories and found that 75.9% goals originate from open-play and only 24.1% from set-pieces. Study III explores how we can identify distinct offensive tactical patterns leading up to goals, by using an unsupervised approach. For that purpose we use our synchronized positional data (see Section 2.2) enriched with several features developed in studies I and II, that either describe the finishing action (e.g. xG values), or the actions leading up to the goal (e.g. passes played during the possession phase) and feed it into an agglomerative clustering approach. The next steps of choosing the number of clusters and contextualizing them based on video footage were done in close collaboration with football practitioners. The final clustering consists of 50 interpretable clusters, each describing a unique offensive pattern leading to a goal and, due to the hierarchical nature, one can see related patterns by looking at the dendrogram depicted in Figure 8.

This clustering also allows for cutoffs at higher levels, where it finds more general categories (e.g. set-piece goals). In the end, football experts evaluated each cluster and were able to identify and name clear patterns the contained goals had in common. Figure 9 shows exemplary goals for the 12 open-play goal clusters they identified. These categories include goals following a long build-up phase finished with the foot or with the head (i.e. typically set up by a cross), as well as goals following counter-pressing (motivating more defensive oriented follow-up research in Bauer and Anzer (2021)).

What separates this study from others exploring goal-scoring patterns (Reep & Benjamin, 1968; Szwarc, 2007; Mitrotasios & Armatas, 2012; Plummer, 2013; González-Ródenas et al., 2019), is its large dataset (3,417 goals), the handcrafted features de-

scribing the underlying offensive actions in greater detail, and the direct involvement of football practitioners throughout the process. Currently, most match analysis departments routinely analyze and categorize the goals they scored/conceded during a season. This manual process can be automated using our clustering methodology, with the added benefit that it not only saves time, but also creates reproducible categories. Furthermore, the large sample size allows for the identification of rare categories (e.g. corner-trick plays). Comparing individual players with averages over all players on the same position reveals characteristic traits that may be exploited when scouting replacements for departing players.

A limitation of this study, as is typical for most unsupervised machine learning approaches, is that the resulting clustering may not be identical to categories a practitioner may demand. While this exploratory approach delivers satisfying categories, some practitioners may prefer predefined categories. For such cases, a supervised approach can be built upon our results as future work. A further common problem, as touched upon in Section 2.3, is that determining the right cluster number is somewhat subjective.

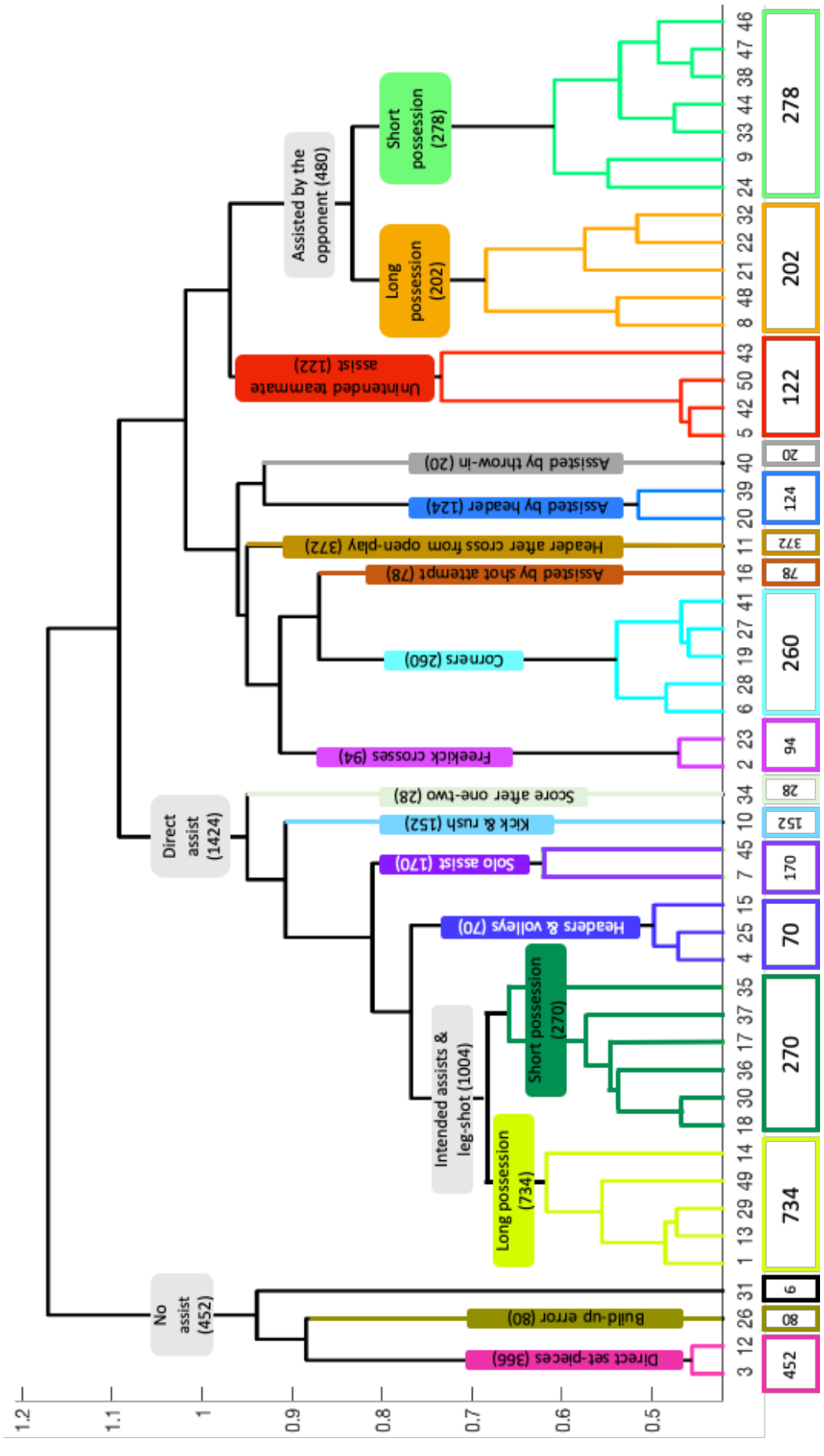

Figure 8: Dendrogram with contextual annotations copied from Anzer, Bauer, and Brefeld (2021) (p.4). The colored names represent the clusters experts grouped together and the numbers at the bottom indicate how many goals are contained within each cluster.

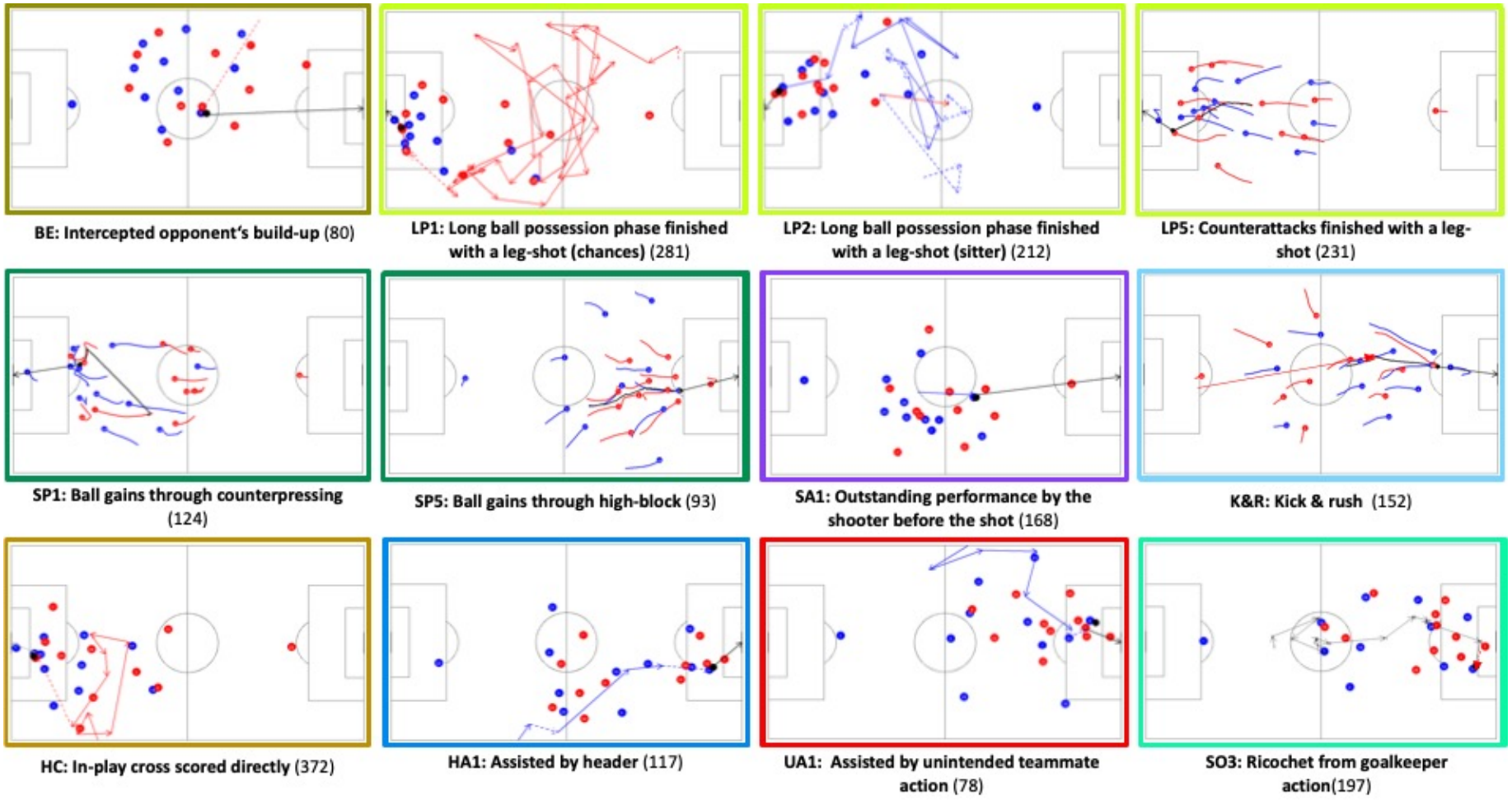

Figure 9: Clusters of open-play goals each with an exemplary case as can be found in Anzer, Bauer, and Brefeld (2021) (p.5) including a more detailed description.

3.4 Study IV: Putting Team Formations in Association Football into Context (Bauer, Anzer, & Shaw 2022)

After quantifying individual offensive actions in studies I and II (Sections 3.1 and 3.2), and identifying offensive scoring patterns in study III (Section 3.3), one of the goals of study IV (Bauer, Anzer, & Shaw, 2022) is to analyze offensive performance on a team strategy level (see Figure 1). This done by measuring the effectiveness of offensive formations against different opposing defensive set-ups on a team level. Which formation to choose, is one of the coach's most important strategic decisions (Wilson, 2009; Wei, Sha, Lucey, Morgan, & Sridharan, 2013; Bialkowski et al., 2014; Müller-Budack et al., 2019), as it affects a team's ability to create scoring opportunities as well as the opponent's. However, a team does not play in the exact same formation throughout the match, but its shape rather depends on the tactical situation (Andrienko et al., 2019; Shaw & Glickman, 2019). Shaw and Glickman (2019) introduced a method to classify team formations depending on the game-state (offensive and defensive), but concluded that further granularity of these game-states would lead to more accurate representations of formations. Therefore, to estimate a team's formation during a certain tactical phase, one must first determine each time when said phase occurred during the match. Typically, a football match can be separated in four distinct match-phases, as illustrated in Figure 10. These four phases can additionally be split into subcategories, for instance the offensive phase can be further differentiated in build-up play and attacking play. The exact definitions of the phases and their subcategories can be found in Bauer (2021). To complicate things, unlike American football or basketball, football cannot easily be separated into distinct possession phases. Therefore, one needs

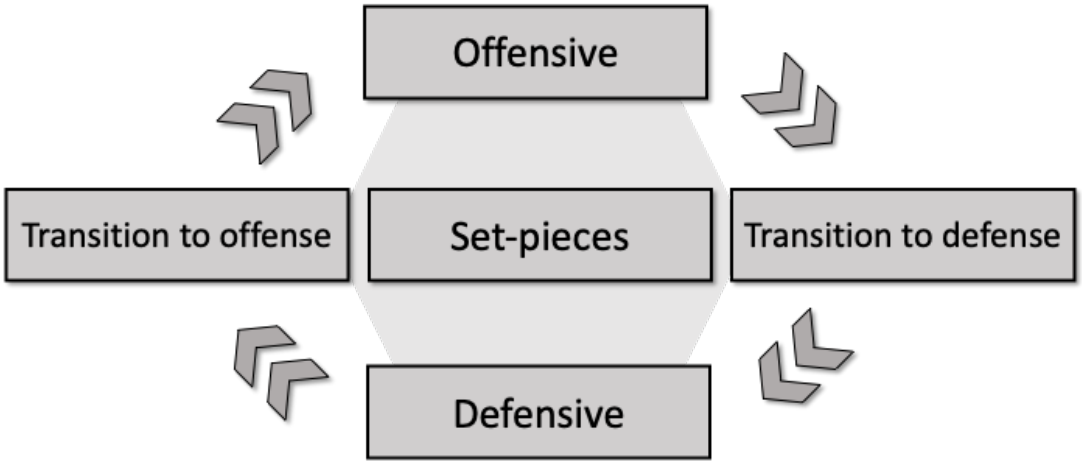

Figure 10: Four main tactical phases, in which open-play is generally divided. This figure and a more thorough explanation of the tactical phases can be found in Bauer (2021).

to not only identify the phases, but also their starting and end points.

The first part of this study aims to determine the active tactical phase for each moment of the game. In order to have a ground truth, 97 Bundesliga matches from the 2018/2019 season were annotated describing for every moment of the game the current tactical main- and sub-phases (in sum 59 hours and 50 minutes). These labels are used to train convolutional neural networks (CNN), taking positional data mapped to 2-D images as input (as in Dick and Brefeld (2019), Zheng, Yue, and Lucey (2016) and K.-C. Wang and Zemel (2016)). With these (predicted) phases now available for all Bundesliga matches from the 2013/2014 to the 2019/2020 season, we can determine what formations teams used within these phases (as in Shaw and Glickman (2019)) and use a hierarchical clustering to determine the ones most frequently used. Again, the choice of the right number of clusters is non-trivial and addressed in close collaboration with football experts evaluating and contextualizing the resulting clusters.

The above steps enable us to answer the strategic question of

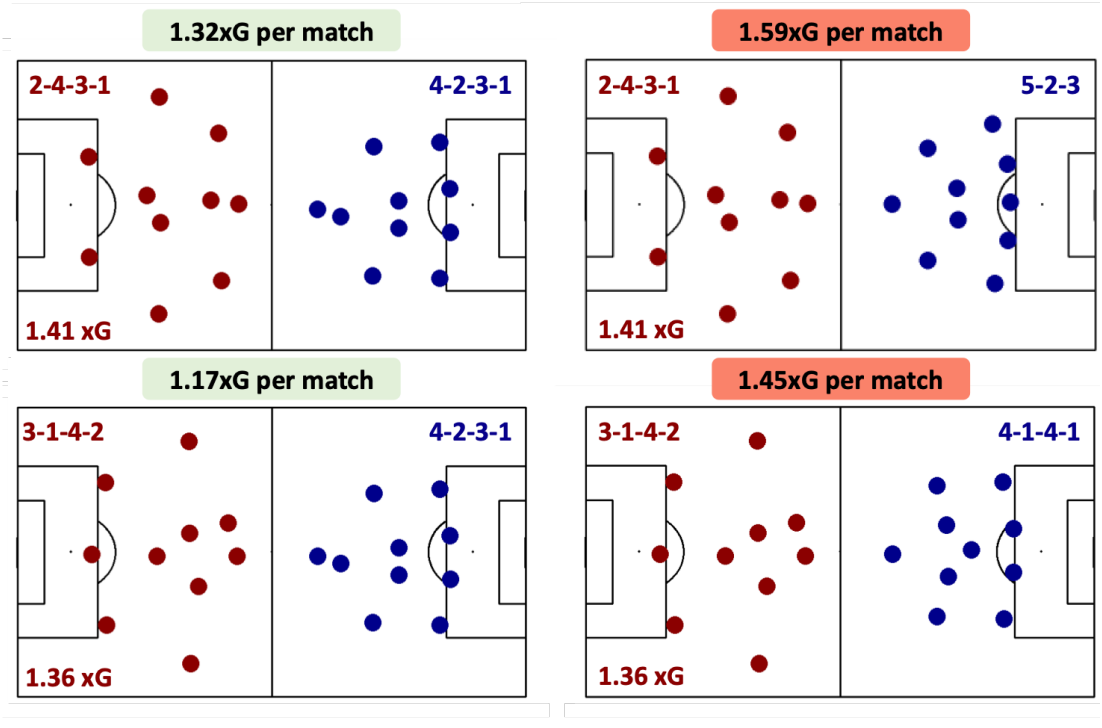

Figure 11: Effectiveness, measured by expected goals, of different formations as introduced in Bauer, Anzer, and Shaw (2022) (p.11).

which formation is the most efficient against an opposing one during a certain phase. Figure 11 shows how well two different build-up formations perform against various mid-block formations employed by the opposing team. The performance is measured by the accumulated xG values (because they, as shown in study I (Section 3.1), can pick up underlying information about offensive performance more quickly than goals) when both formations were used and scaled to full match lengths. As can be seen on the left, against a 4-2-3-1 defensive mid-block formation a 2-4-3-1 build-up is more effective than a 3-1-4-2 (the two most frequently used build-up formations).

Obviously, this rather simple approach of measuring effectiveness of offensive formations comes with several limitations, most notably we did not control for confounding factors like if there is a preference for stronger (or weaker) teams to use a particular formation. Moreover, we kept the definitions of the tactical

phases used for labeling rather general. However, in reality each team may have their own style, so in future work one could explore if team specific models could improve the tactical phase identification. Lastly, as in study III, selecting the right number of clusters of different formations, is non-trivial and greatly affects the results.

4 Discussion

The aim of this thesis is to show how one can objectively analyze offensive performances using synchronized positional and event data. In a first step (see Section 2.2), it introduced a novel methodology to combine the two separate datasets. This reproducible methodology can be applied to any tracking and event data—vendor- and sport-agnostic—and enables researchers to combine the information of both datasets instead of choosing just one. Throughout this thesis, offensive performances were investigated from a very granular action-based level all the way to a higher team based strategy level (see Figure 1). Studies I and II (Sections 3.1 and 3.2) objectively quantified two of the most common offensive actions, passes, and shots. Study III (Section 3.3) also showed how one can measure offensive performance on a tactical pattern level. And finally, study IV (Section 3.4) explored the effectiveness of different offensive formations on a team strategy level. Especially the inclusion of an improved quantification of basic actions (like expected goals), proved to be a vital component in the analysis of tactical patterns and team strategy: in study III, it played an important role in separating goals based on their chance quality and in study IV, it allowed to measure the efficiency of a formation more reliably than would have been possible with basic events like shots or goals. Table 1 gives an overview of the studies included in this thesis. It describes the level on which each study operates (i.e. action, pattern, or strategy), it lists the amount of data used (and from which league)[19] and what machine learning methods were utilized. Moreover, it names some of the key results or applications, as well as some of the limitations of each study.

[19] BL denotes the German Bundesliga and BL2 the German Bundesliga 2.

Table 1: Overview of the studies presented throughout this thesis.

Study	Level	Data	Methods	Results/Applications	Limitations
Study I (Anzer & Bauer, 2021)	Offensive Action: Shot	4,284 Matches (BL & BL2); 105,627 Shots	Supervised ML (XGBoost)	• Shot quality model • Measures finishing skill • Measures assist quality	• Slight shifts in synchronization have large impacts on the results • Important features (e.g. body orientation) not included in the dataset • Non-shot dangerous situation not considered
Study II (Anzer & Bauer, 2022)	Offensive Action: Pass	918 Matches (BL); 840,386 Passes	Supervised ML (XGBoost)	• Pass success probability estimation • Identification of intended pass targets • Uses movement and ball trajectory models • Measures passing skill	• Physical trajectories highly dependent on the accuracy of ball data • Pass decisions not analyzed, due to missing pass reward model • Only available for non blocked passes
Study III (Anzer, Bauer, & Brefeld, 2021)	Tactical Pattern: Goal Creation	1,224 Matches (BL & BL2); 3,457 Goals	Unsupervised ML (HC)	• Categorization of Bundesliga goals • Identification of different goal scoring tactics • Contextualization of scoring trends	• Categories may differ from practitioner needs • Selection of the ideal cluster number partially subjective
Study IV (Bauer, Anzer, & Shaw, 2022)	Offensive Strategy: Formation Selection	2,142 Matches (BL); 97 Fully Labeled Matches	Supervised ML (CNN) Unsupervised ML (HC)	• Detection of tactical phases • Clustering of different formations • Measures offensive effectiveness of varying formation strategies	• Tactical phases may deviate between teams • Selection of the ideal cluster number partially subjective • Formational effectiveness does not account for confounding factors

4.1 Data and Synchronization

The quality of any data-driven analysis depends for the largest part on the underlying data. While both data sources used in this thesis are recorded in detail and oblige strict quality control processes, they are still prone to errors. Human errors can occur in the manual process of collecting event data (e.g. identifying the wrong player as the one who took the shot), as well as in the human supported part of the tracking data collection (e.g. annotating when the play is halted by the referee). Furthermore, even the automated parts of tracking data collection (i.e. the computer vision algorithms) are not infallible for a variety of reasons (e.g. players getting mixed up in huddle situations). An interesting yet untouched area of research could investigate how to automatically identify erroneous data. A general limitation of the current tracking data is, that it simplifies players' movement to two-dimensional positions missing two very important aspects: player orientation and tracking their limbs. Both areas have been addressed in research (Arbués-Sangüesa, Haro, Ballester, & Martín, 2019; Arbues-Sanguesa et al., 2020; Cust, Sweeting, Ball, & Robertson, 2018) and have the potential to unlock completely new avenues of more granular analysis, once this additional data is collected systematically. However, the question remains, if the marginal gain of adding more complexity to the data will be worth the effort given, that the current data sources have not been fully explored yet. Furthermore, the current tracking data technologies require extensive on-premise camera set-ups making it prohibitively expensive for lower leagues with limited financial power (Manafifard et al., 2017). In an attempt to lower the entry barrier and to democratize this data, research (Johnson,

2021) and companies[20] have explored possibilities to extract positional data (at a lower quality) purely from broadcast videos.

Besides the input data, the ground truth labeled data is also essential for any supervised machine learning model. Depending how well-defined the target variable is (often measured by the inter-labeler reliability), the better a model can perform. Obviously, more objective labels like whether a shot ended as a goal, or whether a pass reached a teammate, achieve a higher inter-labeler reliability than more subjective definitions, as what tactical phase is currently active (see Section 3.4) or which pass was more difficult (see Section 3.2). Nevertheless, it should be a goal during the ground truth gathering process to maximize this measure by developing clear and concise definitions, training the labelers, and monitoring their pair-wise accuracy. It is also important not to introduce any biases into the process (e.g. when estimating the difficulty of a pass, humans can be influenced by the result, i.e. whether it was successful).

The most susceptible step to data quality issues is the event synchronization. Being at the intersection of both datasets, if either contains errors, the synchronization may fail. For instance, if in the event data the wrong player was recorded as having taken a shot, the algorithm will most likely fail looking for a situation where this player could have taken a shot in the tracking data. Similarly, if within the tracking data the ball position is stuck at the corner flag for a certain period of time, the algorithm will fail to synchronize any event that happened in the meantime. As shown in Anzer and Bauer (2022), with the high-quality data in the Bundesliga these issues preventing a successful synchronization happen fairly rarely, but the lower the data quality of either

[20] e.g. Skillcorner (https://www.skillcorner.com), or Metrica Sports (https://metrica-sports.com/#)

source, the more often they will occur. Since our rule-based approach to solve the issue of synchronization performed with such a high accuracy, we refrained from using a machine learning based one (see Section 2.3). Nevertheless, it could be interesting future research, if after collecting frame accurate timestamps of the events in the tracking data, one could train a machine learning model to perform this task at a even higher quality level. There has also been ample research on completely automating the manual process of event data collection by either using video footage as input (Ekin, Tekalp, & Mehrotra, 2003; Kolekar, Palaniappan, Sengupta, & Seetharaman, 2009; Pouyanfar & Chen, 2017) or tracking data (Stein et al., 2019; Richly, Bothe, Rohloff, & Schwarz, 2016; Motoi et al., 2012; Gudmundsson & Wolle, 2010). If these automated approaches would reach a satisfying accuracy, they could eventually replace the cost/time intensive process of manual event data collection with the added benefit of removing subjectivity from the process. Throughout this thesis, we have shown the importance of using synchronized data for complex tasks of measuring offensive performances. But the synchronization alone already provides significantly more context to individual actions (see Figure 4) than would be reasonably possible to collect manually. For instance, we can easily obtain a popular currently hand collected metric "Packing",[21] i.e. how many opposing players were bypassed with a single action (e.g. dribble or pass) (Steiner et al., 2019), or we can derive who the closest opponent was for every event, exactly how long a player was in possession of the ball before each action, or several more. In a similar vein, the synchronized data could also be used not only to extend context around recorded events, but rather to create completely new off-ball events. For example, in

[21] https://www.impect.com/en/

Bauer and Anzer (2021), we identify situations when a team is counterpressing and in Fassmeyer et al. (2021), we detect counterattacking situations. This enhanced data could also include several other events, e.g. offensive off-ball runs (Gregory, 2019; Fernandez & Bornn, 2018).

4.2 Machine Learning

Machine learning on football data comes with its own challenges. A general one is that for supervised tasks one typically needs large amounts of labelled data. In Sections 3.1 and 3.2, this was not an issue, since for all the considered shots and passes we already have access to the label whether they were successful or not, but in other studies we conducted (Bauer, Anzer, & Smith, 2022; Fassmeyer et al., 2021; Bauer, Anzer, & Shaw, 2022; Bauer & Anzer, 2021) gathering enough ground truth data was an exhaustive prerequisite to apply supervised machine learning algorithms. We addressed this problem in several ways during the research program: in Bauer, Anzer, and Smith (2022) we used data augmentation to increase our training dataset by a factor of ten, through slightly altering the input data while keeping the labels unchanged. In Fassmeyer et al. (2021), we show that using a semi-supervised approach, where we first learn a meaningful feature representation using variational autoencoders, we can then greatly reduce the amount of annotated data needed for a classifier working in this feature space. Further methods to reduce the required amount of labeled data, such as transfer or active learning (Panigrahi, Nanda, & Swarnkar, 2021; Druck, Settles, & McCallum, 2009), could also be applied to football data.

A problem very specific to the tracking data is determining an appropriate ordering of players to train models in a permutation-

invariant space (Wei et al., 2013). One cannot simply use the raw positional data as input to a machine learning algorithm, because the order in which players appear is not always identical and simple rule-based attempts to order them (e.g. from left to right, or based on jersey number) are neither stable nor scalable. To address this obstacle, in Sections 3.1 and 3.2 we used hand-crafted features and in Section 3.4 we transformed the data to images. Recently graph neural networks were shown to be a very effective approach to solve this problem (Stöckl et al., 2021; Dick, Tavakol, & Brefeld, 2021; Sun, Karlsson, Wu, Tenenbaum, & Murphy, 2019).

On a higher level, a limitation of machine learning is that it uses events from the past, to predict unseen (often future) events. This can cause complications if the underlying process one is interested in, does not remain constant, but changes over time. In football this is clearly the case, as strategies, tactics, and physical capabilities change over time (Wallace & Norton, 2014). Not only that, but as we have shown in Link and Anzer (2021) and Szymski et al. (2021), also an external force as the COVID-19 pandemic and the absence of crowds has changed the nature of the game. As coach turnover is quite common in football, the new coach's requirements and definitions of actions, tactical patterns, and strategies may change. Hence, it increases the importance of having a streamlined and flexible approach in place to deal with those changing needs. Moreover, football players are acting individuals, so a purely external data-based view neglects the complexities determining human decision making processes. Therefore, future work could take sport psychological principals in consideration.

4.3 Interplay between Sport Science and Data Science

Throughout this research program, we placed a high importance on the close collaboration between football and machine learning experts. The benefit of this interplay is shown in this thesis, where rule-based approaches were often outperformed by machine learning methods that use hand-crafted features (as suggested by the domain experts). This finding is confirmed by Goes, Brink, Elferink-Gemser, Kempe, and Lemmink (2020), Herold et al. (2019) and Rein and Memmert (2016) who argue, that this interdisciplinary cooperation is essential for success in the sports analytics domain. The collaboration helped with the unsupervised tasks by contextualizing results, but also with the supervised ones by developing definitions, annotating manual labels, and creating relevant features. Involving the domain experts in the process has also enabled them to gain a basic understanding of the underlying data and what is possible to achieve with it. This transparency created an atmosphere that encouraged back and forth communication and ultimately led to an increase in the usage of data-driven recommendations. The domain of machine learning can also profit from this cooperation: Tuyls et al. (2021) concluded that football data provides an unparalleled data environment for developing and testing new machine learning methods.

4.4 Future Work

The major contribution of this thesis to the research domain is to show how the combination of tracking and event data can help analysis of offensive performances in football. But this thesis is far from a complete and exhaustive exploration of all offensive related actions, tactics, or strategies, as they range from offensive

corner kick analysis (Shaw & Gopaladesikan, 2021) to pitch control metrics (Fernandez & Bornn, 2018; Spearman, 2018). Most previous studies in this area are either based on positional or event data, but hardly any have combined the two. Consequently, instead of covering all offensive related football problems, we used the synchronization and selected some areas of high interest, like shots and passes. Nevertheless, the general concept introduced in Sections 3.1 and 3.2 could be extended to other basic offensive events like dribbles or throw-ins in future work. One of the overarching questions in this space is, if there exists a metric that can precisely measure the value or threat created with each action or movement. While there have been several attempts of measuring it with event data (Decroos et al., 2019; Rudd, 2011) or with tracking data (Fernández et al., 2019; Link et al., 2016) using the synchronization proposed here, could improve these models even further. Such a derived model could even include our xG and xPass approaches as sub models.

Just like in other sports, the development of advanced analytics in football primarily started with analyzing offensive performances. One of the reasons that there is less research on the defensive side is that initially hardly any defensive actions were recorded and to this day in the event data offensive actions typically outnumber defensive ones by a magnitude of ten. Obviously, this does not mean that teams spend less time defending than attacking, but rather that defensive actions are harder to record. Often a sign of good defending are the lack of actions, as the former Italian international Paolo Maldini stated: "If I have to tackle then I have already made a mistake".[22] This quote shows how much harder of a challenge it is to measure defensive perfor-

[22] https://www.thesun.co.uk/sport/football/1197148/paolo-maldini-the-defender-so-good-he-didnt-even-need-to-make-a-tackle/

mances in football. Nevertheless, during this research program we have conducted several studies that also analyze defensive performances: In Bauer and Anzer (2021) we detect the defensive tactical pattern of counterpressing and in Bauer, Anzer, and Smith (2022) we measure defensive marking schemes during opposing corner kicks. Study IV (Section 3.4) can also be used to identify defensive tactical phases (like the height of the block) or to measure the effectiveness of defensive strategies. Moreover, with the synchronization described in Section 2.2 one could evaluate opposing defensive positioning at key moments of the game.

5 Conclusion

In conclusion, this dissertation uses the largest collection of event and tracking in the world to introduce a novel and reproducible approach of combining the two data sources. It showed, how this data can unlock new possibilities to better quantify the two most common offensive actions (passes and shots). These improved actions also allowed to measure and identify offensive tactical patterns and team strategies. Due to the close cooperation with practitioners, every study included details on how the results can be applied in sporting organizations to improve processes, as well as in the media to enrich data story telling. There, the two enhanced offensive metrics were developed into statistics that are now shown to a global audience during all German Bundesliga matches, introducing sports analytics to the mainstream fan. Due to the sheer complexity of football this dissertation is not meant to quantify every aspect of offensive performance. Therefore, future work could build on the concepts introduced here to answer one of the still open questions of how to accurately value offensive actions in football. Furthermore, as defensive performance was mostly left untouched by this dissertation, one could analogously to this work focus on the defensive side of football.

References

Alguacil, F. P., Fernandez, J., Arce, P. P., & Sumpter, D. (2020). Seeing in to the future: using self-propelled particle models to aid player decision-making in soccer. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1-23.

Anderson, C., & Sally, D. (2013). *The numbers game: Why everything you know about football is wrong*. Penguin UK.

Andrienko, G., Andrienko, N., Anzer, G., Bauer, P., Budziak, G., Fuchs, G., ... Wrobel, S. (2019). Constructing spaces and times for tactical analysis in football. *IEEE Transactions on Visualization and Computer Graphics*, *27*, 2280-2297. Retrieved from https://ieeexplore.ieee.org/document/8894420 doi: 10.1109/TVCG.2019.2952129

Andrienko, G., Andrienko, N., Budziak, G., Dykes, J., Fuchs, G., von Landesberger, T., & Weber, H. (2017). Visual analysis of pressure in football. *Data Mining and Knowledge Discovery*, *31*, 1793-1839. doi: 10.1007/s10618-017-0513-2

Andrzejewski, M., Chmura, J., Pluta, B., & Konarski, J. M. (2015). Sprinting activities and distance covered by top level europa league soccer players:. *http://dx.doi.org/10.1260/1747-9541.10.1.39*, *10*, 39-50. Retrieved from https://journals.sagepub.com/doi/abs/10.1260/1747-9541.10.1.39 doi: 10.1260/1747-9541.10.1.39

Antipov, E. A., & Pokryshevskaya, E. B. (2020). Interpretable machine learning for demand modeling with high-dimensional data using gradient boosting machines and shapley values. *Journal of Revenue and Pricing Management*, *19*, 355-364. Retrieved from https://doi.org/10.1057/s41272-020-00236-4 doi: 10.1057/s41272-020-00236-4

Anzer, G., & Bauer, P. (2021). A goal scoring probability model based on synchronized positional and event data. *Frontiers in Sports and Active Learning (Special Issue: Using Artificial Intelligence to Enhance Sport Performance)*, *3*, 1-18. Retrieved from https://www.frontiersin.org/articles/10.3389/fspor.2021.624475/full doi: 10.3389/fspor.2021.624475

Anzer, G., & Bauer, P. (2022). Expected passes—determining the difficulty of a pass in football (soccer) using spatio-temporal data. *Data Mining and Knowledge Discovery, Springer US*. doi: 10.1007/s10618-021-00810-3

Anzer, G., Bauer, P., & Brefeld, U. (2021). The origins of goals in the german bundesliga. *Journal of Sport Science*. Retrieved from https://www.tandfonline.com/doi/full/10.1080/02640414.2021.1943981 doi:

10.1080/02640414.2021.1943981

Anzer, G., Bauer, P., & Höner, O. (2021). The Identification of Counterpressing in Football. In D. Memmert (Ed.), *Match analysis—how to use data in professional sport* (1st Editio ed., pp. 228–235). New York: Routledge. doi: https://doi.org/10.4324/9781003160953

Arbues-Sanguesa, A., Martin, A., Fernandez, J., Ballester, C., & Haro, G. (2020). Using player's body-orientation to model pass feasibility in soccer. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2020-June*, 3875-3884. doi: 10.1109/CVPRW50498.2020.00451

Arbués-Sangüesa, A., Haro, G., Ballester, C., & Martín, A. (2019). Head, shoulders, hip and ball... hip and ball! using pose data to leverage football player orientation. *Barça sports analytics summit*, 1-13.

Bauer, P. (2021). Automated detection of complex tactical patterns in football using positional and event data—using machine learning techniques to identify tactical behavior. *(Unpublished doctoral dissertation)*.

Bauer, P., & Anzer, G. (2021). Data-driven detection of counterpressing in professional football—a supervised machine learning task based on synchronized positional and event data with expert-based feature extraction. *Data Mining and Knowledge Discovery*. Retrieved from https://doi.org/10.1007/s10618-021-00763-7 doi: 10.1007/s10618-021-00763-7

Bauer, P., Anzer, G., & Shaw, L. (2022). Putting Team Formations in Association Football into Context. *Journal of Sports Analytics (submitted)*.

Bauer, P., Anzer, G., & Smith, J. W. (2022). Individual role classification for players defending corners in football (soccer). *Journal of Quantitative Analysis in Sports (submitted)*.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, 1-9.

Bialkowski, A., Lucey, P., Carr, P., Matthews, I., Sridharan, S., & Fookes, C. (2016). Discovering team structures in soccer from spatiotemporal data. *IEEE Transactions on Knowledge and Data Engineering*, *28*, 2596-2605. doi: 10.1109/TKDE.2016.2581158

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., & Matthews, I. (2014). Large-scale analysis of soccer matches using spatiotemporal tracking data. *IEEE International Conference on Data Mining, ICDM (Proceeding)*, 725-730. doi: 10.1109/ICDM.2014.133

Borrie, A., Jonsson, G. K., & Magnusson, M. S. (2002). Temporal pattern analysis and its applicability in sport: An explanation and exemplar data. *Journal of Sports Sciences*, *20*, 845-852. doi: 10.1080/026404102320675675

Bradley, P. S., Lago-Peñas, C., Rey, E., & Diaz, A. G. (2013). The effect of high and low percentage ball possession on physical and technical profiles in english fa premier league soccer matches. *Journal of Sports Sciences*, *31*, 1261-1270. doi: 10.1080/02640414.2013.786185

Brander, J. A., Egan, E. J., & Yeung, L. (2014). Estimating the effects of age on nhl player performance. *Journal of Quantitative Analysis in Sports*, *10*, 241-259. Retrieved from https://www.degruyter.com/document/doi/10.1515/jqas-2013-0085/html doi: 10.1515/JQAS-2013-0085

Bransen, L., & Haaren, J. V. (2019). Measuring football players' on-the-ball contributions from passes during games. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11330 LNAI*, 3-15. doi: 10.1007/978-3-030-17274-91

Bransen, L., & van Haaren, J. (2020). Player chemistry: Striving for a perfectly balanced soccer team. *arXiv*, 1-24.

Brefeld, U., Lasek, J., & Mair, S. (2019). Probabilistic movement models and zones of control. *Machine Learning*, *108*, 127-147. Retrieved from https://doi.org/10.1007/s10994-018-5725-1 doi: 10.1007/s10994-018-5725-1

Brooks, J., Kerr, M., & Guttag, J. (2016). Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining*, *9*, 338-349. doi: 10.1002/sam.11318

Buchheit, M., Allen, A., Poon, T. K., Modonutti, M., Gregson, W., & Salvo, V. D. (2014). Integrating different tracking systems in football: multiple camera semi-automatic system, local position measurement and gps technologies. *Journal of Sports Sciences*, *32*, 1844-1857. Retrieved from http://dx.doi.org/10.1080/02640414.2014.942687 doi: 10.1080/02640414.2014.942687

Carling, C. (2011). Influence of opposition team formation on physical and skill-related performance in a professional soccer team. *European Journal of Sport Science*, *11*, 155-164. doi: 10.1080/17461391.2010.499972

Chang, Y.-H., Maheswaran, R., Su, J., Kwok, S., Levy, T., Wexler, A., & Squire, K. (2014). Quantifying shot quality in the nba. *Proceedings of the 8th annual MIT Sloan sports analytics conference, Boston (USA)*, 1-8.

Chawla, S., Estephan, J., Gudmundsson, J., & Horton, M. (2017). Classifi-

cation of passes in football matches using spatiotemporal data. *ACM Transactions on Spatial Algorithms and Systems*, *3*. doi: 10.1145/3105576

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17*, 785-794. doi: 10.1145/2939672.2939785

Cust, E. E., Sweeting, A. J., Ball, K., & Robertson, S. (2018). Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. *https://doi.org/10.1080/02640414.2018.1521769*, *37*, 568-600. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/02640414.2018.1521769 doi: 10.1080/02640414.2018.1521769

Decroos, T., Bransen, L., van Haaren, J., & Davis, J. (2020). Vaep: An objective approach to valuing on-the-ball actions in soccer (extended abstract). *IJCAI International Joint Conference on Artificial Intelligence*, *2021-Janua*, 4696-4700. doi: 10.24963/ijcai.2020/648

Decroos, T., Haaren, J. V., Bransen, L., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1851-1861. doi: 10.1145/3292500.3330758

Decroos, T., Haaren, J. V., & Davis, J. (2018). Automatic discovery of tactics in spatio-temporal soccer match data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 223-232. doi: 10.1145/3219819.3219832

Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal*, *20*, 364-366. Retrieved from https://academic.oup.com/comjnl/article/20/4/364/393966 doi: 10.1093/COMJNL/20.4.364

Dewancker, I., McCourt, M., & Clark, S. (2016). Bayesian optimization for machine learning : A practical guidebook. Retrieved from https://arxiv.org/abs/1612.04858v1

Dick, U., & Brefeld, U. (2019). Learning to rate player positioning in soccer. *Big Data*, *7*, 71-82. doi: 10.1089/big.2018.0054

Dick, U., Tavakol, M., & Brefeld, U. (2021). Rating player actions in soccer. *Frontiers in Sports and Active Learning (Special Issue: Using Artificial Intelligence to Enhance Sport Performance)*, *3*, 1-14. doi: 10.3389/fspor.2021.682986

Druck, G., Settles, B., & McCallum, A. (2009). Active learning by labeling features. *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical*

Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009, 81-90. doi: 10.3115/1699510.1699522

Ekin, A., Tekalp, A. M., & Mehrotra, R. (2003). Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, *12*, 796-807. doi: 10.1109/TIP.2003.812758

Fassmeyer, D., Anzer, G., Bauer, P., & Brefeld, U. (2021). Toward Automatically Labeling Situations in Soccer. *Frontiers in Sports and Active Living*, *3*(November). doi: 10.3389/fspor.2021.725431

Fernandez, J., & Bornn, L. (2018). Wide open spaces: A statistical technique for measuring space creation in professional soccer. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1-19.

Fernández, J., Bornn, L., & Cervone, D. (2019). Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1-18. Retrieved from https://lukebornn.com/sloan_epv_curve.mp4

Fernández, J., Bornn, L., & Cervone, D. (2020). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Arxiv Preprint*. Retrieved from https://arxiv.org/abs/2011.09426

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*, 367-378. doi: 10.1016/S0167-9473(01)00065-2

Goes, F. R., Brink, M. S., Elferink-Gemser, M., Kempe, M., & Lemmink, K. A. (2020). The tactics of successful attacks in professional association football—large-scale spatiotemporal alanalysis of dynamic subgroups using position tracking data. *Journal of Sports Sciences*, *39*, 523-532. doi: 10.1080/02640414.2020.1834689

Goes, F. R., Kempe, M., Meerhoff, L. A., & Lemmink, K. A. (2019). Not every pass can be an assist: A data-driven model to measure pass effectiveness in professional soccer matches. *Big Data*, *7*, 57-70. doi: 10.1089/big.2018.0067

Goes, F. R., Meerhoff, L. A., Bueno, M. J., Rodrigues, D. M., Moura, F. A., Brink, M. S., ... Lemmink, K. A. (2020). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, *0*, 1-16. doi: 10.1080/17461391.2020.1747552

González-Ródenas, J., López-Bondia, I., Aranda-Malavés, R., Desantes, A. T., Sanz-Ramírez, E., & Malaves, R. A. (2019). Technical, tactical and spatial

indicators related to goal scoring in european elite soccer. *Journal of Human Sport and Exercise*, *15*, 1-16. doi: 10.14198/jhse.2020.151.17

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Retrieved from http://www.deeplearningbook.org/front_matter.pdf

Gould, P., & Gatrell, A. (1979). A structural analysis of a game: The liverpool v manchester united cup final of 1977. *Social Networks*, *2*, 253-273. doi: 10.1016/0378-8733(79)90017-0

Gregory, S. (2019). Ready player run: Off-ball run identification and classification sam gregory. *Barça sports analytics summit*, 1-19.

Gréhaigne, J. F., Godbout, P., & Bouthier, D. (1999). The foundations of tactics and strategy in team sports. *Journal of Teaching in Physical Education*, *18*, 159-174. doi: 10.1123/jtpe.18.2.159

Gudmundsson, J., & Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys*, *50*. doi: 10.1145/3054132

Gudmundsson, J., & Wolle, T. (2010). Towards automated football analysis: Algorithms and data structures. *Proc. 10th Australas. Conf. Math. Comput. Sport*.

Hedar, S. (2020). Applying machine learning methods to predict the outcome of shots in football outcome of shots in football. *Thesis Uppsala University*.

Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., & Meyer, T. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science and Coaching*, *14*. doi: 10.1177/1747954119879350

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527-1554. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.76.1541 doi: 10.1162/neco.2006.18.7.1527

Ibrahim, L., Mesinovic, M., Yang, K. W., & Eid, M. A. (2020). *Explainable prediction of acute myocardial infarction using machine learning and shapley values.* doi: 10.1109/ACCESS.2020.3040166

James, B. (1988). *The bill james historical baseball abstract*. Random House Incorporated.

Johnson, N. (2021). Extracting player tracking data from video using non-stationary cameras and a combination of computer vision techniques. *MIT Sloan Sports Analytics Conference, Boston (USA)*. Retrieved from http://www9.cs.tum.edu/projects/aspogamo

Kolekar, M. H., Palaniappan, K., Sengupta, S., & Seetharaman, G. (2009).

Semantic concept mining based on hierarchical event detection for soccer video indexing. *Journal of Multimedia*, *4*, 298-312. doi: 10.4304/jmm.4.5.298-312

Król, M., Konefał, M., Chmura, P., Andrzejewski, M., Zając, T., & Chmura, J. (2017). Pass completion rate and match outcome at the world cup in brazil in 2014. *Polish Journal of Sport and Tourism*, *24*, 30-34. doi: 10.1515/pjst-2017-0004

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. Norton.

Link, D. (2018a). *Data analytics in professional soccer*. Springer.

Link, D. (2018b). Sports analytics: How (commercial) sports data create new opportunities for sports science. *German Journal of Exercise and Sport Research*, *48*, 13-25. doi: 10.1007/s12662-017-0487-7

Link, D., & Anzer, G. (2021). How the covid-19 pandemic has changed the game of soccer. *International Journal of Sports Medicine*. doi: 10.1055/a-1518-7778

Link, D., Lang, S., & Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. *PloS one*, *11*(12), e0168768.

Link, D., & Weber, H. (2017). Effect of ambient temperature on pacing in soccer depends on skill level. *Journal of Strength and Conditioning Research*, *31*, 1766-1770. Retrieved from https://journals.lww.com/nsca-jscr/Fulltext/2017/07000/Effect_of_Ambient_Temperature_on_Pacing_in_Soccer.2.aspx doi: 10.1519/JSC.0000000000001013

Linke, D., Link, D., & Lames, M. (2020). Football-specific validity of tracab's optical video tracking systems. *PLoS ONE*, *15*, 1-17. doi: 10.1371/journal.pone.0230179

Lucey, P., Bialkowski, A., Monfort, M., Carr, P., & Matthews, I. (2014). "quality vs quantity": Improved shot prediction in soccer using strategic features from spatiotemporal data. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1-9. Retrieved from http://www.sloansportsconference.com/?p=15790

Lucey, P., Oliver, D., Carr, P., Roth, J., & Matthews, I. (2013). Assessing team strategy using spatiotemporal data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Part F1288*, 1366-1374. doi: 10.1145/2487575.2488191

Lundberg, S. M., & Lee, S. I. (2017). Consistent feature attribution for tree ensembles. *arXiv*.

Manafifard, M., Ebadi, H., & Moghaddam, H. A. (2017). A survey on player tracking in soccer videos. *Computer Vision and Image Understanding*, *159*, 19-46. doi: 10.1016/j.cviu.2017.02.002

Martens, F., Dick, U., & Brefeld, U. (2021). Space and control in soccer. *Frontiers in Sports and Active Learning (Special Issue: Using Artificial Intelligence to Enhance Sport Performance)*, *3*, 1-13. doi: 10.3389/fspor.2021.676179

McHale, I. G., & Relton, S. D. (2018). Identifying key players in soccer teams using network analysis and pass difficulty. *European Journal of Operational Research*, *268*, 339-347. doi: 10.1016/j.ejor.2018.01.018

Mehrasa, N., Zhong, Y., Tung, F., Bornn, L., & Mori, G. (2018). Learning person trajectory representations for team activity analysis. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1-8. Retrieved from http://arxiv.org/abs/1706.00893

Meng, Y., Yang, N., Qian, Z., & Zhang, G. (2020). What makes an online review more helpful: An interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*, 466-490. doi: 10.3390/jtaer16030029

Mitrotasios, M., & Armatas, V. (2012). Analysis of goal scoring patterns in the 2012 european football championship. *The Sport Journal*, 1-9. Retrieved from http://thesportjournal.org/article/analysis-of-goal-scoring-patterns-in-the-2012-european-football-championship/

Motoi, S., Misu, T., Nakada, Y., Yazaki, T., Kobayashi, G., Matsumoto, T., & Yagi, N. (2012). Bayesian event detection for sport games with hidden markov model. *Pattern Analysis and Applications*, *15*, 59-72. doi: 10.1007/s10044-011-0238-6

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*, 86-97. Retrieved from https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.53 doi: 10.1002/WIDM.53

Müller-Budack, E., Theiner, J., Rein, R., & Ewerth, R. (2019). "does 4-4-2 exist?" – an analytics approach to understand and classify football team formations in single match situations. *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports (Nice, France)*, *MMSports '*, 25-33. doi: 10.1145/3347318.3355527

Oliver, D. (2004). *Basketball on paper: Rules and tools for performance analysis*. Brassey's.

Olkin, I., & Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra and Its Applications*, *48*,

257-263. doi: 10.1016/0024-3795(82)90112-4

Panigrahi, S., Nanda, A., & Swarnkar, T. (2021). A survey on transfer learning. *Smart Innovation, Systems and Technologies*, *194*, 781-789. doi: 10.1007/978-981-15-5971-6

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, *6*, 1-15. Retrieved from http://dx.doi.org/10.1038/s41597-019-0247-7 doi: 10.1038/s41597-019-0247-7

Pettersen, S. A., Johansen, D., Johansen, H., Berg-Johansen, V., Gaddam, V. R., Mortensen, A., ... Halvorsen, P. (2014). Soccer video and player position dataset. *Proceedings of the 5th ACM Multimedia Systems Conference, MMSys 2014 (Singapore, March 2014)*, 18-23. doi: 10.1145/2557642.2563677

Plummer, B. T. (2013). Analysis of attacking possessions leading to a goal attempt, and goal scoring patterns within men's elite soccer. *Journal of Sports Science*, *1*, 1-038. Retrieved from https://learnzone.loucoll.ac.uk/sportres/BScDissertations/BPlummer_dle21_data_temp_turnitintool_1509080854._1910_1363351553_718.pdf

Pollard, R., & Reep, C. (1997). Measuring the effectiveness of playing strategies at soccer. *Journal of the Royal Statistical Society Series D: The Statistician*, *46*, 541-550. doi: 10.1111/1467-9884.00108

Pouyanfar, S., & Chen, S.-C. (2017). Semantic event detection using ensemble deep learning. , 203-208. doi: 10.1109/ism.2016.0048

Power, P., Ruiz, H., Wei, X., & Lucey, P. (2017). "not all passes are created equal:" objectively measuring the risk and reward of passes in soccer from tracking data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Part F1296*, 1605-1613. doi: 10.1145/3097983.3098051

Qian, G., Sural, S., Gu, Y., & Pramanik, S. (2004). Similarity between euclidean and cosine angle distance for nearest neighbor queries. *Proceedings of the ACM Symposium on Applied Computing*, *2*, 1232-1237. doi: 10.1145/967900.968151

Reep, C., & Benjamin, B. (1968). Skill and chance in association football author. *Journal of the Royal Statistical Society*, *131*, 581-585. Retrieved from https://www.jstor.org/stable/2343726?seq=1

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, *5*.

doi: 10.1186/s40064-016-3108-2

Richly, K., Bothe, M., Rohloff, T., & Schwarz, C. (2016). Recognizing compound events in spatio-temporal football data. *IoTBD 2016 - Proceedings of the International Conference on Internet of Things and Big Data*, 27-35. doi: 10.5220/0005877600270035

Robberechts, P., & Davis, J. (2020). How data availability affects the ability to learn good xg models. *Communications in Computer and Information Science*, *1324*, 17-27. Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-64912-8_2 doi: 10.1007/978-3-030-64912-8_2

Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, *34*, 1013-1026. Retrieved from https://doi.org/10.1007/s10822-020-00314-0 doi: 10.1007/s10822-020-00314-0

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386-408. doi: 10.1037/h0042519

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. doi: 10.1016/0377-0427(87)90125-7

Rowlinson, A. (2020). Football shot quality visualizing the quality of soccer/football shots. Retrieved from www.aalto.fi

Roy, M. V., Robberechts, P., chi Yang, W., Raedt, L. D., & Davis, J. (2018). Leaving goals on the pitch : Evaluating decision making in soccer.

Rudd, S. (2011). A framework for tactical analysis and individual offensive production assessment in soccer using markov chains. *New EnglandSymposium on Statistics in Sports*. Retrieved from http://nessis.org/nessis11/rudd.pdf

Samuel, A. L. (1959). Some studies in machine learning. *IBM Journal of Research and Development*, *3*, 210-229. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392560

Sarmento, H., Anguera, T., Campaniço, J., & Leitão, J. (2010). Development and validation of a notational system to study the offensive process in football. *Medicina*, *46*, 401-407. doi: 10.3390/medicina46060056

Seidl, T. (2019). Radio-based position tracking in sports—validation, pattern recognition and performance analysis. *Dissertation thesis, TU München*.

Shaw, L., & Glickman, M. (2019). Dynamic analysis of team strategy in professional football. *Barça sports analytics summit*, 1-13.

Shaw, L., & Gopaladesikan, S. (2021). Routine inspection: A playbook for corner kicks. *MIT Sloan Sports Analytics Conference, Boston (USA)*.

Sibson, R. (1973). Slink: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, *16*, 30-34. Retrieved from https://academic.oup.com/comjnl/article/16/1/30/434805 doi: 10.1093/COMJNL/16.1.30

Sokal, C. D. M. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, *38*, 1409-1438. Retrieved from https://ci.nii.ac.jp/naid/10004143217 doi: 10.24517/00055549

Spearman, W. (2018). Beyond expected goals. *12th Annual MIT Sloan Sports Analytics Conference, Boston (USA)*, 1-17. Retrieved from https://www.researchgate.net/publication/327139841

Spearman, W., Basye, A., Dick, G., Hotovy, R., & Pop, P. (2017). Physics-based modeling of pass probabilities in soccer. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1-14. Retrieved from https://www.researchgate.net/profile/William-Spearman/publication/315166647_Physics-Based_Modeling_of_Pass_Probabilities_in_Soccer/links/58cbfca2aca272335513b33c/Physics-Based-Modeling-of-Pass-Probabilities-in-Soccer.pdf

Stein, M., Janetzko, H., Seebacher, D., Jäger, A., Nagel, M., Hölsch, J., ... Grossniklaus, M. (2017). How to make sense of team sport data: From acquisition to data modeling and research aspects. *Data*, *2*, 2. doi: 10.3390/data2010002

Stein, M., Seebacher, D., Karge, T., Polk, T., Grossniklaus, M., & Keim, D. A. (2019). From movement to events: Improving soccer match annotations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11295 LNCS*, 130-142. doi: 10.1007/978-3-030-05710-7_11

Steiner, S., Rauh, S., Rumo, M., Sonderegger, K., & Seiler, R. (2019). Outplaying opponents—a differential perspective on passes using position data. *German Journal of Exercise and Sport Research*, *49*, 140-149. doi: 10.1007/s12662-019-00579-0

Stöckl, M., Seidl, T., Marley, D., & Power, P. (2021). Making offensive play predictable - using a graph convolutional network to understand defensive performance in soccer. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1-19.

Sun, C., Karlsson, P., Wu, J., Tenenbaum, J. B., & Murphy, K. (2019). Stochastic prediction of multi-agent interactions from partial observations. *Seventh*

International Conference on Learning Representations (ICLR), New Orleans (USA), 1–15.

Szwarc, A. (2007). Efficacy of successful and unsuccessful soccer teams taking part in finals of champions league. *Research Yearbook*, *13*, 221-225. Retrieved from http://journals.indexcopernicus.com/abstracted.php?icid=838944

Szymski, D., Anzer, G., Alt, V., Meyer, T., Gärtner, B., & Krutsch, W. (2021). Contact times in professional football before and during the sars-cov-2 pandemic: Tracking data from the german bundesliga. *European Journal of Sport Science (Submitted)*.

Tenga, A., Holme, I., Ronglan, L. T., & Bahr, R. (2010). Effect of playing tactics on goal scoring in norwegian professional soccer. *Journal of Sports Sciences*, *28*, 237-244. doi: 10.1080/02640410903502774

Thomson, W., & Roth, A. E. (1991). *The shapley value: Essays in honor of lloyd s. shapley.* (Vol. 58). doi: 10.2307/2554979

Tuyls, K., Omidshafiei, S., Muller, P., Wang, Z., Connor, J., Hennes, D., ... Hassabis, D. (2021). Game plan: What ai can do for football, and what football can do for ai. *Journal of Artificial Intelligence Research*, *71*, 41-88. doi: 10.1613/JAIR.1.12505

Vračar, P., Štrumbelj, E., & Kononenko, I. (2016). Modeling basketball play-by-play data. *Expert Systems with Applications*, *44*, 58-66. doi: 10.1016/j.eswa.2015.09.004

Wallace, J. L., & Norton, K. I. (2014). Evolution of world cup soccer final games 1966-2010: Game structure, speed and play patterns. *Journal of Science and Medicine in Sport*, *17*, 223-228. doi: 10.1016/j.jsams.2013.03.016

Wang, K.-C., & Zemel, R. (2016). Classifying nba offensive plays using neural networks. *MIT Sloan Sports Analytics Conference, Boston (USA)*.

Wang, Q., Zhu, H., Hu, W., Shen, Z., & Yao, Y. (2015). Discerning tactical patterns for professional soccer teams: An enhanced topic model with applications. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2015-Augus*, 2197-2206. doi: 10.1145/2783258.2788577

Ward, T., & Joe, H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236-244.

Wei, X., Sha, L., Lucey, P., Morgan, S., & Sridharan, S. (2013). Large-scale analysis of formations in soccer. *2013 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2013*. doi: 10.1109/DICTA.2013.6691503

Werbos, P. (1994). *The roots of backpropagation: from ordered derivatives to neural networks.* John Wiley and Sons Inc.

Wilson, J. (2009). *Inverting the pyramid, a history of football tactics*. Orion.

Zhang, Q., Zhang, M., Chen, T., Sun, Z., Ma, Y., & Yu, B. (2019). Recent advances in convolutional neural network acceleration. *Neurocomputing*, *323*, 37-51. doi: 10.1016/j.neucom.2018.09.038

Zheng, S., Yue, Y., & Lucey, P. (2016). Generating long-term trajectories using deep hierarchical networks. *Advances in Neural Information Processing Systems*, 1551-1559.

Łukasz Szczepański, & Mchale, I. (2016). Beyond completion rate: Evaluating the passing ability of footballers. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *179*, 513-533. doi: 10.1111/rssa.12115

A Appendix—Study I: A Goal Scoring Probability Model for Shots based on Synchronized Positional and Event Data in Football (Soccer)

frontiers in Sports and Active Living

ORIGINAL RESEARCH
published: 29 March 2021
doi: 10.3389/fspor.2021.624475

A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer)

Gabriel Anzer[1,2*†] *and Pascal Bauer*[2,3†]

[1] *Sportec Solutions AG, Subsidiary of the Deutsche Fußball Liga (DFL), Munich, Germany,* [2] *Institute of Sports Science, University of Tübingen, Tübingen, Germany,* [3] *DFB-Akademie, Deutscher Fußball-Bund e.V., Frankfurt am Main, Germany*

Due to the low scoring nature of football (soccer), shots are often used as a proxy to evaluate team and player performances. However, not all shots are created equally and their quality differs significantly depending on the situation. The aim of this study is to objectively quantify the quality of any given shot by introducing a so-called *expected goals (xG)* model. This model is validated statistically and with professional match analysts. The best performing model uses an extreme gradient boosting algorithm and is based on hand-crafted features from synchronized positional and event data of 105, 627 shots in the German Bundesliga. With a ranked probability score (RPS) of 0.197, it is more accurate than any previously published expected goals model. This approach allows us to assess team and player performances far more accurately than is possible with traditional metrics by focusing on process rather than results.

Keywords: expected goals, XG, positional data, event data, applied machine learning, football, soccer, sports analytics

OPEN ACCESS

Edited by:
Arno Knobbe,
Leiden University, Netherlands

Reviewed by:
José Luis Felipe,
European University of Madrid, Spain
Laurentius Antonius Meerhoff,
Leiden University, Netherlands

****Correspondence:***
Gabriel Anzer
gabrielanzer@gmail.com

†ORCID:
Gabriel Anzer
orcid.org/0000-0003-3129-8359
Pascal Bauer
orcid.org/0000-0001-8613-6635

Specialty section:
This article was submitted to Sports Science, Technology and Engineering, a section of the journal Frontiers in Sports and Active Living

Received: *31 October 2020*
Accepted: *15 February 2021*
Published: *29 March 2021*

Citation:
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 3:624475. doi: 10.3389/fspor.2021.624475

1. INTRODUCTION

In professional football (soccer), only 1% of all attacking plays and only around 10% of all shots taken end up in a goal (Pollard and Reep, 1997; Tenga et al., 2010; Lucey et al., 2014). However, goals alone decide the outcome of a game and are the most common metric to judge both a team's and individual player's performance. For example, both the best goal scorers[1] and the players with the most assists[2] receive a lot of attention from experts and the media. Nevertheless, judging performances solely based on this binary metric (*goal* or *no goal*) loses a lot of information and places results over process. For example, the performance from an outstanding creative player could be made void by strikers missing all their chances.

For this reason, in football as well as in other sports, it has become typical to consider more granular process-based metrics. In baseball, scouts and experts focused their attention on *homeruns* or *hits* for decades until more complex evaluation metrics changed the assessment procedure of hitters' performance significantly (James, 1985). Another famous example is basketball: By

[1]https://www.goal.com/en-us/lists/cristiano-ronaldo-lionel-messi-pele-who-are-the-top-goal-scorers-/ynctx2o9fa371vi1x0dsgr0np (accessed July 10, 2020).
[2]https://www.givemesport.com/1534019-the-top-10-players-with-the-most-assists-in-europes-top-five-leagues-this-decade (accessed July 8, 2020).

calculating scoring probabilities of different shot locations (Reich et al., 2006; Chang et al., 2014; Harmon et al., 2016; Jagacinski et al., 2019), the NBA's shooting behavior changed significantly[3]. The high scoring nature of basketball enables clubs to go even further and to apply individual shooting efficiency models (Beshai, 2014). Similar shot prediction models were also developed for ice hockey (Macdonald, 2012) as well as for return plays in tennis (Wei et al., 2016) and table tennis (Draschkowitz et al., 2015).

The fact that football is the lowest scoring game of the above-mentioned sports, makes it harder to develop such models, because of the scarcity of data. Consequently, the rareness and therefore importance of goals makes such a metric even more relevant when assessing teams and players. As another consequence of this low-scoring nature, the role of shots as a success proxy within several studies in football is fortified (Spearman et al., 2017). However, assessing shots just by being successful or not is a too rough abstraction that warps reality. An *expected goals model* (hereafter *xG model*) tries to estimate the probability of any given shot being converted to a goal based on various different factors describing the shot. These probabilities can then be added up per team and yield a "result-agnostic" description of the teams' performance. The xG metric is well-established in the football analytics community (see Davis and Robberechts, 2020)[4,5,6,7]. Although to the best of our knowledge, no peer-reviewed journal publication has introduced a positional data-driven xG model, valuable work has been done in "gray literature" like master theses (Hedar, 2020; Rowlinson, 2020) and conference proceedings (Lucey et al., 2014). Rathke (2017) analyzed in total around 18,000 shots from one season of Bundesliga and Premier League based on manually acquired shot annotations. Differentiating between four different shooting types *(open play footed shot, header, freekick, or penalty shot)*, Ruiz et al. (2017) built a multi-layer perceptron to predict shot outcomes based on roughly 10,000 shots. Using a similar approach, Fairchild et al. (2018) tried to predict the goal scoring probabilities of 1,115 non-penalty shots from 99 Major League Soccer matches, again solely based on event data.

Recent developments in technology allows us not only to make use of manually annotated event data *(shots, passes, goals with a manually assigned location)* but also accurate positions of all 22 players and the ball at up to 25 times a second. It is quite intuitive that the positioning of the defensive team, especially of the goalkeeper, has a crucial influence on the shot outcome (Lucey et al., 2014; Schulze et al., 2018). **Figure 1** displays the positioning of relevant players during two shots occurring at similar spots. In the left figure, both a defender and the goalkeeper are in good position to block the shot, while in the right figure the attacker has already dribbled past the goalkeeper (#38) and defenders, and faces an easy tap-in into an empty goal[8]. However, this information is not covered in event data and thus not taken into consideration in the previously listed xG models. Lucey et al. (2014) were the first to estimate goal probabilities using event and positional data in their model. They used 10,000 shots of the English Premier League.

In this paper, we will introduce a shot prediction model, utilizing event and positional data. The accuracy of this model is evaluated both statistically and based on the discussion with professional match analysts. We also incorporate their expertise both when defining the model's features and when interpreting their influence on the prediction. Additionally, we show how our model can support coaching staffs by introducing various use cases and applying them on one season worth of Bundesliga data.

The remainder of this paper is structured as follows. In section 2, we introduce the data and definitions. How event and tracking data are synchronized is described in section 3. Section 4 describes how the supervised prediction model is build, and finally, section 5 consists of two parts: practical applications (5.1) of our approach based on a season of German Bundesliga and a critical discussion of the results (5.2).

2. DATA AND DEFINITIONS

Like in most other professional football competitions, the German Bundesliga systematically collects positional and event data on a league-wide level in a pre-defined and thus consistent format. *Positional data*—often also referred to as tracking or movement data (Stein et al., 2015)—provides the positions of all players, referees, and the ball related to the pitch boundaries with a frequency of 25 Hz. These data are gathered by an optical tracking system, which captures high-resolution video footage from different camera perspectives. On the other hand, *event data* are manually acquired by trained operators live during the match. Among other things, this event data contain many details about basic events, such as passes, shots, fouls, saves, and so on including the involved players or special characteristics.

Since shots are an important statistic in football, the event data in the Bundesliga describe them with more than 20 attributes. For example, the collector differentiates between three basic shot types *(leg, header, other)* or six different scenarios how a player controlled the ball before taking a shot *(direct, volley, two touches, dribbling* $> 10\,m$*, dribbling* $< 10\,m$*, set-piece)*.

In this investigation, we make use of 105,627 shots from German Bundesliga and 2nd Bundesliga of the seasons 2013/2014 until 2019/2020. The event data were collected according to the official Bundesliga match-data catalog[9], and the optical tracking data were provided by Chyronhego's TRACAB system[10].

[3]https://fivethirtyeight.com/features/how-mapping-shots-in-the-nba-changed-it-forever/ (accessed July 10, 2020).

[4]https://www.americansocceranalysis.com/home/2017/3/6/validating-the-asa-xgoals-model (accessed October 24, 2020).

[5]http://www.northyardanalytics.com/blog/2015/08/22/pitfalls-of-measuring-shooting-and-saving-skill/ (accessed October 24, 2020).

[6]https://www.optasports.com/services/analytics/advanced-metrics/ (accessed October 24, 2020).

[7]https://differentgame.wordpress.com/2014/05/19/a-shooting-model-an-expglanation-and-application/ (accessed October 24, 2020).

[8]The situation in the right plot is also displayed in **Figure 2**. The respective video can be found here: https://www.youtube.com/watch?v=UdvrKfsJISY&feature=onebox&t=1m08s (accessed October 24, 2020).

[9]https://s.bundesliga.com/assets/doc/10000/2189_original.pdf (accessed September 10, 2020).

[10]https://chyronhego.com/wp-content/uploads/2019/01/TRACAB-PI-sheet.pdf (accessed September 10, 2020).

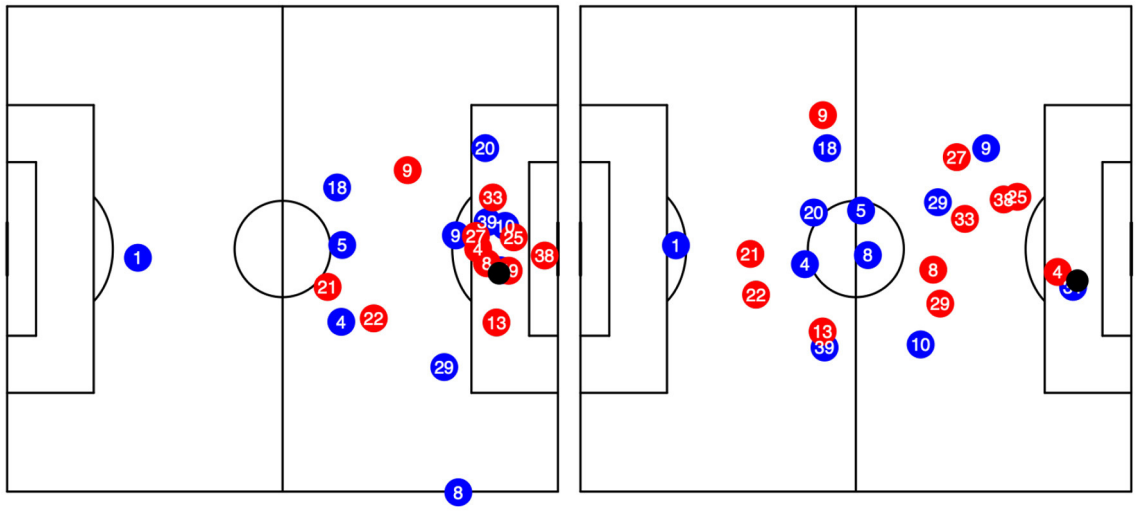

FIGURE 1 | Player positions of two shots from roughly the same location, but different surrounding environments. In both cases, the blue team is playing from left to right.

Due to a growing availability of optical tracking systems in football, several studies have been conducted to evaluate their accuracy (Redwood-Brown et al., 2012; Linke et al., 2018, 2020; Linke, 2019; Taberner et al., 2019). In Linke et al. (2020), the two versions of the TRACAB system (*Gen 4*/*Gen 5*)[11] were compared to an accurate ground truth measurement[12]. Both systems achieved a diversion of $<$ 10 cm from the ground truth system (RMSE Gen 4: 0.09 cm, Gen 5: 0.08 cm). A non-peer reviewed study confirmed these results[13]. All above-mentioned evaluation studies focused on player detection, whereas the detection of the ball—probably the hardest challenge for optical tracking systems—is not covered.

To the best of our knowledge, no scientific study evaluated the quality of event data. However, in the German Bundesliga the acquisition follows an elaborate quality assurance process. Critical information is double-checked manually live (e.g., goals and red cards). Finally, an independent person inspects and adds additional information (e.g., event locations) to all acquired event data after the match.

3. MAKING USE OF BOTH POSITIONAL AND EVENT DATA

3.1. Synchronizing Shots With Tracking Data

A major challenge when attempting to use both tracking and event data is that they are generally not aligned. This is due to the fact that they come from different data providers and/or acquisition methods, one specialized in logging events manually according to catalog of set definitions (i.e., what is considered a shot or a tackling) and the other focusing on extracting player positions through, for example, computer vision algorithms. This leads to two potential issues when synchronizing the data:

(a) The manual collected event time stamps are prone to human errors, e.g., reaction time, distractions, and decision time, leading to time offsets of up to 20 s based on our investigations.
(b) The two systems use their own clock, causing systematic offsets between the two sources.

For these reasons, a "naive" synchronization—using the time stamp from the event data—to identify player positions at the time of an event leads to large inaccuracies. The upper plots in **Figure 2** display the coordinates of the players and the ball at the different moments of the scenario from **Figure 1** (right plot). The scene describes Kevin Volland's (Bayer Leverkusen) 1:0 against Borussia Dortmund (BVB) at the 14th matchday in the 2017/2018 season:[14] The upper right plot in **Figure 2** displays the shot time stamp tagged in the event data, which is roughly 2 s after the time stamp our synchronization suggests the shot took place (upper middle plot). The upper left plot in **Figure 2** shows the positioning of the players 2 s prior to that. As one can see, the

[11]Note that the *Gen 5* system has been in use since season 2019/2020, while all prior Bundesliga seasons were tracked using the *Gen 4* TRACAB version.
[12]The ground truth was measured by a VICON system, using an optoelectronic motion capture system based on markers placed on the tracked objects. Further details about this system can be found here: https://www.vicon.com/. An evaluation study of that system can be found in Merriaux et al. (2017).
[13]The study was conducted by the *Fédération Internationale de Football Association (FIFA)* in close cooperation with the Victoria University (Melbourne, Australia). An overview of the study can be found here: https://football-technology.fifa.com/en/media-tiles/fifa-quality-performance-reports-for-epts/, the report of the *Gen 5* system can be found here: https://football-technology.fifa.com/media/172171/chyronhegoopt-fifa-epts-report-nov2018.pdf (accessed December 26, 2020).
[14]https://www.youtube.com/watch?v=UdvrKfsJISY&feature=onebox&t=1m08s (accessed September 10, 2020).

situations are drastically different ranging from a distant dribble to a player celebrating his goal. The figure underpins that a shift of a few seconds in the synchronization can have a massive impact on the features used for the xG calculation, like the shot location or the goalkeeper position.

Therefore, we developed a synchronization algorithm tackling both issues. As a first step, we shift all tracking time stamps by the time difference between the kick-offs in both data sets. This resolves issue (b) and furthermore reduces a potential systematic delay in the manual event collection. In order to tackle issue (a), we compute several features that help to determine when a particular shot could have happened in the tracking data. First, we determine when the shooting player was in ball possession. We define potential individual ball possession sequences as the time interval when the player is in close proximity to the ball—our subject experts suggested 2 m as a cut-off, which is in line with Linke et al. (2018). Next, within each possession window, we identify the frame with the maximum acceleration of the Euclidean distance between player and ball. This aims to identify the exact moment where a shot occurred. Lastly, since there are potentially many situations that fulfill the above-mentioned criteria, we identify which best matches the event description. For that we compute Euclidean distances between the player and ball, the player and the manual collected event location as well as between the ball and the manual collected event location. Additionally, we compute the time difference between the (shifted) tracking time stamps and the manual collected event time stamp. We compute a weighted sum of these features, and the one frame out of the solution space that minimizes this weighted sum is chosen. The weights were obtained by performing a grid-based search that aimed to optimize accuracy of the synchronization on a manual labeled test set. The lower part of **Figure 2** shows how these features behave in the 20 s before and after the exemplary shot described above. When we applied this synchronization algorithm on the full data set of six seasons, the event shot times had an average absolute offset of 2.3 s (≈57 frames) from the synchronized frame. **Figure 3** displays histograms of the differences in timing (left) and locations (right) of each shot.

3.2. Evaluation of the Synchronization

In order to evaluate the accuracy of the synchronization, we manually annotated the timing of total 219 shots of the nine matches from matchday one of Bundesliga season 2018/2019. First, a full 90 min video animation of the 2D tracking data was created for each match. As a ground truth, we used a tactical video feed, which is filmed manually with an angle to capture all outfield player (and the most relevant goalkeeper). Additionally, for each match a xml-file[15] containing all shot-events, and the kick-off was produced. Next, we used the kick-offs in all three data sources to synchronize them manually as accurately as possible using Hudl Sportscode[16]—a dedicated tool for football video analysis with functionalities to combine different video sources and data sources (i.e., event data can be imported via xml-files). For each shot, we stop the video at the exact moment the shot occurred—defined as the first frame when the ball left the shooter—and extract this time point using Sportscode functionalities.

We now use these labeled shot timestamps as the ground truth and compare them with both, the results from our synchronization, and the event timestamps. Our synchronization displays an average absolute offset of 0.23 (±0.49) s, while the event timestamps differ by 1.82 (±4.06) s. Out of the 219 shots, we were able to synchronize 218, and 210 (95.9%) of these shots were < 0.3 s apart from the ground truth[17]. In contrast, only 63 (28.8%) of the event timestamps were within 0.3 s of the ground truth. It is evident that generally this synchronization is far superior to event timestamps. Two exemplary situations for a successful and an unsuccessful shot synchronization can be found here[18,19].

When a shot cannot be synchronized, it is typically due to either tracking data quality issues (e.g., the ball is poorly tracked, and never gets close to the player taking the shot, or two players were swapped in the tracking data) or event data quality issues (e.g., the wrong shooter is identified). To ensure that the quality of the input data is as high as possible, all shots that could not be synchronized at all were excluded from further analysis. Over the entire data set, this was the case in 3.4% of the shots.

All together, the synchronization of positional and event data presents a tremendous improvement for the analysis of shots, and could potentially be extended, using a similar algorithm, to other event types, like passes or tacklings. As we have seen above, misidentifying the shot time just slightly can cause a stark misrepresentation of its surrounding circumstances, and consequently affect the xG value significantly.

4. EXPECTED GOALS MODELING

4.1. Hand-Crafted Feature Extraction

To feed the supervised machine learning model, features influencing the goal scoring opportunity were defined together with professional match analysts from Bundesliga clubs and the German national team. A description of all features can be found in **Table 1**. In order to make full use of the synchronization of our two data sources, the features are based on both event and tracking data. The goalkeeper positioning is included in two features: We check whether they are in the line of shot, defined as the triangle between the shot location and the two posts, which is also the baseline for our shot angle calculation. Second, the distance between the goalkeeper and the goal is used as features

[15]Xml stands for *eXtensible Markup Language* and is an established format to transfer complex data files.

[16]https://www.hudl.com/products/sportscode (accessed June 20, 2020).

[17]We use a range here, because both, harmonizing the different video and data sources and the manual selection of the shot timestamp, may cause slight time discrepancies.

[18]In the first sequence, actual match-footage of a scene is shown. The second shows a 2D animation of the same scene, with a frame-counter on top. This frame counter counts down till 0 where the shot happened and increases afterwards again. The third sequence combines both video sources together (see **Supplementary Video 1**).

[19]See **Supplementary Video 2**.

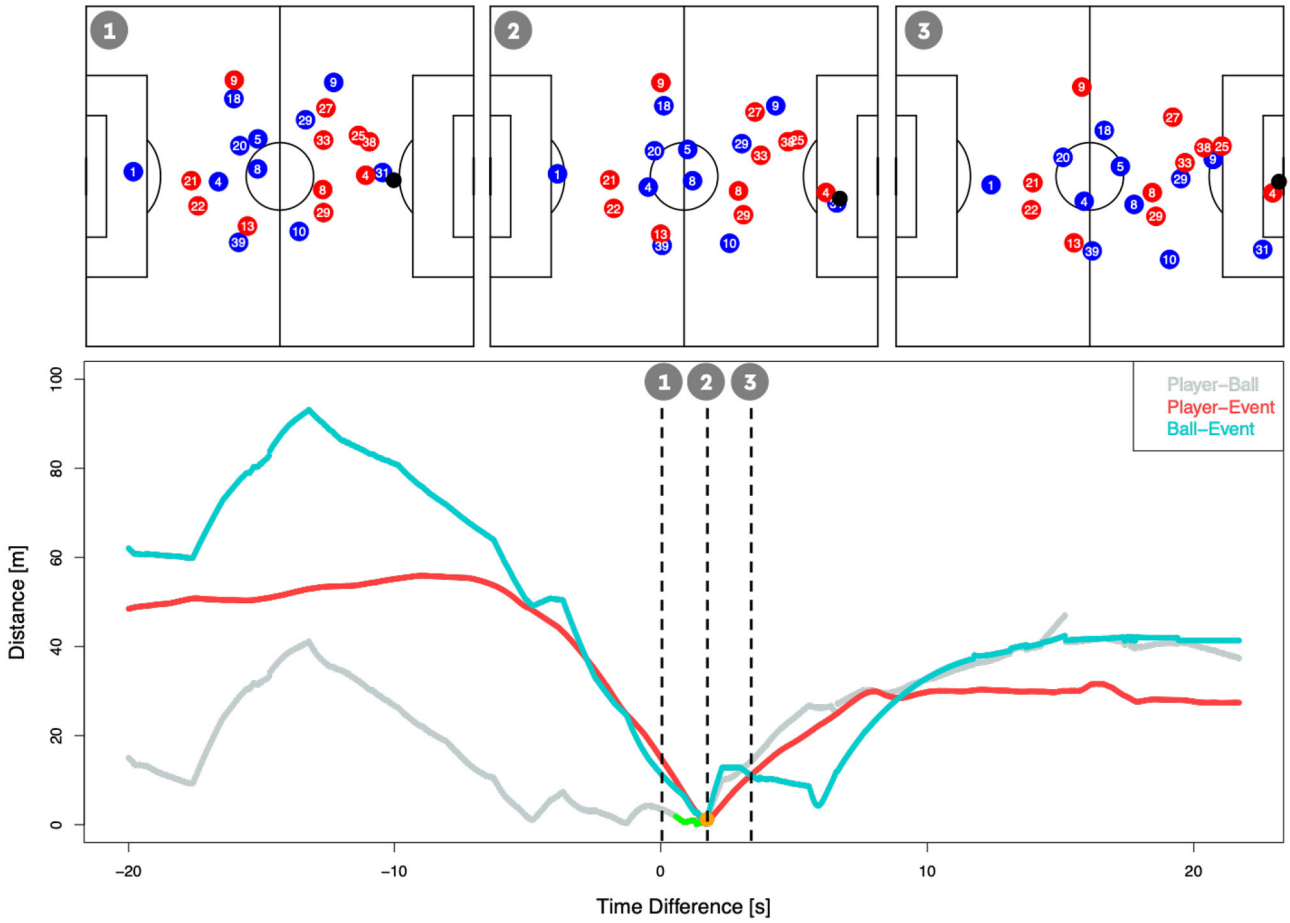

FIGURE 2 | Relevant metrics for the synchronization over time. The green points highlight the time interval where we detect a potential individual ball possession sequence. The orange point indicates where the shot event was finally detected.

in our model. The defending players' positions, either threatening to block the shot or applying pressure on the shooter, are also taken into consideration. Similarly to the goalkeeper feature, we count the number of defenders in the line of shot. Based on the logic from Andrienko et al. (2017), we calculate the total amount of pressure on the shooter aggregated over all defending players, as well as the maximum individual pressure on the shot-taking player. For both pressure metrics, we additionally compute the differences to the expected pressures given the shot location. Furthermore, the speed of the shooter, while taking the shot, is integrated in our model.

4.2. Predict the Scoring Probability as a Supervised Machine Learning Task

For a total of 105, 627 shots, all features from **Table 1** were calculated based on the synchronized positional and event data. Since the features *shot type* and *freekick* significantly influence the contribution of all other features, we split our problem into three subtasks: the prediction of goal scoring probabilities of open play leg-shots, headers, and direct freekicks. Per subtask, the optimal set of features was explored. Consequently, for all three subtasks we trained several supervised machine learning models based on 81, 462 open play leg-shots, 18, 748 headers and 5, 417 direct freekicks, respectively, labeled by the information whether the shot ended up in a goal (1) or not (0). For each subtask, the shots were randomly split into 60% training, 20% validation, and 20% test data sets. To avoid over representing teams or scores, this split was conducted for every match separately. The final model, shown in **Table 1** (row 5), describes the combination of our three submodels. To investigate the efficiency of the division into the three subgroups, another model is trained based on all 105, 627 shots taking all features from **Table 2** including the information whether the shot was a header, a leg-shot from open play or a direct freekick.

Various standard supervised machine learning models were trained on the training data set, hyperparameters were optimized on the validation data set and the models' accuracy's were evaluated on the test data set. Naturally, the necessary hyperparameters depend on the machine learning algorithm. In the case of the extreme gradient boosting model (hereafter referred to as XGBoost), the parameters we optimized are as follows: *Learning rate*: controls the step size used per update; *Max*

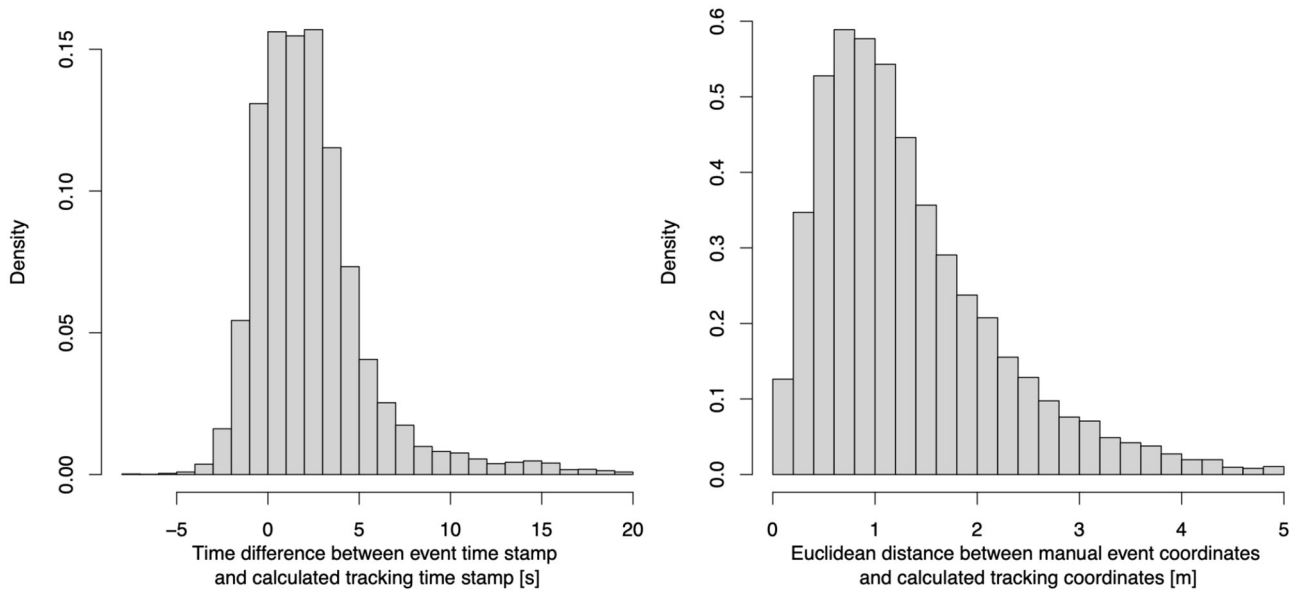

FIGURE 3 | Time stamp **(left)** and shot location **(right)** differences between event and synchronized time stamps.

TABLE 1 | Features derived from synchronized positional and event data used to train our model.

Feature	Value	Description
Shot location	Numeric	The x, y and the z-coordinate of the ball at the time of the shot are used for several features, such as angle and distance to goal center.
Speed of player taking the shot	Numeric	The speed of the player attempting the shot, at the time of the shot (in *km*/*h*).
Defenders in the line of the shot	Numeric	The number of defenders in the line of the shot.
Goalkeeper position	Numeric	The position of the goalkeeper is used for two different features, describing whether they are in the line of shot and their distance to the goal.
Pressure on the player taking the shot	Numeric	Various metrics describing the pressure that the player was under while attempting the shot, at the time of the shot Andrienko et al., 2017.
Type of shot	Categorical	Describing the body part used for the shot *(Head, leg* or *other)*.
Taker ball-control	Categorical	Describes how the player taking the shot gained control of the ball before/when taking the shot *(volley, controlShot, dribblingLess10m, dribblingMore10m, setPiece)*.
After freekick	Categorical	Indicates whether the shot followed a freekick.
Freekick	Categorical	Describes whether the shot is a direct freekick or not.

depth: limits the depth of the tree; *Subsample*: controls number samples applied to the tree; *Min child weight*: controls instance weight of a node. For the optimization, we applied Bayesian tree-structured Parzen Estimator hyperparameter optimization approaches for the gradient boosting model (Bergstra et al., 2011; Dewnacker et al., 2016; Wang, 2019).

For several models in **Table 2**, we calculated SHAP values per feature (Roth and Thomson, 1988; Lundberg and Lee, 2017; Rodríguez-Pérez and Bajorath, 2020). In several applications, using SHAP values[20] instead of standard gain values has proven to be beneficial (Antipov and Pokryshevskaya, 2020; Ibrahim et al., 2020; Meng et al., 2020).

In order to get a better understanding of the resulting model's accuracy, we implemented two simple models as a baseline models. The first one uses an attribute that is collected for every shot (*chance quality*). This manually collected attribute can contain one of the following two values: sitter or chance. The very simple model now assigns each shot the average conversion rate of the corresponding class. So all shots labeled as *chances* are assigned a value of 0.063, while the remaining shots labeled as *sitters* receive a value of 0.548. The second baseline model uses all the event data based features from **Table 1** (namely *Shot location*, *Type of shot*, *Taker ball-control*,

[20]The abbreviation **SHAP** stands for **SH**apley **A**dditive ex**P**lanation.

TABLE 2 | Statistical evaluation of the expected goal model outcome.

	Model	Precision	Recall	AUC	RPS
1	Gradient boosting (all situations)	0.646	0.181	0.822	0.196
2	Logistic regression	0.611	0.108	0.807	0.160
3	ADA boost	0.548	0.201	0.816	0.076
4	Random forest	0.611	0.163	0.794	0.165
5	Gradient boosting combined	0.665	0.164	0.823	0.197
	Leg-shot model	0.668	0.171	0.825	0.201
	Header model	0.655	0.161	0.813	0.187
	Direct freekick model	–	0	0.830	0.099
6	Chance evaluation model	0.516	0.420	0.688	0.170
7	Event data based model	0.587	0.098	0.772	0.118

After freekick, and *Freekick*), and train a XGBoost model using these features.

4.3. Statistical Evaluation of the Shot Prediction Model

The first two validation metrics (precision and recall) presented in **Table 2** evaluate the outcome of a classification problem. A goal classified with an xG above 50% is classified as a true positive, whereas an unsuccessful shot with an xG below that threshold is defined as a true negative. Thereafter, a recall of 1 could simply be achieved by assigning each shot an xG value above 50%. To incorporate both the true positive and the false positive rate depending on the threshold into our evaluation, we also use the area under the receiving operator curve (*AUC*) as an error function (Daskivich et al., 2018). However, it is our objective to assess the accuracy of the underlying goal scoring probabilities and not just of a binary classification (goal or no goal). While this is possible with the AUC, using the ranked probability score (*RPS*), as presented in Murphy (1970), fulfills this purpose better, especially for imbalanced data sets.

By splitting up the shots into two groups (chances and sitters), the chance evaluation model (**Table 2**, row 6) achieves a good balance between precision and recall. While this relatively simple model already achieves a somewhat satisfactory RPS of 0.170, the human-made classifications are possibly biased by the shot outcomes. This label is therefore not used as a feature for the remaining prediction models. For the event data based model, the extremely low recall can be interpreted as follows: The model predicts xG value below 50% for most of the shots that actually end up as goals. However, the AUC shows that the event-based model yields more granular predictions than the chance evaluation model. In the direct freekick submodel, no xG prediction exceeds 50%, and therefore its precision is undefined.

Shots are non-deterministic, at the time of the shot, meaning that no model can have a 100% accuracy predicting whether any given shot will score. But what we can expect from our model predictions is that they converge over a large sample. To verify this, we looked at the first 54 matches (matchday one through three) of the 2020/2021 season in Bundesliga and 2nd Bundesliga. Out of the 1,357 shots, 150 found the back of the net and our model predicted an aggregated xG value of 151.6.

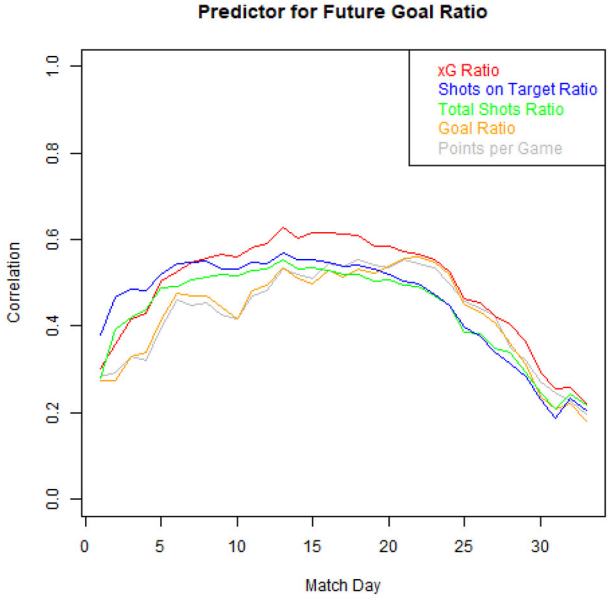

FIGURE 4 | Correlation between a team's future goal ratio after a certain matchday and an aggregated metric before said matchday (average of all seasons 2013/2014–2019/2020).

Estimating a team's true strength or its future performances is a crucial unsolved problem in football with many potential use cases (Goes et al., 2019). Both shots on target, two well-established metrics in the literature, have been used for this context (Lamas et al., 2014). **Figure 4** displays in which scenarios our xG values fulfills this task better than traditional approaches. It looks at how well you can predict a team's future rest of the season goal ratio (defined as the difference between goals scored and goals conceded) after a certain matchday, by only taking into account one aggregated metric before said matchday. On the *y*-axis, the correlation between the future goal ratio and the respective metrics (see legend) before that matchday (*x*-axis) is shown. Consistently, over all considered seasons a team's historic xG values are able to predict future results better than traditional metrics, especially between matchday 10 and 20. Additionally, we found that in 73.3% of all matches (excluding draws), the winner had a higher xG value[21], while only in 56.2% of these games, the winning team had more shots, than its opponent.

Next, we analyze the features' influence on the predicted goal scoring probability. In the following, we discuss the overall feature importance of our gradient-boosting model trained on all shots with the subcategories as features (**Table 2**, row 1). **Figure 5** displays the overall influence according the respective SHAP values per feature on the right, which can be interpreted as an aggregated quantification of the feature's influence. The SHAP values show that the most crucial factors are the shot location

[21]On a match and team level the overall xG balance between the two teams is considered here. For both teams, we sum up the xG values per team of all their shots.

(Goal Distance, Angle) and the goalkeeper position *(Distance Goalkeeper to Goal)*. *Maximum Individual Pressure Diff*, defined as the difference between the actual pressure and the average pressure given the shot location, has the third highest influence on the predicted values. In **Figure 5** (left plot), the x-value of each colored dot displays how a feature influences the model, whereas the color scaling describes the value of the respective feature. Both a flat line and a smooth change of colors (from left to right or vice versa) indicates a roughly linear correlation between the feature value and the model outcome. In **Figure 6**, this relationship between the feature values (x-axis) and influence on the model (y-axis) is shown more granularly. Although the red line shows a regression, the dispersion of the blue dots provide a deeper insight. Both the left plot in **Figure 5** (smooth decrease of the colored dots from left to right) and **Figure 6** (red line) shows that the goal distance has an almost linear impact on the predicted values. However, if the distance to the goal is very high, influence relies more on other features, as can be seen by the growing dispersion of the blue dots. The importance of the number defenders in the line of the shot (here *Defenders*) underpins the relevance of using positional data, including all opposing players' positions. Looking deeper into the SHAP distributions of this feature, **Figure 6** shows an almost linear decrease of the average SHAP value over all shots from zero to four defenders in the line of shot. For more defenders in the line of shot, the average SHAP value—describing a proxy for the features influence—remains mostly constant. In **Figure 6**, the feature *Goalkeeper in the goal* underpins our practitioners' intuitive assumption and can be interpreted as follows: If the goalkeeper is not in the line of shot, it increases the xG value significantly.

Again, most of this information would not be available in event data, which highlights the benefit of using both event and positional data once more.

4.4. Evaluation by Subject Matter Expertise

In several workshops with match analysts from Bundesliga clubs and the German national team, the features were defined and ranked according to the estimated influence. These estimations were compared with the above calculated feature importance. Additionally, the SHAP value dispersions and interpretations were discussed in detail. Besides a lot of agreement from practitioners, some statistical results—, e.g., the influence of 4–10 players in the line of shot—were discussed intensively among experts. To evaluate the plausibility of our model from a practitioners perspective, a workshop with selected (assistant) coaches of Bundesliga and 2nd Bundesliga clubs was conducted. For the recently concluded season, the coaches were asked to classify their matches into four categories: *deserved* or *undeserved* victories, draws, or losses as in **Figure 8**. Afterwards, we compared their labels to the ones produced from our xG model. With a category-accordance of more than 85% (in total 102 matches with 293 goals), practitioners characterized our approach as a helpful tool to assess individual shot qualities and the overall performance of a team.

5. APPLICATION AND DISCUSSION

5.1. Applications

For the following section, we consider the 2019/2020 season of the German Bundesliga, with in total 306 matches, 954 goals, and 5,450 shots. We describe how the goal scoring probability $xG(S)$ model for a given shot S is aggregated over a season to evaluate teams and players further:

$$xG_{\text{agg}}(\text{Team/Player}) = \sum_{S_i \in \text{Shots}} xG(S_i)$$

Own goals are not a subtype of a shot event, but rather a separate event type with different attributes. Therefore, they are excluded from our xG calculation. Penalties are assigned an xG value of 0.766, which is the average conversion rate in the Bundesliga history. In the case of so-called double-chance, situations in which a first shot is blocked, but is immediately followed up by a rebound shot, we calculate xG values for each shot. But when we aggregate the team level xG values, we do not want to simply add them up, because it could lead to situations where a teams xG value for small time-window could exceed 1. Therefore, given a double-chance S, defined as two shots within 5 s, we compute the overall probability as:

$$xG(S) = xG(S_1) + \left(xG(S_2) * xG(\bar{S_1})\right)$$

5.1.1. Teams

Figure 7 displays how many goals each team scored and conceded in comparison to the aggregated xG values our model computed. Consequently, for the 2019/2020 season, BVB (sixth place in the left ranking of **Figure 7**) scored roughly 30 more goals than the sum of all the respective shots' xG values would suggest. **Figure 8** provides a closer look at BVB efficiency on a match level. Comparing actual goal differences to the xG differences, the upper right quadrant could be interpreted as *deserved* wins, where BVB created more promising shot opportunities than their opponents. Matches on the lower right could be interpreted as lucky wins, e.g., the return match[22] against Borussia Mönchengladbach (black and white hatched diamond logo in the bottom right of the left figure).

Another match, where our model would have predicted a different result is displayed in **Figure 9**[23]. The graph shows the aggregated xG values per team over the course of a match. Although SC Freiburg displayed an extraordinary shooting efficiency, by scoring three goals out of three difficult situations, Eintracht Frankfurt created several high quality chances but only converted three of them.

Furthermore, our model can help match analysts examine a teams' shooting behavior. **Figure 10** presents the number of shots taken vs. the average xG-value per team (left) and for the most scoring strikers (right). Although Fortuna Düsseldorf (red/white logo furthest left in **Figure 10**) had an average xG value (∅(xG)) of 0.08 in the 2019/2020 season, Borussia Mnchengladbach

[22]https://www.youtube.com/watch?v=RUaORAinaoc&feature=onebox (accessed October 2, 2020).

[23]https://www.youtube.com/watch?v=jl1C0KsIqaQ (accessed October 2, 2020).

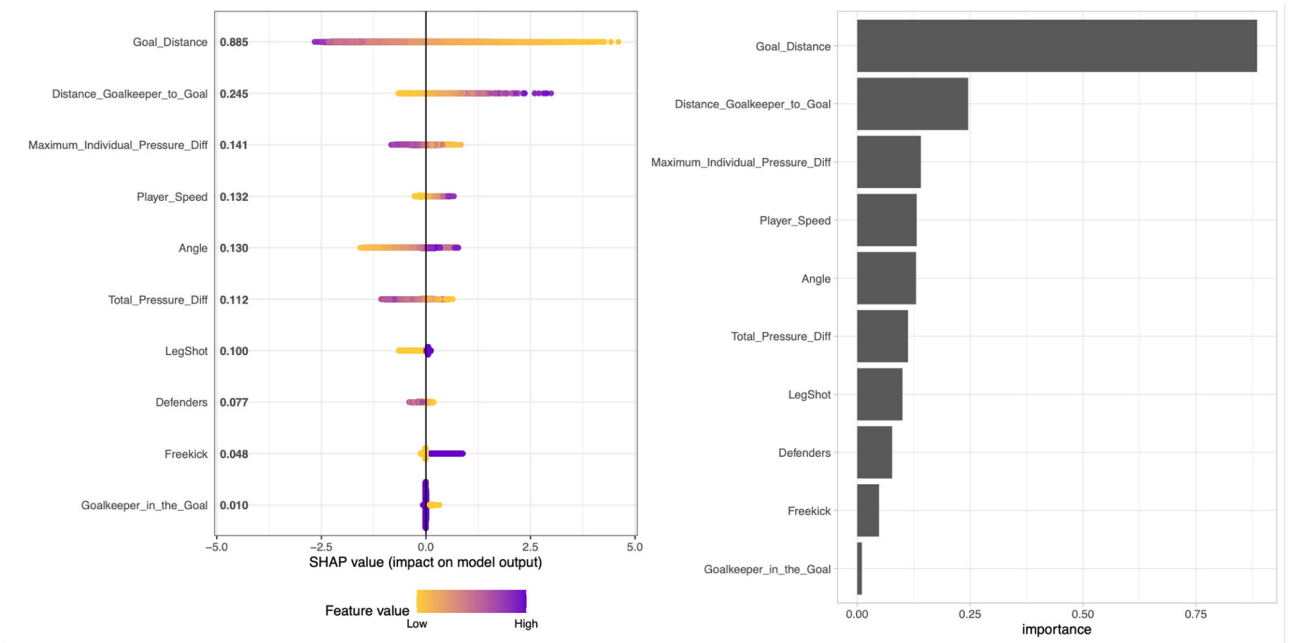

FIGURE 5 | Feature importance according to Shapley values displayed as a SHAP summary plot **(left)** and global feature contributions by the mean SHAP value across all samples **(right)**.

seems to take their shots only in cases of a clear scoring opportunity ($\varnothing(\text{xG}) = 0.14$). FC Bayern Munich (red/blue/white logo top right in **Figure 10**), takes by far the most shots per game. However, with around four less shots per match, Borussia Mönchengladbach has a higher quality of attempts according to our xG model. Comparing FC Augsburg (red/white/green logo with *FCA* inscription) to Werder Bremen (green diamond logo with a white *W* as an inscription) shows two distinct patterns. While both teams had a similar number of aggregated xGs over the whole season (see **Figure 7**), Bremen tends to take more shots in less promising situations, while FC Augsburg emphasizes more on taking their shots in situations with a higher goal scoring probability. Having this information for the next opponent prior to each match can help teams to adapt their defending strategy depending on the opponent's shooting preferences.

5.1.2. Players

Additionally, we can use player aggregated xG values, both for individual player performance analysis as well as scouting. Comparing Jadon Sancho to Serge Gnabry shows that both players—playing in similar positions and both with very successful teams—have strongly differing shooting patterns. Although Serge Gnabry (top left in **Figure 10**) takes the second most shots per match, Jadon Sancho (lowest in **Figure 10**) takes the fewest shots out of the top 10 scorers, but often in more promising situations according to the xG-values. Besides an overview of strikers shooting behavior in **Figure 10**, xG provides a lot more applications to quantify a player's offensive contribution more granularly than traditional metrics.

Since our xG model can be seen as an average across all Bundesliga players' shot efficiency, it can also be used to find players that convert shots at an above average rate. Using this approach, we see that Robert Lewandowski (upper right in **Figure 10**) outscored his aggregated xG value (29.6) by about four goals, scoring a total of 34 in the season out of his 140 shots (**Table 3**, row 12). While this is already an impressive feat, there were in total 11 players, outscoring their xG totals by a larger margin. Jadon Sancho (17 goals/53 shots/8.49 xG_{agg}) and Erling Haaland (13 goals/34 shots/7.59 xG_{agg}) lead this category and showed an extraordinary scoring efficiency.

5.2. Discussion

We present an xG model that performs better than any of the approaches discussed in the introduction. Rathke (2017) split the pitch into eight zones and trained a logistic regression on each, indirectly taking shot location and angle into consideration. However, their analysis was neither tested on unseen data nor took the positions of defenders and goalkeepers into consideration. By contrast, Lucey et al. (2014) did not only make use of positional data, but also displayed the improvements of the model accuracy. They split all shots into six different game-context situations *(open play, counterattack, corner, penalties, freekicks, set pieces)* and also learned a regressor for each. Their average error across all shots and scenes is 0.1439. In our final combined model (**Table 2**, row 5), this average error is 0.0928. As a combination of the larger data set (more than 100,000 shots), our novel synchronization approach (see section 3) and the expert crafted features (see section 4.1) are possible reasons for this improvement.

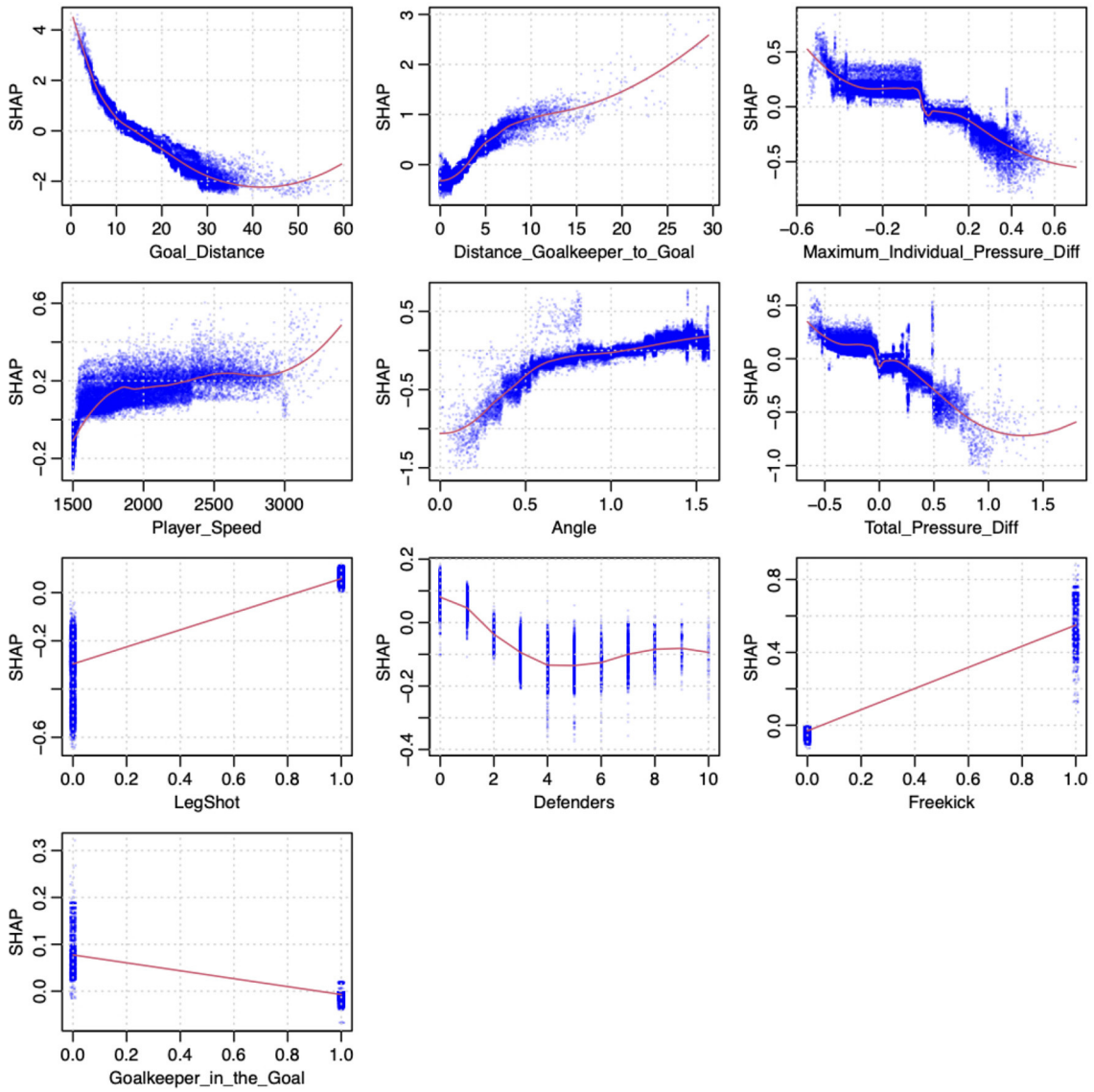

FIGURE 6 | SHAP dependence Plot. For each shot, the respective feature value is plotted on the x-axis vs. the corresponding Shapley values on the y-axis (distance is displayed in meter, and speed is shown in meter per hour).

However, xG models in football are not without flaws. An often criticized point is that they are not evaluating dangerous situations where no shot took place. While this criticism certainly has merits, most offensive actions end up in shots. The official Bundesliga event data include an event type *chance without a resulting shot*, describing situations, where a team was in a scoring position, but failed to attempt a shot. In our data set, this event occurs on average only 0.93 times per match, underlining that the impact non-shot opportunities have for measuring team performance is rather small. Additionally, as seen in section 5.1, evaluating team strength is not the only application of xG. Shot conversion on team/player level, average shot quality or even on a goalkeeper analysis are insightful use cases that only depend on actual shots taken. Nevertheless, several studies aim to tackle this problem, of noteworthy goal-scoring opportunities without shots, by computing so-called expected possession values (Link et al., 2016; Spearman, 2018; Fernández et al., 2019), but even these concepts are often build upon a well-calibrated xG model.

Following the logic of expected possession values, it is definitely a potential next step to break the contribution to a goal scored further down to the participating players and their actions. For instance, in the situation described in **Figure 2** by assuming shots at several time-points, a simple rule-based approach using our xG model can quantify how much xG Volland added through his dribbling. Another popular extension of xG are expected assists (xA), which measure the likelihood that a pass leading to a shot becomes an assist, by assigning it the resulting xG value. This allows to quantify a player's

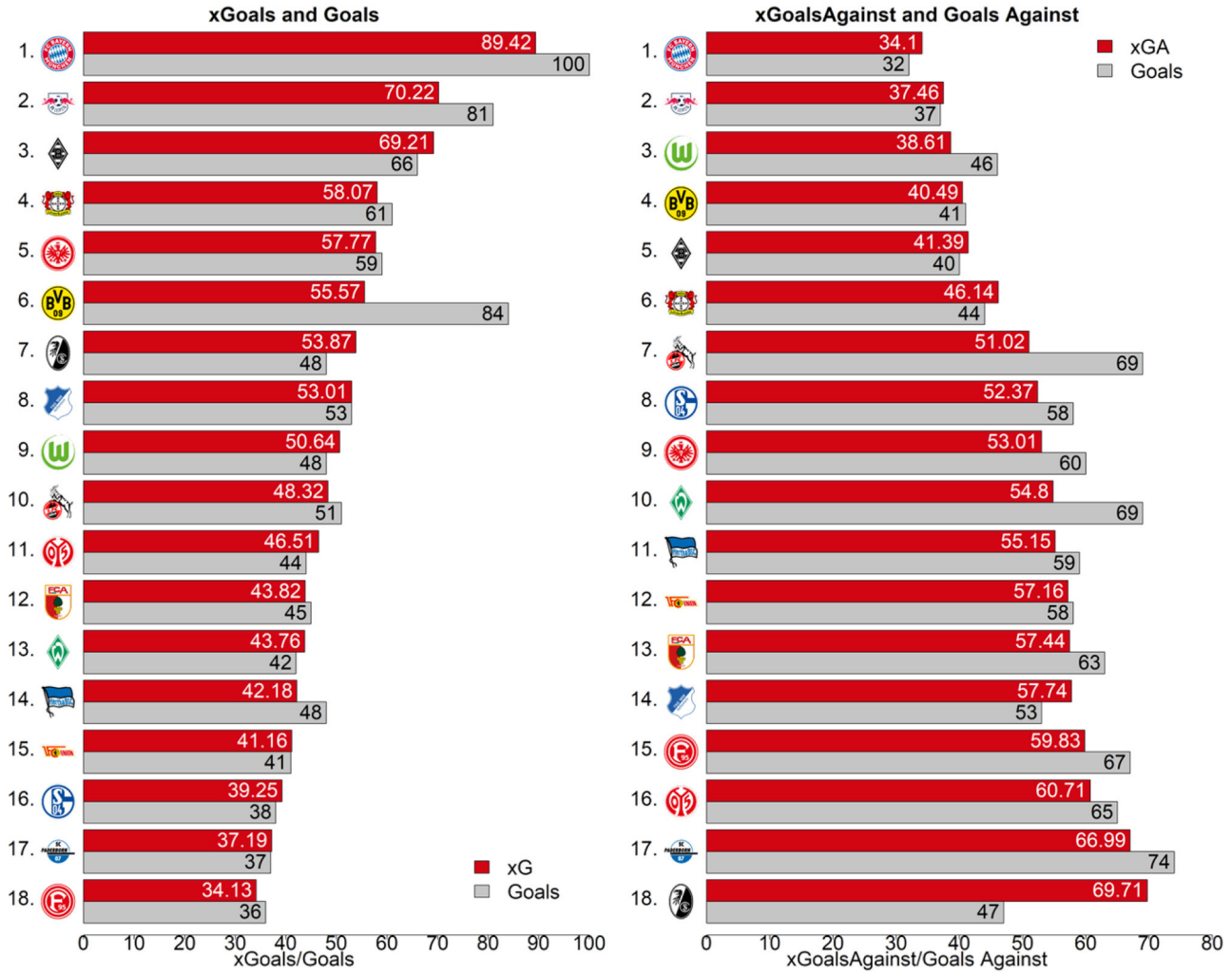

FIGURE 7 | Bundesliga 2019/2020 season ranking with aggregated xG and the actual number of goals (xG red, actual goals gray).

shot assisting qualities independent of the final shooter's ability to score.

Both the synchronization and the inputs for the xG model heavily rely on the quality of the underlying data. Even for purely event data based xG models, Robberechts (2019) showed that their usefulness strongly depends on the event data quality. One of the parameters causing the biggest inaccuracy in the current model is the ball height. Small objects—like a ball—are hard to track based on video footage, especially due to confusion with replacement balls or other small white objects occurring in the stadium. For header shots, little differences in the ball height have a large impact on the ability of a player to control the placement of a shot causing inaccuracies for our current header model (**Table 2**, row 7). With a steady increase of video camera resolutions and object detection algorithms, we expect a significant improvement for ball tracking. This increase in data quality would likely improve shot synchronization results even further (see section 3.2) and consequently result in even more accurate xG models. Nevertheless, both for tracking data (including ball tracking) and for event data additional evaluation studies to ensure a high data quality for similar projects is essential. Although latest positional and event data provide accurate and detailed information about players, their body orientation and limb tracking could further improve the model's accuracy. For the header model in particular, heights and jumping altitude capacities could be taken into consideration as well.

The harmonization of tracking and event data is not a problem unique to football, which has been barely explored in the literature. In basketball, for instance, the two data sources[24] are mainly used independently of one another (Tian et al., 2020), but as Manisera et al. (2019) noted the combination of both data sources is a crucial future issue. While our algorithm is optimized

[24]In basketball, event level data are often referred to as *play-by-play* data.

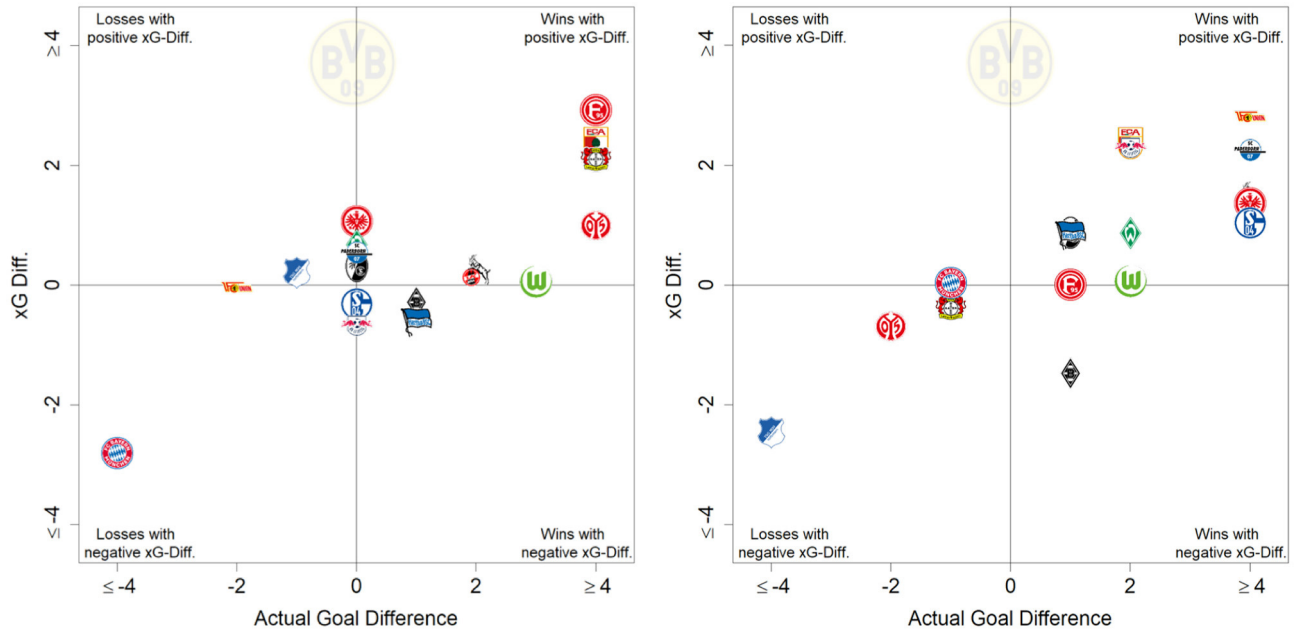

FIGURE 8 | Season report of BVB in season 2019/2020 showing efficiency of BVB matches according to the underlying xG values.

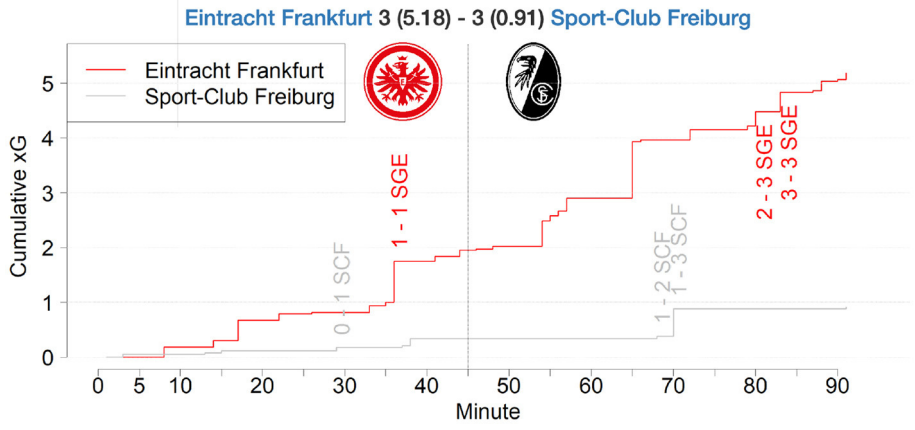

FIGURE 9 | xG match report of a Bundesliga match between SC Freiburg and Eintracht Frankfurt in season 2019/2020.

for football events, it could be adapted and applied to several other sports where both data sources are available.

An accurate expected goals model provides tremendous decision-making support for clubs: Creating many high-quality shooting situations is a crucial indicator of a good performance. To which extent these situations actually end up in goals often depend on random factors or luck. Consequently, a single final match result may not represent the actual team performance accurately. By quantifying a team's conversion rate (goals vs. xG) separately from their aggregated offensive contribution (created xG), clubs can evaluate the performance of their players, teams, and coaches objectively. Future research could even go one step further and explore how this work could affect the way the game is played. One could use our goal probabilities to determine numerically in which situations it is beneficial to shoot, and when one is better of risking an additional dribble or pass. Another area where the use of xG could be explored further are media applications: Recently, media and broadcasting have included xG values in their match coverage. For each goal occurring in German Bundesliga, different broadcasters have chosen to display our xG value seconds after the goal occurred[25].

[25]https://www.dfl.de/en/news/bundesliga-and-amazon-web-services-to-develop-next-generation-football-viewing-experience/ (accessed September 10, 2020).

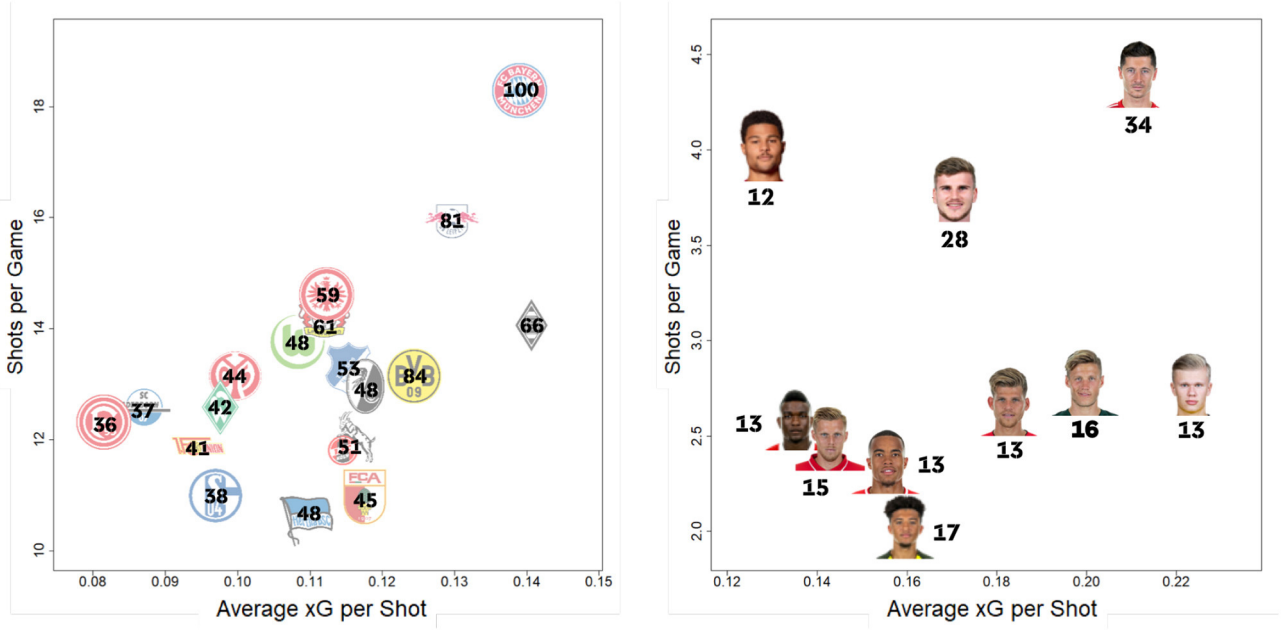

FIGURE 10 | Quality vs. quantity of shots taken per team (left) and player (right). The total number of goals scored over the whole season per team and player is displayed in black.

TABLE 3 | Players with the highest scoring efficiency in the German Bundesliga 2019/2020 season.

	Name	Club	Minutes	xG	Goals	Shots	CR
1	J. Sancho	Dortmund	2,386	8.49	17	53	2.002
2	E. Haaland	Dortmund	1,117	7.59	13	34	1.712
3	J. Cordoba	Köln	2,107	8.41	13	60	1.545
4	R. Hennings	Düsseldorf	2,598	9.92	15	71	1.512
5	A. Kramaric	Hoffenheim	1,496	8.61	12	42	1.393
6	T. Werner	Leipzig	2,934	20.79	28	122	1.346
7	R. Quaison	Mainz	2,727	10.72	13	69	1.212
8	A. Silva	Frankfurt	1,671	9.91	12	55	1.210
9	K. Havertz	Leverkusen	2,570	10.13	12	56	1.184
10	M. Reus	Dortmund	1,568	9.31	11	47	1.181
11	N. Petersen	Freiburg	2,588	9.44	11	54	1.165
12	R. Lewandowski	Bayern	2,888	29.57	34	140	1.149
13	S. Andersson	Union Berlin	2,821	11.68	12	64	1.027
14	S. Gnabry	Bayern	2,288	12.74	12	100	0.941
15	W. Weghorst	Wolfsburg	2,898	17.59	16	88	0.909
16	F. Niederlechner	Augsburg	2,858	14.93	13	82	0.870

CR describes the conversion rate from xG to goals.

Now that the amount of data-driven approaches to support tactical analysis in football is increasing (Goes et al., 2020), more qualitative studies might help to underpin the statistical evaluation of models like xG. Although we present a first attempt toward an expert-based evaluation of our approaches (see sections 3.2 and 4.4), there is a lot of potential for further investigations, which could also serve to establish data-driven methods in the sport science and football community.

6. CONCLUSION

We present a meaningful proxy for goals scored in football, which helps to evaluate players' and teams' performance more accurately and objectively. Our xG model is based on a huge data set of cutting-edge and consistently acquired positional and event data that we combined using our own synchronization algorithm.

It exceeds traditional metrics significantly when evaluating strikers' (**Table 3**) and teams' (**Figure 7**) scoring efficiency, when evaluating single match performances (i.e., teams with higher xG win 73.3% of all not-drawn matches) and even when predicting future match results (**Figure 3**). It also allows us to evaluate assist performances of players independent of the striker's final touch. Additionally, several future potentials are shown for sport and data science research.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: data are property of DFL/DFB e.V. and thus can not be shared publicly. Requests to access these datasets should be directed to Sportec Solutions AG, DFL e.V., DFB e.V.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

GA was responsible for the implementation of the approach and involved in all discussions with practitioners. PB focused on the practical evaluation and the communication with practitioners. Both authors conducted the scientific studies together and were equally involved in the writing process of the manuscript.

ACKNOWLEDGMENTS

This work would not have been possible without the perspective of professional match analysts from world class teams who helped us to define relevant features and spend much time evaluating (intermediate) results. We would cordially like to thank Dr. Stephan Nopp and Christofer Clemens (match analysts of the German National team), Jannis Scheibe (head match-analyst of the German U21 mens national team) as well as Sebastian Geißler (former match-analyst of Borussia Mönchengladbach). Additionally, the authors would like to thank Dr. Hendrik Weber and Deutsche Fußball Liga (DFL)/Sportec Solutions AG for providing the positional and event data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspor.2021.624475/full#supplementary-material

We also provide an example of the media live application in German Bundesliga[26].

[26]https://www.youtube.com/watch?v=5flVB9ef0uM.

REFERENCES

Andrienko, G., Andrienko, N., Budziak, G., Dykes, J., Fuchs, G., von Landesberger, T., et al. (2017). Visual analysis of pressure in football. *Data Mining Knowl. Discov.* 31, 1793–1839. doi: 10.1007/s10618-017-0513-2

Antipov, E. A. and Pokryshevskaya, E. B. (2020). Interpretable machine learning for demand modeling with high-dimensional data using gradient boosting machines and shapley values. *J. Rev. Pricing Manage.* 19, 355–364. doi: 10.1057/s41272-020-00236-4

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011* (Granada), 1–9.

Beshai, P. (2014). *Buckets: Basketball Shot Visualization*. Semantic Scholar Preprint, 1–14.

Chang, Y. H., Maheswaran, R., Su, J., Kwok, S., Levy, T., Wexler, A., et al. (2014). "Quantifying shot quality in the NBA," in *MIT Sloan Sports Analytics Conference* (Boston, MA), 1–8.

Daskivich, T., Luu, M., Noah, B., Fuller, G., Anger, J., and Spiegel, B. (2018). Differences in online consumer ratings of health care providers across medical, surgical, and allied health specialties: observational study of 212,933 providers. *J. Med. Internet Res.* 20, 29–36. doi: 10.2196/jmir.9160

Davis, J., and Robberechts, P. (2020). "How data availability affects the ability to learn good xG models," in *7th International Workshop of Machine Learning and Data Mining for Sports Analytics* (Ghent). doi: 10.1007/978-3-030-64912-8_2

Dewnacker, I., McCourt, M., and Clark, S. (2016). Bayesian optimization for machine learning. a practical guidebook. *arXiv* 2–5.

Draschkowitz, L., Draschkowitz, C., and Hlavacs, H. (2015). Using video analysis and machine learning for predicting shot success in table tennis. *EAI Endorsed Trans. Creat. Technol.* 2:150096. doi: 10.4108/eai.20-10-2015.150096

Fairchild, A., Pelechrinis, K., and Kokkodis, M. (2018). Spatial analysis of shots in MLS: a model for expected goals and fractal dimensionality. *J. Sports Anal.* 4, 165–174. doi: 10.3233/JSA-170207

Fernández, J., Bornn, L., and Cervone, D. (2019). "Decomposing the Immeasurable Sport: a deep learning expected possession value framework for soccer," in *MIT Sloan Sports Analytics Conference*, 1–18.

Goes, F., Kempe, M., and Lemmink, K. (2019). "Predicting match outcome in professional Dutch football using tactical performance metrics computed from position tracking data," in *MathSport International Conference* (Athens), 4–5. doi: 10.29007/4jjb

Goes, F. R., Meerhoff, L. A., Bueno, M. J. O., Rodrigues, D. M., Moura, F. A., Brink, M. S., et al. (2020). Unlocking the potential of big data to support tactical performance analysis in professional soccer: a systematic review. *Eur. J. Sport Sci.* doi: 10.1080/17461391.2020.1747552. [Epub ahead of print].

Harmon, M., Lucey, P., and Klabjan, D. (2016). Predicting shot making in basketball learnt from adversarial multiagent trajectories. *arXiv*.

Hedar, S. (2020). *Applying machine learning methods to predict the outcome of shots in football outcome of shots in football* (Thesis), Uppsala University, Uppsala, Sweden.

Ibrahim, L., Mesinovic, M., Yang, K.-W., and Eid, M. A. (2020). Explainable prediction of acute myocardial infarction using machine learning and shapley values. *IEEE Access* 8, 210410–210417. doi: 10.1109/ACCESS.2020.3040166

Jagacinski, R. J., Newel, K. M., and Isaac, P. D. (2019). Predicting the success of a basketball shot at various stages of execution. *J. Sport Psychol.* 1, 301–310. doi: 10.1123/jsp.1.4.301

James, B. (1985). *The Historical Baseball Abstract*.

Lamas, L., Barrera, J., Otranto, G., and Ugrinowitsch, C. (2014). Invasion team sports: strategy and match modeling. *Int. J. Perform. Anal. Sport* 14, 307–329. doi: 10.1080/24748668.2014.11868723

Link, D., Lang, S., and Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLoS ONE* 11:e0168768. doi: 10.1371/journal.pone.0168768

Linke, D., Link, D., and Lames, M. (2018). Validation of electronic performance and tracking systems EPTS under field conditions. *PLoS ONE* 13:e0199519. doi: 10.1371/journal.pone.0199519

Linke, D., Link, D., and Lames, M. (2020). Football-specific validity of TRACAB's optical video tracking systems. *PLoS ONE* 15:e0230179. doi: 10.1371/journal.pone.0230179

Linke, D. M. (2019). *Validation of methodology, design & applications* (Ph.D. thesis), Technische Universität München, Munich, Germany.

Lucey, P., Bialkowski, A., Monfort, M., Carr, P., and Matthews, I. (2014). ""Quality vs Quantity": improved shot prediction in soccer using strategic features from spatiotemporal data," in *MIT Sloan Sports Analytics Conference*, 1–9.

Lundberg, S. M., and Lee, S. I. (2017). "Consistent feature attribution for tree ensembles," in *Proceedings of the 34th International Conference on Machine Learning* (Sydney), 1–9.

Macdonald, B. (2012). "An expected goals model for evaluating NHL teams and players," in *MIT Sloan Sports Analytics Conference 2012* (Boston, MA), 1–8. doi: 10.1515/1559-0410.1447

Manisera, M., Metulini, R., and Zuccolotto, P. (2019). Basketball analytics using spatial tracking data. *Springer Proc. Math. Stat*. 288, 305–318. doi: 10.1007/978-3-030-21158-5_23

Meng, Y., Yang, N., Qian, Z., and Zhang, G. (2020). What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values. *J. Theor. Appl. Electron. Comm. Res*. 16, 466–490. doi: 10.3390/jtaer16030029

Merriaux, P., Dupuis, Y., Boutteau, R., Vasseur, P., and Savatier, X. (2017). A study of vicon system positioning performance. *Sensors* 17, 1–18. doi: 10.3390/s17071591

Murphy, A. H. (1970). The ranked probability score and the probability score: a comparison. *Mon. Weather Rev*. 98, 917–924. doi: 10.1175/1520-0493(1970)098<0917:TRPSAT>2.3.CO;2

Pollard, R., and Reep, C. (1997). Measuring the effectiveness of playing strategies at soccer. *J. R. Stat. Soc. D Stat*. 46, 541–550. doi: 10.1111/1467-9884.00108

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *J. Hum. Sport Exerc*. 12, S514–S529. doi: 10.14198/jhse.2017.12.Proc2.05

Redwood-Brown, A., Cranton, W., and Sunderland, C. (2012). Validation of a real-time video analysis system for soccer. *Int. J. Sports Med*. 33, 635–640. doi: 10.1055/s-0032-1306326

Reich, B. J., Hodges, J. S., Carlin, B. P., and Reich, A. M. (2006). A spatial analysis of basketball shot chart data. *Am. Stat*. 60, 3–12. doi: 10.1198/000313006X90305

Robberechts, P. (2019). "Valuing the art of pressing," in *StatsBomb Innovation in Football Conference 2019* (London), 11.

Rodríguez-Pérez, R., and Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *J. Comput. Aided Mol. Des*. 34, 1013–1026. doi: 10.1007/s10822-020-00314-0

Roth, A. E., and Thomson, W. (1988). *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press. Available online at: https://www.hbs.edu/faculty/Pages/item.aspx?num=6946

Rowlinson, A. (2020). *Football shot quality* (Master thesis), Aalto University, Espoo, Finland.

Ruiz, H., Power, P., Wei, X., and Lucey, P. (2017). ""The Leicester City Fairytale?": utilizing new soccer analytics tools to compare performance in the 15/16 & 16/17 EPL seasons," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS), 1991–2000. doi: 10.1145/3097983.3098121

Schulze, E., Mendes, B., Maurício, N., Furtado, B., Cesário, N., Carriço, S., et al. (2018). Effects of positional variables on shooting outcome in elite football. *Sci. Med. Football* 2, 93–100. doi: 10.1080/24733938.2017.1383628

Spearman, W. (2018). "Beyond expected goals," in *MIT Sloan Sports Analytics Conference* (Boston, MA), 1–17.

Spearman, W., Basye, A., Dick, G., Hotovy, R., and Pop, P. (2017). "Physics-based modeling of pass probabilities in soccer," in *MIT Sloan Sports Analytics Conferece* (Boston, MA), 1–14.

Stein, M., Häußler, J., Jäckle, D., Janetzko, H., Schreck, T., and Keim, D. A. (2015). Visual soccer analytics: understanding the characteristics of collective team movement based on feature-driven analysis and abstraction. *ISPRS Int. J. Geoinform*. 4, 2159–2184. doi: 10.3390/ijgi4042159

Taberner, M., O'Keefe, J., Flower, D., Phillips, J., Close, G., Cohen, D. D., et al. (2019). Interchangeability of position tracking technologies; can we merge the data? *Sci. Med. Football* 4, 76–81. doi: 10.1080/24733938.2019.1634279

Tenga, A., Ronglan, L. T., and Bahr, R. (2010). Measuring the effectiveness of offensive match-play in professional soccer. *Eur. J. Sport Sci*. 10, 269–277. doi: 10.1080/17461390903515170

Tian, C., De Silva, V., Caine, M., and Swanson, S. (2020). Use of machine learning to automate the identification of basketball strategies using whole team player tracking data. *Appl. Sci*. 10:24. doi: 10.3390/app10010024

Wang, Y. (2019). A Xgboost risk model via feature selection and bayesian hyper-parameter optimization. *arXiv*. doi: 10.5121/ijdms.2019.11101

Wei, X., Lucey, P., Morgan, S., Reid, M., and Sridharan, S. (2016). "The Thin Edge of the Wedge: accurately predicting shot outcomes in tennis using style and context priors," in *MIT Sloan Sports Analytics Conference* (Boston, MA), 1–11. doi: 10.1145/2783258.2788598

Conflict of Interest: GA was employed by the company Sportec Solutions AG and PB was employed by the company DFB-Akademie (Deutscher Fußball-Bund e.V).

B Appendix—Study II: Expected Passes: Determining the Difficulty of a Pass in Football (Soccer) Using Spatio-Temporal Data

In the following, Anzer and Bauer (2022) is reproduced with permission from Springer.

Data Mining and Knowledge Discovery
https://doi.org/10.1007/s10618-021-00810-3

Expected passes

Determining the difficulty of a pass in football (soccer) using spatio-temporal data

Gabriel Anzer[1,2] · **Pascal Bauer**[2,3]

Received: 18 April 2021 / Accepted: 21 October 2021

Abstract
Passes are by far football's (soccer) most frequent event, yet surprisingly little meaningful research has been devoted to quantify them. With the increase in availability of so-called positional data, describing the positioning of players and ball at every moment of the game, our work aims to determine the difficulty of every pass by calculating its success probability based on its surrounding circumstances. As most experts will agree, not all passes are of equal difficulty, however, most traditional metrics count them as such. With our work we can quantify how well players can execute passes, assess their risk profile, and even compute completion probabilities for hypothetical passes by combining physical and machine learning models. Our model uses the first 0.4 seconds of a ball trajectory and the movement vectors of all players to predict the intended target of a pass with an accuracy of 93.0% for successful and 72.0% for unsuccessful passes much higher than any previously published work. Our extreme gradient boosting model can then quantify the likelihood of a successful pass completion towards the identified target with an area under the curve (AUC) of 93.4%. Finally, we discuss several potential applications, like player scouting or evaluating pass decisions.

Responsible editor: Albrecht Zimmermann.

Gabriel Anzer
gabriel.anzer@sportec-solutions.de

Pascal Bauer
pascal.bauer@dfb.de

1 Sportec Solutions AG, subsidiary of the Deutsche Fußball Liga (DFL), Munich, Germany

2 Institute of Sports Science, Department of Sport Psychology and Research Methods, University of Tübingen, Tubingen, Germany

3 DFB-Akademie, Deutscher Fußball-Bund e.V. (DFB), Frankfurt, Germany

Published online: 04 January 2022

Keywords Expected passes · Sports analytics · Football · Soccer · Positional data · Event data · Machine learning

1 Introduction

Passes are a crucial part of modern football (soccer) matches. However, traditionally player's passing performance is quantified using a binary pass completion metric. This means that — regardless of the quality or difficulty of a pass — completed passes are rewarded a "1/+", and incomplete passes rewarded a "0/−". A player's pass completion rate is thus calculated as the ratio of completed passes to total passes played. This ratio neglects to take the complexity or the reward of a pass into consideration. Nevertheless, pass completion rates are regularly used as performance indicators on team and player levels in literature (Bradley et al. 2013; Król et al. 2017) and in the daily business of professional football teams. Whenever a player is in possession of the ball, they may choose to pass to any of their teammates — and each option comes with a unique set of risks and rewards. This decision can only be evaluated by considering both the risk and the reward of each option.

The relevance of passes in football was investigated with annotational analysis of passing patterns (Reep and Benjamin 1968) and through experimental studies analyzing influencing factors for passes (Williams 2000). The increasing availability of granular football data unlocked new avenues for the analysis of passes. Event data, following the idea of Reep and Benjamin (1968), describes a log of all on-the-ball-actions (e.g. shots, passes, tackles) and are systematically acquired in most professional football leagues. Several studies used this event data to analyze passes on a much larger scale than previous experimental studies would allow. For example, Szczepański et al. (2016) used 253, 090 open-play passes and McHale and Relton (2018) analyzed 960, 000 events including passes. While manually collected event data provides relevant information about one or two players involved in the current ball action, recent improvements in computer vision allow to accurately track the positions of all 22 players and the ball at any time of the match. This type of data is typically referred to as tracking, positional or movement data (Stein et al. 2017; Andrienko et al. 2019; Bauer and Anzer 2021; Anzer et al. 2021).

While some studies quantified the reward of a pass using event data only (Brooks et al. 2016; Power et al. 2017; Bransen et al. 2019), combining the manually tagged event data with the automatically acquired positional data allows for a more granular analysis of the reward of a pass. Several studies addressed this reward-quantification of passes in different ways (Rein et al. 2017; Chawla et al. 2017; Goes et al. 2019; Gómez-Jordana et al. 2019; Steiner et al. 2019; Anzer and Bauer 2021), but they typically measure how much a pass would increase the chance of scoring if successful.

As highlighted in Power et al. (2017) and Goes et al. (2021) the quantification of pass decisions has two dimensions: The reward of a pass, as discussed above, and the difficulty of the pass, usually measured in the completion probability. This risk of a pass is often referred to as expected pass (xPass) values in literature (Spearman et al. 2017; Power et al. 2017; Fernández et al. 2020; Arbués-Sangüesa et al. 2020; Alguacil et al. 2020; Stöckl et al. 2021). An xPass model tries to estimate the probability of

a given pass being successfully completed to a teammate, based on various factors describing the pass — usually derived from positional and/or event data. Furthermore, Li et al. (2019) and Vercruyssen et al. (2016) have explored the target identification of passes as a standalone problem. To quantify the risk, Power et al. (2017) built a logistic regressor for 571, 278 passes, whereas Spearman et al. (2017) modelled 10, 875 passes as Bernoulli trials. In order to retrieve the missing information regarding the intended receiver of a pass (at the moment the pass was played), they modelled both ball and player trajectory based on physical simulations first. This allows them to calculate an xPass value, as the predicted probability of a pass being completed, at the moment when the pass is played. The physics-based models were slightly improved by Alguacil et al. (2020) through taking friction for ground-passes into consideration.

Stöckl et al. (2021) later slightly improved the accuracy of the xPass model by using Graph Neural Networks (Battaglia et al. 2018) to overcome both the feature extraction and the ordering-problem of using spatio-temporal tracking data in a dynamic sport like football. Arbués-Sangüesa et al. (2020) showed that a player's body orientation (typically not included in off-the-shelf tracking data) has a significant influence on pass completion probabilities as well. Several further extensions, built on top of xPass models, exist in the literature: Fernandez et al. (2018) and Spearman et al. (2017) include xPass models as central ingredients for computing their expected possession values, and Hubáček et al. (2018) use it to try to predict which pass will be played next in any given situation.

But overall the literature is lacking a thoroughly described method of synchronizing pass events with tracking data, a highly accurate intended receiver estimation and a properly (manually) evaluated xPass model. Our work fills this gap, while keeping the individual modules completely separated and introduces novel concepts, like blocking probabilities.

Our goal is to train a machine learning model on the binary classification, of whether a pass will be successful or not using all the information available at the time of the pass. While the data set (described in Sect. 2) is extremely detailed, it is missing one piece of essential information, namely the targeted recipient of unsuccessful passes. Our work consists of the following four steps:

(1) **Synchronization of pass events:** We synchronize both the location and the exact timing of pass events from manually annotated event data with automatically acquired tracking data (similar to the method introduced for shot events in Anzer and Bauer (2021)). Details of the approach can be found in the Appendix A.
(2) **Estimate the intended receiver:** First, we use a state-of-the-art movement model to derive the potential positions of all players within a certain time window according Brefeld et al. (2019) (see Sect. 3.2), and second, combine this with a physics-based ballistic ball trajectory model as described in Spearman et al. (2017) (see Sect. 3.1). Given the ball positions within the first 0.4 seconds of a pass, this model uses the results from aerodynamic investigations (Asai et al. 2007; Oggiano and Satran 2010) to predict the trajectory of the ball. The combination of both steps provides us an accurate prediction of the intended receiver for unsuccessful passes (see Sect. 3.3).

(3) **Pass probability:** In Sect. 4, we train a machine learning model to estimate the probability of a pass (that was not blocked immediately) based on the information derived from (2) and from expert-based features describing the pass.
(4) **Blocking model:** In order to get unbiased estimates for the probabilities of all passes, we further calculate the likelihood that a pass is blocked (see Sect. 5). This is also approached using a supervised machine learning model with hand-crafted features.

Finally, we can compute the probability of any potential pass being completed. By combining and slightly improving previor work, we exceed the accuracy of all previously presented results for the prediction of the pass receiver as well as the classification of played passes being successful or not.

2 Data and definitions

In the official match-data catalog of the German Bundesliga,[1] a pass is defined the attempt to switch ball control from one player to a teammate. For each pass detected, trained operators annotate a variety of sub-attributes describing the pass in detail. Among others, they annotate who played and (in case of a successful pass) received the ball, whether it was a *high* or a *low played* pass, as well as, whether the pass was played over a *short*, *medium*, or *long* distance. Of course, all of the sub-attributes underlay detailed definitions, defining high passes as passes played above knee height and setting thresholds to differentiate short passes ($< 10\,\text{m}$), passes of medium length ($10 - 30\,\text{m}$) and long passes ($> 30\,\text{m}$). All attributes are collected for both successful and intercepted passes, meaning that the intended height and the intended length is estimated by the human operator in case of intercepted passes. While this manually acquired event data underlays strict quality checks, especially for incomplete passes it can be quite subjective.

More objective and more granular information can be found in the positional data, capturing the positions of all 22 players and the ball at 25 Hz. In each Bundesliga-stadium, up to 20 installed HD-cameras record any action on the pitch and serve as input for computer vision algorithms estimating the 2D-positions of all players as well as the 3D positions of the ball. In the Bundesliga, data from Chyronhego's optical Tracab system is collected.[2] Several studies evaluated the accuracy of this data (Linke et al. 2020, 2018).

We excluded fair-play passes, in which a player voluntarily relinquishes his team's ball-control, passes that accidentally end up with a teammate who was not the intended target, as well as throw-ins from our analysis. This information is captured within the event data and can thus be simply filtered out for our investigation. We end up with positional and event data of 840, 386 passes from 918 Bundesliga games from the 2017/2018, 2018/2019 and 2019/2020 seasons, with an average completion rate of 85.2%.

[1] https://s.bundesliga.com/assets/doc/10000/2189_original.pdf (accessed March 27, 2021).
[2] https://tracab.com/products/tracab-technologies/tracab-optical/ (accessed April 14, 2021).

The necessity to synchronize the two independently acquired data-sources, is detailed in the literature (Anzer and Bauer 2021; Spearman et al. 2017). We apply a slightly modified methodology to synchronize pass events as described for shots in Anzer and Bauer (2021). The outcome of the synchronization is manually evaluated in Sect. 6 as a part of the xPass evaluation, finding that 99.1% of all pass events are identified correctly. Further details on the synchronization methodology as well as a more thorough validation study are provided in the Appendix A.

For reproduction, Pettersen et al. (2014) present a publicly available set of positional data, and open source event data can be found in Pappalardo et al. (2019).[3]

3 Estimating the target

While the receivers of successful passes are included in the event data, the intended target of unsuccessful passes is missing. This is a crucial point of information necessary for determining the difficulty of a pass, since otherwise only the surrounding circumstances of the passer could be taken into consideration. Therefore, we need to determine who the intended receiver was, to later extract features for both successful and unsuccessful passes. For that purpose we first use a physics-based approach to estimate the ball trajectory based on the first couple of frames after the pass is played (Sect. 3.1). Second, we compute a movement model, to estimate the area on the pitch, players could potentially reach in the next n frames, based on their movement direction and velocity (Sect. 3.2). Third, by combining both the estimated ball trajectory and the reachable area, we identify the teammate most likely to reach the ball first as the intended recipient of the pass (Sect. 3.3). Furthermore, we discuss (Sect. 3.4) how this can be used to derive physics-based features describing the difficulty of a pass.

3.1 Modelling the ball trajectory

Knowing that a football adheres to physical laws, we can use these laws to determine the path a ball will travel on (until it is touched again) based on its initial direction and velocity. As suggested in Spearman et al. (2017), we use the first 10 frames (equivalent to 0.4 seconds) after a pass was played, to receive a stable estimate of its initial direction (x, y, z) and exit velocity. Therefore, we exclude all passes blocked within the first 0.4 seconds, since we are unable to determine the necessary starting values for them reliably. Using a physical trajectory model, including gravity, air drag and rolling drag (with the simplification that as soon as a ball lands, it is grounded), we can estimate for every following frame, where the ball will be. As presented in Spearman et al. (2017), the trajectory of the ball is consequently modelled as:

$$\ddot{r} = -g\hat{z} - \frac{1}{2m}\rho C_D A \dot{r}\dot{r}$$

[3] Other (non-scientific) open-source data sets can be accessed from Skillcorner (https://github.com/SkillCorner/opendata), Metrica sports (https://github.com/metrica-sports/sample-data) or Statsbomb (https://github.com/statsbomb/open-data).

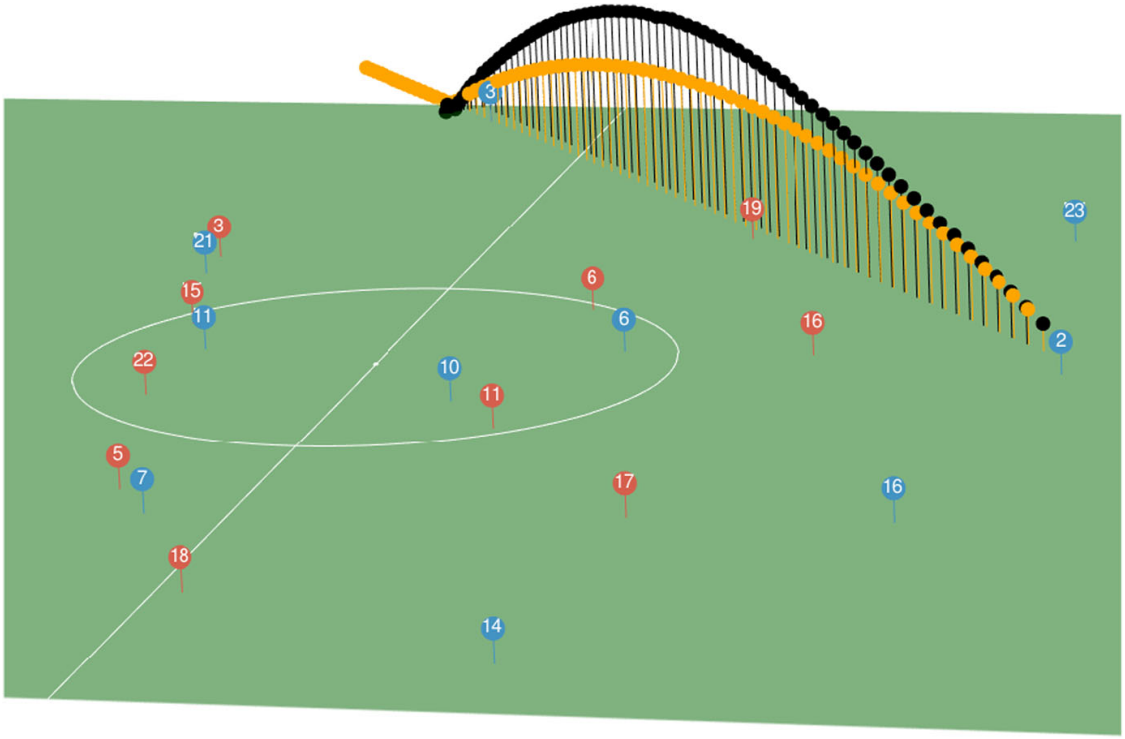

Fig. 1 Estimated ball trajectory (yellow dots) compared to the measured data-points from the tracking data. The video footage of the pass can be found here: https://dfb-my.sharepoint.com/:v:/g/personal/pascal_bauer_dfb_de/EUJra9f8i6BCl2-mzpuHtacBvgHNx7cnCH9P8y5taozDnQ?e=nIhgkC (Color figure online)

All physical values are set to the respective standard.[4] The Bundesliga-ball has a weight of 0.4 kg (m) and a cross-sectional area of 0.038 m^2 Further background information regarding the aerodynamics of balls in football can be found in Asai et al. (2007); Oggiano and Satran (2010).

Figure 1 shows both the observed ball path from the tracking data (black dots) and the estimated ball trajectory from our physics-based model (yellow dots) for a played pass. The physics-based model yields a smooth and realistic ball path, while the observed ball path shows some jumps (e.g. around the highest point of the trajectory), frequently present when tracking small fast moving objects from large distances. Furthermore, it can be used to model where a ball might have ended up, had it not been deflected or intercepted.

3.2 Movement model

The movement model predicts what area of the pitch a player can reach within a defined time-window. Again, Spearman et al. (2017) presented a physics-based approach for this. Additionally, they gave a first outlook towards data-driven movement models which were later built upon by several studies. The positions a player can reach within a certain time frame depend on his current speed and direction (Brefeld et al. 2019; Fernandez et al. 2018). With these assumptions we use movement data from three

[4] Gravitational force $g = 9.8 \frac{m}{s^2}$; air density $\rho = 1.22 \frac{kg}{m^3}$; drag coefficient $C_D = 0.25$.

seasons of Bundesliga data. First, we transform the data, so that all players are traveling in the same direction. Next we compute the convex hull of all observed locations players traveling in a certain speed interval were able to reach after n-frames. Due to our large data set of tracking data, we are able to use much smaller speed intervals (of 0.5 km/h) compared to Brefeld et al. (2019). With this information we fit our movement model to estimate the center of the circle and its diameter, based on speed and time.

Now we can calculate for any player on the pitch what area they could theoretically cover in the next seconds based on their movement vector. This is displayed for some players (#18/#22 red team; #7 blue team) in Fig. 3. Each circle represents the area the respective player can reach within 0.5, 1, 1.5 and 2 seconds.

3.3 Target estimation

To estimate the intended target of a given pass, we combine the physics-based ball trajectory model with the data-driven player movement model. To incorporate the ball height, we additionally assume that a ball is only reachable below a height of 1.5 m. This threshold was obtained by optimizing for the accuracy of the intended receiver prediction for successful passes on the training data set introduced in Sect. 4. Thus, we can calculate which team member of the passer could theoretically be the first to reach the pass, and declare them as the intended receiver.

We are able to predict the correct player for successfully completed passes with an accuracy of 93.1%. For unsuccessful passes, we conducted an evaluation study, described in Sect. 6, showing that we are able to predict the estimated target with an accuracy of 72.0%.

3.4 Physics-based passing features

We can quantify what direction and how fast a pass would need to be played to arrive at the target receiver. For that purpose we compute hypothetical passes by varying the initial starting parameters of a pass, i.e. initial velocity and initial direction of a pass. Combined with the movement model, we can determine if the target player is still the most likely player to receive each hypothetical pass. This step is done by performing a grid based search varying the velocity and the direction of the pass noting for every (reasonable) combination[5] if the intended target is likely the first player to potentially reach the pass. From this we can compute the direction window, defined as the width of the reachable angles. The speed window is defined as the difference between the maximal relative increase and decrease of a baseline exit velocity, with which hypothetical passes would still be reachable by the intended receiver. An example of a direction window is indicated in Fig. 2. Hypothetical passes, with slightly modified x-/y-directions (assuming an ideal speed and launch-angle) that could be received by Emre Can (#3 of the blue team) according to our model are displayed as grey lines. The total width of theses potential pass angles amounts to 21 degrees in this example.

[5] The velocity range is between $[-100\%, 100\%]$ of the average pass speed, and the direction window range is between $[-25°, 25°]$ of the direct connection line between the passer and the intended target.

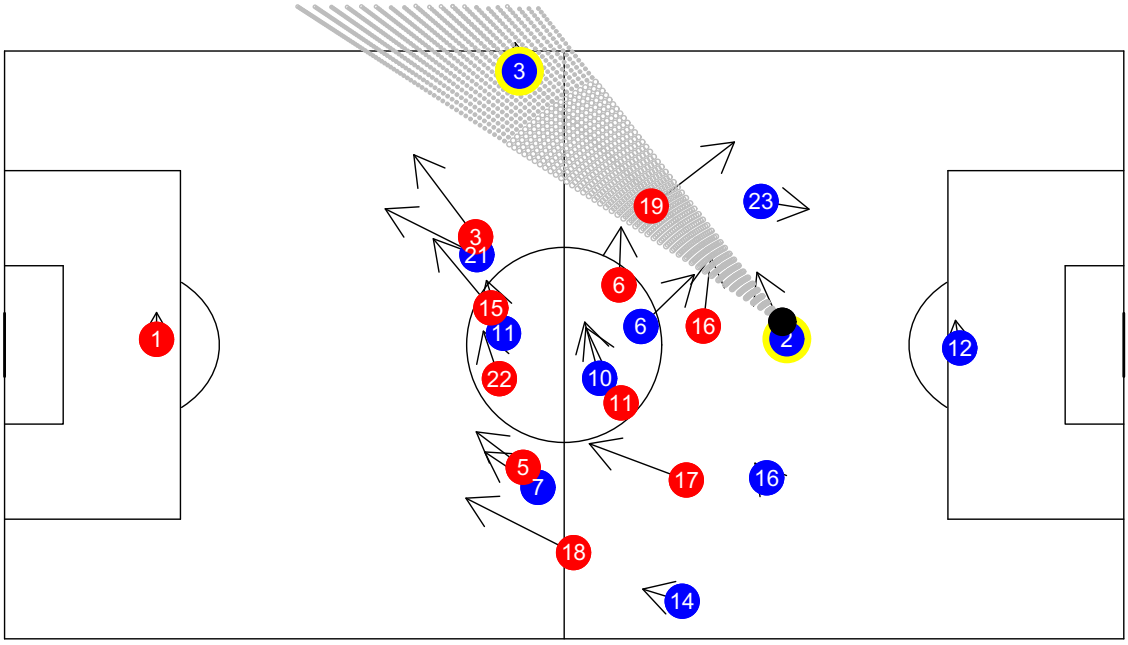

Fig. 2 Visualization of potential pass angles reaching the intended target. Players' movement vectors are displayed as arrows. The passing player (Robin Koch, #2 of the blue team) as well as the receiving player (Emre Can, #3 of the blue team) are highlighted in yellow. The same pass is displayed as in Fig. 1 and in this video: https://dfb-my.sharepoint.com/:v:/g/personal/pascal_bauer_dfb_de/EUJra9f8i6BCl2-mzpuHtacBvgHNx7cnCH9P8y5taozDnQ?e=nIhgkC (Color figure online)

Table 1 Speed scalar window width (as percentage) and direction window (in degrees) of passes

Passe Outcome	Speed window	Direction window	Number of Passes
Successful	0.99 (±0.48)	32.7 (±16.3)	291, 700
Unsuccessful	0.26 (±0.36)	11.6 (±12.6)	56, 570

The observed standard deviation is denoted in parentheses

Table 1 shows that unsuccessful passes have both a much narrower window of potential directions (in degrees) as well as in speed values (in percentage difference compared to a baseline speed value). This aligns with expert opinions that the less accurate a pass needs to be played, the easier it is and the higher chance that it will be completed.

The interplay of the target prediction using the physics-based ball trajectory model and the data-driven movement model is displayed in Fig. 3. In this situation Robin Koch (#2 of the blue team) plays a diagonal ball to his teammate Julian Draxler (#7 of the blue team). The curvature of the ball trajectory (yellow dots) shows the trajectory of the played diagonal pass from the mid-point (player #2) to the left attacker of the blue team (player #7).[6] Due to its height, the ball can only be reached towards the end of the projected trajectory. Matching possible intersections of the trajectory after n frames with each teammate's reachable area after the same time period, reveals, that the first player to possibly reach the ball is the attacker on the left wing after 2.48

[6] The video of the pass can be found here: https://dfb-my.sharepoint.com/:v:/g/personal/pascal_bauer_dfb_de/EWIWkaF8Gp5CjPCdQRs5KXsB6Rt0LKHoKomXUFogNsR2Wg?e=u6EziX..

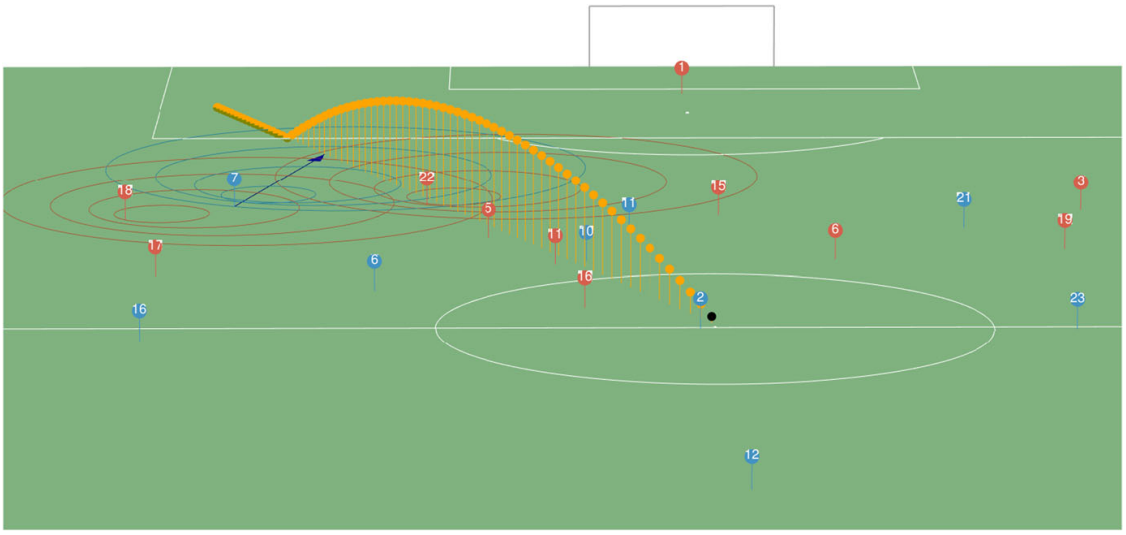

Fig. 3 Estimated target of a pass with ball-trajectory and movement models. The combination of the estimated ball trajectory (yellow dots) and the player movement model (blue and red circles) predict Julian Draxler (#7 of the blue team) reaching the ball first, with the arrow indicating the first point where he could potentially intercept the pass. The respective video sequence can be viewed here: https://dfb-my.sharepoint.com/:v:/g/personal/pascal_bauer_dfb_de/EWIWkaF8Gp5CjPCdQRs5KXsB6Rt0LKHoKomXUFogNsR2Wg?e=u6EziX (Color figure online)

seconds. The point where he could first reach the ball is indicated by the black arrow, but this does not mean that it is always the optimal strategy to do so.

4 Pass probability estimation

For successful passes we know the recipient of a pass from the event data. With the approach described in Sect. 3, we can identify the intended target of unsuccessful passes (as long as it is not blocked). This allows us to compute tailored features influencing the pass difficulty and train supervised machine learning models estimating pass completion probabilities. We build on the features describing passes presented recently (Power et al. 2017; Spearman et al. 2017; Mchale and Lukasz 2014; Hubáček et al. 2018). Table 2 shows an overview of all features we compute for every pass. As in Spearman et al. (2017), we are interested to train a predictive model, i.e. all features must be available at the time of a pass. This will allow us later to compute hypothetical pass probabilities, or evaluate if a player is over-/under-performing. The first eight features describe the pass origin and the situation around the passer, e.g. where on the pitch the passer is located, how far the next opponent is away from them and how much pressure they are receiving according the pressure-model introduced in Andrienko et al. (2017). The next set of features (rows 9 – 14, Table 2) describe the receiver and their surrounding environment. The third block of features (rows 15 – 21) describe further context information around the pass itself taking full advantage of the positional data. The manual collected features *height* and *distance* (rows 22 and 23) are described in Sect. 2. The last two features in Table 2 (rows 24 and 25) are calculated based on the logic described in Sect. 3.

Table 2 Hand-crafted features used to train the xPass-model

	Feature	Description	B
1	Location pass	x- and y- coordinate of the pass.	x
2	Distance sideline	Distance between the location of the pass and the closest sideline.	x
3	Distance goal	Distance between the location of the pass and the opposing goal.	x
4	Distance opponent	Distance between the location of the pass and the nearest opposing player at the time of the pass.	x
5	Speed passer	Speed of the passing player at the time of the pass.	x
6	Ball height	Ball height at the time of the pass.	
7	Opponents closer to goal	Number of opponents closer to their own goal than the passer at the time of the pass.	x
8	Pressure on passer	Pressure exerted on the passer according Andrienko et al. (2017).	
9	Location receiver	x- and y-coordinate of pass receiver at the time of the pass.	
10	Distance receiver sideline	Distance between the receiving player and the closest sideline at the time of the pass.	
11	Distance receiver goal	Distance between the receiving player and the opposing goal at the time of the pass.	
12	Distance receiver opponent	Distance between the receiving player and the nearest opposing player at the time of the pass.	
13	Opponents closer to goal receiver	Number of opponents closer to their own goal than the receiver at the time of the pass.	
14	Speed receiver	Speed of the receiving player at the time of the pass.	
15	Possession phase	The time the passer was in ball possession before attempting the pass.	x
16	Bypassed opponents	Opposing players that would be bypassed with the pass (Steiner et al. 2019).	
17	Angle	Directional angle of the pass compared to the playing direction (i.e. 0 is directly towards the opposing goal line, and 180 would be backwards towards the own goal line).	
18	Opponents in path	Number of opposing players in the passing path. The path being defined as a corridor between the pass location and the receiver location with a width of 10 meters.	
19	Nearest defender pass line	Nearest defender to a straight line connecting the pass location and the receiver location.	
20	Distance pass	Distance between passer and receiver when the pass is played.	
21	Dead ball	Binary information whether the pass originated from a set-piece (e.g. freekick, goalkick, ...) or not.	
22	Height	Manually annotated binary feature describing the intended ball height included in the event data.	x

Table 2 continued

	Feature	Description	B
23	Distance	Manually annotated intended pass length (short, medium, long) included in the event data.	x
24	Speed window	Mean speed window as defined in Sect. 3.4.	
25	Direction window	Mean direction window as defined in Sect. 3.4.	

Column "B" notes all features that are also used for the blocking model described in Sect. 5

For our model training we use 840, 386 passes from 918 Bundesliga games. We split the data into training (504, 232 passes), validation (168, 077 passes) and test set (168, 077 passes) and use different subsets of features from Table 2 to train various supervised machine learning models (logistic regression, extreme gradient boosting, random forest). For each feature set the best performing models on the test set were extreme gradient boosting (XGBoost) models (Chen et al. 2016). For all XGBoost models we applied Bayesian hyperparameter optimization on the validation set (Nazareth 2004). The accuracy metrics of the XGBoost models for the different feature sets are displayed in Table 3. Since precision, recall and F_1-score are not ideally suited to evaluate probabilities in an imbalanced data set, we focus on the metric area under the receiving operator curve (AUC), mean square error (MSE) and the Brier skill score (BSS, Brier (1950)). All relevant metrics indicate that the model using the full set of features (line 1, Table 3) provides the best results.

We implemented three simple baseline models for comparison: First, we trained a model using only the features that can be derived from event data (row 4, Table 3). In order to evaluate the necessity of identifying the intended receiver, we trained a receiver-agnostic model (row 5, Table 3). Lastly, row 6 in Table 3 presents a trivial baseline model, that assumes a constant pass success probability of 85.2% (the completion rate in the training data set). Table 3 shows, that both using positional data and the estimation of the targeted receiver (see Sect. 3) significantly improve the prediction accuracy.

The final hyper-parameter configuration for the model using all features is provided in the Appendix B (Table 8). In the complete model, according to the overall SHAP-values (Lundberg et al. 2017), the possible speed window in which the pass could be completed has by far the highest influence on the success prediction followed by the distance of the closest opponent to the receiver. Purely absolute position related features like the x-/y-coordinates of the pass origin and the receiver position, as well as the distance to the sideline/goal exhibited the lowest influence on the prediction. More details regarding the feature importance and SHAP-values can be found in Appendix B.

5 Blocking model

In order to get a reasonably reliable target identification in the previous sections we focused on passes that were not blocked immediately. However, this inflates the likeli-

Table 3 Outcome of the XGBoost models on the test set using different feature sets

	Model	Precision	Recall	F_1-score	AUC	MSE	BSS
1	All features included	0.944	0.958	0.951	0.955	0.059	0.528
2	Speed and direction window only	0.887	0.963	0.923	0.873	0.091	0.270
3	Speed and direction window excluded	0.923	0.953	0.937	0.920	0.075	0.397
4	Event-data only	0.894	0.960	0.926	0.832	0.095	0.241
5	Receiver Agnostic	0.898	0.957	0.927	0.864	0.091	0.027
6	Average pass probability	–	1	–	0.5	0.126	0

Table 4 Accuracy metrics for the blocking model on the test set

	Model	Precision	Recall	F_1-score	AUC	MSE	BSS
1	Gradient Boosting	0.643	0.031	0.060	0.753	0.059	0.084
2	Naive Blocking Model	–	0.000	–	0.500	0.065	0.000

hood of a pass being completed, since it ignores that about 3.12% of the passes in our data set are blocked. Therefore, we need to adjust our conditional passing probabilities by the likelihood that it is not blocked: For all successful passes A and all blocked passes B, the probability of any non-blocked pass being completed, $P(A \cap \bar{B})$, can be computed as follows:

$$P(A \cap \bar{B}) = P(A|\bar{B}) * P(\bar{B})$$

The probability of a pass success — provided that it is not blocked — $P(A \cap \bar{B})$, is calculated in Sect. 4. However, at the time of the pass we do not know whether it will be blocked or not. Consequently, to get an unbiased pass completion probability, we need to calculate the probability that it is blocked, $P(B)$. Rather then simply discounting all passes with the average rate in which passes passes are blocked (3.12%), we determine the likelihood of each pass being blocked individually, based on some of the features described earlier. We define a blocked pass, as a pass where an opposing player touches it within the first 0.4 seconds. A problem is that the exact initial direction of blocked passes cannot be accurately derived from tracking data. Therefore, we simply assume that if a pass was blocked the intended direction was towards the point where the opponent touched it.

Consequently, a pass can only be blocked (according to our definition) if an opposing player is located in the passing direction and could reach a pass within 0.4 seconds — assuming the average speed of a pass this roughly translates to a 5 meter radius of the passing origin. In all cases where this criteria is not fulfilled, we set the probability of the pass being blocked to zero. For the remaining passes (6% of them were blocked) we trained a XGBoost model to estimate the likelihood that a pass will be blocked. We used several features introduced in Table 2 (marked in column B) like the proximity of the nearest opposing player within the passing direction ($+/-$ 90 degrees), the location on the pitch (i.e. x/y-coordinates), the time of possession and the intended distance. The final blocking model was trained on 312, 413 passes (9, 372 blocked) with a split into 60% training, 20% validation and 20% test data. The outcome of the prediction is presented in Table 4. For comparison it includes a naive model using the average block probability (6.5%) as a baseline (row 2). The final hyperparameters of the blocking model can be found in Appendix B (Table 8).

6 Manual validation

We described statistical evaluations for each component of the entire approach in the respective chapters (i.e. Tables 3, 4). However, to further validate our results, we performed three separate expert-based validation studies of the following components:

(1) Synchronization of passing events with positional data
(2) Detection of the intended receiver for unsuccessful passes
(3) Outcome of the final xPass model

For each of the validation studies three different football experts looked at 3,600 passes from 10 different games, with one game shared amongst all three, to gather the inter-rater reliability.

In order to evaluate (1) in the context of passes, the football experts were presented with identified time stamps, and they were tasked to annotate, whether the timestamps are correct. Overall they identified 99.1% of the timestamps as correct and had a pairwise inter-rater reliability of 99.3% However, since this approach is very binary (and potentially biased), we conducted a separate thorough evaluation study of the pass synchronization, described in Appendix A.

Since we can only systematically assess the accuracy of the intended receiver identification for successful passes (see Sect. 2), in (2) the subjects were tasked to identify the intended recipient of unsuccessful passes. Of the 1,307 unsuccessful passes, our prediction agrees with the human labels in 72.0% of the cases and the inter-rater reliability is 96.2%.

The third and most relevant validation study evaluates, how well we can judge the difficulty of a pass (3). This is especially relevant because our final xPass values result from a combination of different machine learning models, each with their own inaccuracies. Therefore, the final outcome was evaluated manually by football experts. Estimating pass probabilities is a very challenging task for humans (even for football experts). To circumvent this issue, we provide experts with sets of two passes and let them assess which of the two is more difficult. Comparing passes with very similar xPass values is likely not a very reliable ground truth, and comparing passes with large xPass differences should be a trivial task with a high accordance between experts and our model. Therefore, in order to minimize the human-labeling effort, we group pairs of passes in three different categories based on their absolute xPass differences:

- Small difficulty difference ($< 10\%$),
- Medium difficulty difference ($10 — 30\%$),
- Large difficulty difference ($> 30\%$).

Per match we select 300 pass comparisons and the majority of them (90%) in the second category, 7% in the first category and 3% in the last category.[7]

Limited by the inter-labeler accordance, especially in critical situations, Table 5 shows that our model achieves satisfactory results. To investigate how much the addition of the blocking model helps the predictions, we further compute the accuracy of the model without a superimposed blocking model. This simpler model has a lower accuracy of 71.1% over the entire data set.

[7] The pass comparisons were randomly selected with the above described distribution, so the final numbers are subject to randomness and slightly deviate from the target distribution.

Table 5 The average pairwise accuracy are depicted and the number of pairs in a given subset are in the brackets

	Evaluation	Labeler accordance	xPass Model
1	Small difficulty difference	0.786 (70)	0.640 (286)
2	Medium difficulty difference	0.812 (787)	0.715 (3175)
3	Large difficulty difference	1.000 (30)	0.983 (118)
4	All	0.786 (887)	0.718 (3,579)

7 Discussion

A general limitation of our approach is its sensitivity to positional data accuracy. While the quality of tracking data has been increasing continuously over the past decade, the accuracy of ball tracking has not been properly validated in the literature yet (Anzer and Bauer 2021). The spatio-temporal synchronization of positional and event data — typically acquired through independent systems — presents crucial improvement for the analysis of passes. By training various models on different feature sets, we show how much each additional set increases the model's quality. Spearman et al. (2017) also pointed out the necessity of this synchronization step, but did not provide any details, nor an evaluation of their implemented approach. By adopting the methodology from Anzer and Bauer (2021) (synchronization of shot events) to passes, we use a reproducible approach (independent of the event-/tracking-data provider) and evaluate its accuracy manually in two independent experiments (Sect. 6 (1) and Appendix A).

Both, the player-movement model (Sect. 3.2) and the ball trajectory model (Sect. 3.1) draw heavily from previously published work (Brefeld et al. 2019; Spearman et al. 2017). We combine both to estimate the target of a pass and made only minor adjustments in order to improve the prediction accuracy on our data set. One thing we found regarding the movement model, is that the tighter the speed interval, the more the shape of the resulting hull is circular instead of elliptical, contrasting the findings of Brefeld et al. (2019), that finds oval shapes while using broader speed ranges. This could imply that movement ranges for particular initial speeds are circular, but when using a wide range of initial speeds, the total observed range is a combination of the movement circles along the movement direction, thus taking an elliptical shape.

Similar, as in Spearman et al. (2017) we ignore wind, rotation of the ball, and the Magnus force in the ball trajectory estimation. Our approach struggles to identify the intended receiver, when the underlying pass attempt fails completely. Fortunately, this case happens very rarely in the highest professional environments.

Implementing a separate blocking model guarantees that we have an unbiased estimation of pass probabilities. Furthermore, the manual validation (Sect. 6) shows, that it also more accurately coincides with expert assessments regarding the pass difficulty. The relatively low predictive power of the blocking model is likely caused by the nature and quality of the tracking data. The players' x/y-coordinates merely describe their center of gravity and, especially at the moment of the pass, centimeters may decide whether a pass is blocked or not. Therefore, as long as so-called limb-

tracking (recordings of players' entire bodies) does not become more widely available, it will remain hard to estimate if an opponent can extend their leg to block a pass.

Probabilistic metrics (e.g. expected goals, xPass) are hard to manually evaluate, since even experts cannot estimate a ground truth percentage reliably. For this reason we developed an evaluation study design delivering a useful ground truth while maintaining a high inter-labeler reliability. In previous research the quality of pass difficulty models was purely measured by the accuracy of the binary pass success classification. Our work goes one step further through a manual validation study with football experts, allowing us to also evaluate (1) the synchronisation of positional and event data, (2) the receiver estimation for unsuccessful passes, and (3) the pass difficulty. While in (2) we achieved, a reasonable accuracy of 72.0%, the experts showed a very high inter-rater reliability of 96.2%. This can in part be explained, by the fact that they were given video sequences of the passes extending far further than the 0.4 seconds our estimation uses. When examining the cases with the differences, we found that this is mostly caused by long balls (e.g. goal kicks, half-field crosses) where multiple players could be the target, but only one of them gets involved in an aerial duel. The human observers then chose the teammate that lost the aerial duel. But for the purpose of our work, in these cases the possible target players are very close to each other, meaning that the feature calculation and, therefore, their xPass values are very similar. Apart from that erroneous ball tracking data can lead to wrong target predictions (e.g. when the ball has a sharp cut, often called "elbow", in its trajectory after 0.4 seconds, without being touched). The much lower accuracy of the intended target identification achieved by Li et al. (2019) (for successful passes: 27.87%) and Vercruyssen et al. (2016) (for successful passes: 50.00%; for all passes: 41.00%) shows how difficult of a problem this generally is. However, this comparison is not completely fair, since they only consider information at the time of the pass, while we use the first 0.4 seconds after the pass as well.

We are able to increase the accuracy of successful pass estimation on a team level (Spearman et al. (2017): 80.5%, our approach: 91.5%) as well as for the task of predicting the receiving player (Spearman et al. (2017): 67.9%, our approach 89.9%). Power et al. (2017) presents a pass prediction with a root mean square error (RMSE) of 0.2483 which is slightly improved by our approach (RMSE: 0.2428). While our approach uses hand-crafted features, Stöckl et al. (2021) show in their work that using Graph Neural Networks one can forgo extensive feature crafting and achieve similar accuracy results for a variety of football related machine learning tasks. As one of their applications they compare how well they can predict if a pass will be completed using a GNN, without (hardly) any feature crafting, to a simple xPass model based on standard features and find both models to achieve a similar accuracy of 0.86 and 0.85. While their accuracy is below the one our model achieves (0.92), their work still shows, that GNN's are capable of quickly working with unstructured football data, and yet achieve a relatively high accuracy.

Overall, the major benefit of an expected pass model, is that it enables more granular analysis of passing behaviour than would be possible with simple pass completion rate metrics. It can be used to quantify players' and teams' performances (Spearman et al. 2017; Power et al. 2017), by looking at their risk profiles or their efficiency. For example, the players over-performing their expected completion percentages in the

Table 6 Top 10 xPass over-performers in Bundesliga Season 2020/2021 from matchday 1–15 (at least 200 passes)

	Player	Passing Performance	Average xPass
1	Kingsley Coman	5,3 %	85,44%
2	Raphaël Guerreiro	4,9 %	86,92%
3	Max Kruse	4,7 %	82,32%
4	Christopher Nkunku	4,5 %	83,49%
5	Sebastian Rudy	4,4 %	80,69%
6	Ritsu Doan	4,3 %	76,48%
7	Daniel Caligiuri	4,1 %	77,64%
8	Josip Brekalo	4,1 %	81,33%
9	Rafael Czichos	4,1 %	84,40%
10	Joshua Kimmich	4,1 %	88,84%

Performance is defined as the difference between completion rate and average xPass values

Bundesliga season 2020/2021 (up to matchday 15) the most are shown in Table 6. The column ”Passing Performance” indicates how much a player’s actual completion rate exceeds his average xPass values.

Another application is to evaluate possible pass options, and with that a player’s decision making skills. At any given time while a player is in possession we can calculate success probabilities for hypothetical passes to teammates as shown in Fig. 4.[8] This is done by using a combination of the full model (Table 3, row 1) and the blocking model (Table 4, row 1). To find the ideal exit angle and velocity we perform a grid based search over "sensible" combinations and maximise for the xPass values. The resulting probabilities are shown for each teammate. We can see that Robin Koch (#2 of the blue team) chose one of the hardest pass options with a completion probability of 52.7%. The additional hypothetical passes — shown as lines with the respective success probabilities — come with some limitations: First, the probabilities are based on a data set, where players actively opted for a pass and since we can assume a certain amount of rationality in the decision making, values for hypothetical passes might be skewed as a consequence. Second, we assume that passes can be played at any time in any direction, without the need to properly set up before, which obviously warps reality. For instance, in the displayed situation the passing player decides to play a diagonal ball across the pitch to #7. The pass option of another long diagonal ball to number #3 — on the right side of the pitch — would require some preparation allowing opponents, especially #3 and #19 (of the red team), to get into a better defending positions. Furthermore, and this holds true in general for our xPass model, we simply compute the probability that a pass is successful, i.e. arrives at the intended target. It does not tell us if after the first touch, the teammate can hold the ball or looses it immediately thereafter.

[8] The video footage of the situation can be found here https://dfb-my.sharepoint.com/:v:/g/personal/pascal_bauer_dfb_de/EWIWkaF8Gp5CjPCdQRs5KXsB6Rt0LKHoKomXUFogNsR2Wg?e=u6EziX. The ball trajectory of the chosen pass is displayed in Fig. 1.

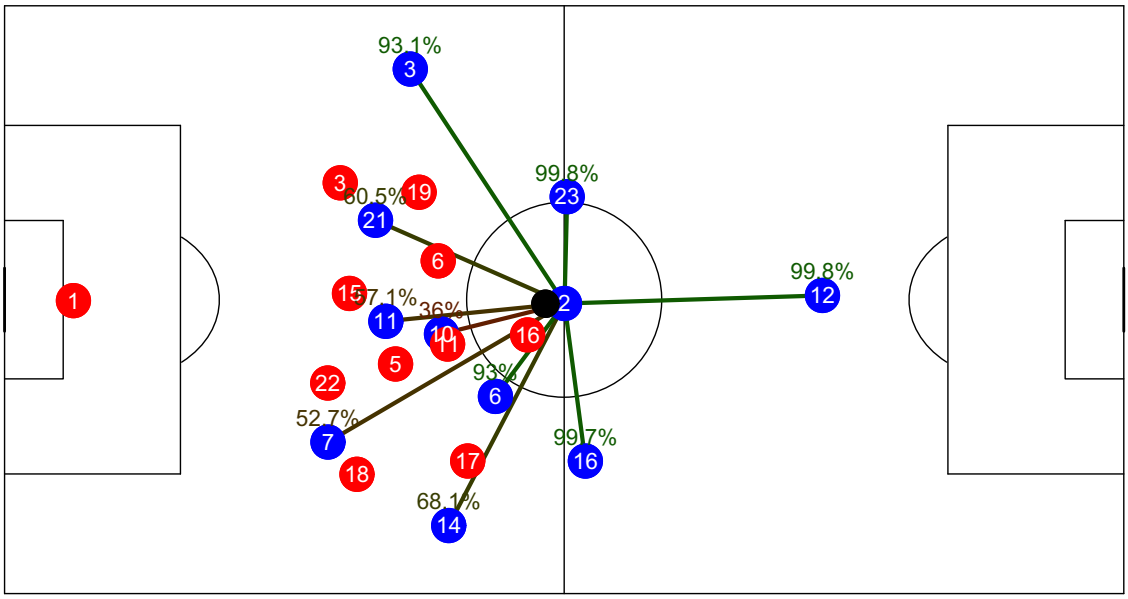

Fig. 4 Pass probabilities of hypothetical passes. This is the same situation as displayed and described in Fig. 3. The video sequence can be found here: https://dfb-my.sharepoint.com/:v:/g/personal/pascal_bauer_dfb_de/EWIWkaF8Gp5CjPCdQRs5KXsB6Rt0LKHoKomXUFogNsR2Wg?e=u6EziX

All together, we present a novel methodology for quantifying football's most relevant actions while addressing some of the shortcomings of previously published work and compare the results with existing literature. Our metric can be used to scout players outperforming their expected completion rates, identify and target weak spots in opposing teams, or show players alternative passing options they may have missed. To even better evaluate the decision making of a player, one would need to combine our risk model with a reward model (e.g. Steiner et al. (2019); Goes et al. (2019); Fernandez et al. (2018)) to not only assess a player's risk profile, but also whether they are making the best possible decision.

Acknowledgements This work would not have been possible without the perspective of professional match-analysts from world class teams who helped us to define relevant features and spend much time evaluating (intermediate) results. We would cordially like to thank Dr. Stephan Nopp and Christofer Clemens (head match-analysts of the German mens National team), Jannis Scheibe (head match-analyst of the German U21 mens national team) as well as Sebastian Geißler (former match-analyst of Borussia Mönchengladbach). Additionally, the authors would like to thank Dr. Hendrik Weber and Deutsche Fußball Liga (DFL) / Sportec Solutions AG for providing the positional and event data.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Ethics and Reproducability By informing all participating players, all tracking is compliant to the general data protection regulation (GDPR)(https://gdpr-info.eu/, accessed 07/20/20.). An ethics approval for wider research program using the respective data is authorized by the ethics committee of the Faculty of Economics and Social Sciences at the University of Tübingen. The data are property of the DFL e.V. / DFB e.V. and cannot be shared public. However, interested researchers can request samples of data under non-disclosure agreement constraints at the respective institutions. With the description of the respective tracking

Table 7 Time shift of data synchronisation against manual label

Time Shift	00.00	00.04	00.08	00.12 – 00.48	00.52 – 02.00
Dispersion	41.6%	38.3%	8.6%	15.6%	4.1%

With our frequency of 25 Hz one frame equals 00.04 seconds

vendors and systems, peers working in the football industry can reproduce the results by using any kind of professional football data.

A The synchronization of pass events

Our algorithm to synchronize passes is based on the work introduced by Anzer and Bauer (2021) to synchronize shots. Since passes occur much more frequently (915 per game) than shots (23 per game), we need to adjust the algorithm to ensure we identify the right one. For that we add one additional rule: if the algorithm finds multiple pass moments in the considered time window by the passing player, we require that the actual receiver of the pass (if there is one), must be within a 2 m distance of the ball within 5 s after the pass moment.

For one Bundesliga match (FC Bayern München vs Borussia Dortmund, matchday 24 of the 2020/2021 season) football experts gathered frame-accurate timestamps from watching the video footage for every pass (1, 088 in total). We then compared these timestamps to the ones our synchronization algorithm produced. Overall only twelve passes showed a deviation of more than two seconds. After further inspection we found that seven of them were wrongly annotated in the manual collection process, and in the other five our algorithm identified a wrong timestamp, due to either faulty tracking data (1), blocked passes (2) or identifying the wrong of two options (2). Table 7 shows how much the time stamps differ for the remaining 1, 076 passes. As we can see in about 80% of the passes the synchronization finds either the same frame or the one next to it as the human.

B Details on XGBoost expected pass and blocking model

Table 8 shows the selected hyperparameters of our final xPass model (Table 3, row 1) and the hyperparameters of the blocking model (Table 4, row 1). Additionally, Figure 5 shows the feature importance of the xPass model using SHAP-values[9] (Lundberg et al.

[9] The abbreviation **SHAP** stands for **SH**apley **A**dditive ex**P**lanation.

Table 8 Final choice of hyperparameters for the xPass model (Table 3, row 1) and the blocking model (Table 4, row 1)

	Hyperparameter	Description	Range	XPass	Blocking
1	Learning rate	Controls the step size used per update	[0, 1]	0.098	0.049
2	Max depth	Limits the depth of the tree	[0, 10]	8	6
3	Subsample	Controls number samples applied to the tree	(0, 1]	0.593	0.695
4	Min child weight	Controls instance weight of a node	[0, 10]	1.641	9.45
5	Class balancer	Controls the balance of negative and positive weights (Number of negative cases / Number of positive cases)	$(0, \infty)$	1	1

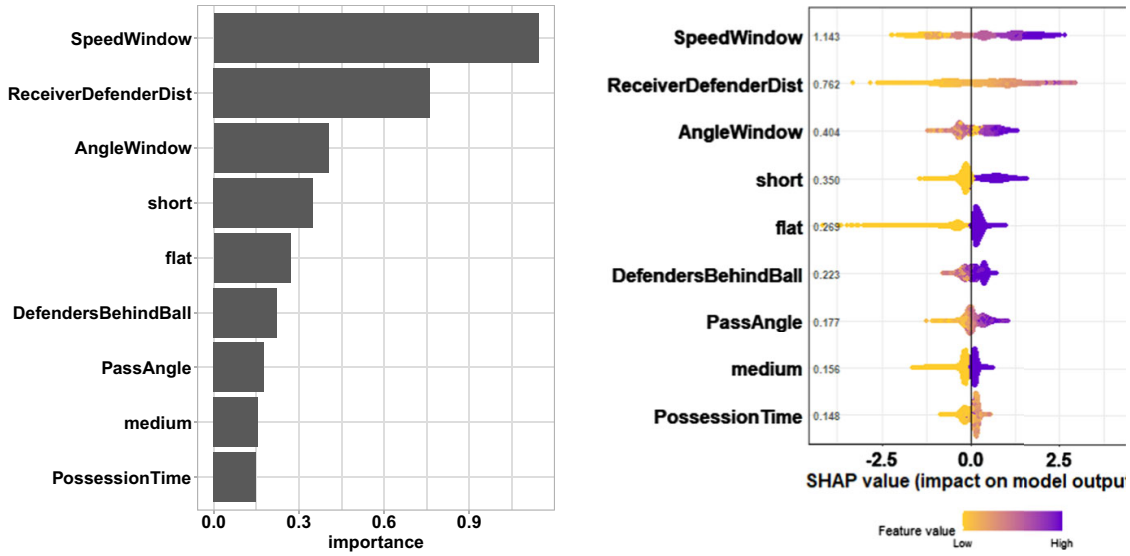

Fig. 5 Feature influence to the xPass model (Table 3, row 1) based on SHAP-values

2017). The left plot shows the absolute, overall influence of the respective features on our prediction. As discussed in Sect. 4, the speed window (*SpeedWindow*) in which the pass can be received by the intended target has the highest influence. The right plot in Fig. 5, where each colored dot (from yellow to violet) indicates the contribution of the feature to the model, shows that this relation is almost linear. The larger the speed window (i.e. violet plots), the higher the xPass value. This illustration also shows the influence of the binary features like *flat, medium* or *short*. The continuous feature, time of possession shows a dispersion of the dots similar to the binary features. This is caused by the fact, that direct played passes (i.e. 'one touch') are implicitly separated from all other passes by the model.

References

Andrienko G et al. (2017) Visual analysis of pressure in football. In: Data Mining and Knowledge Discovery 31.6, pp. 1793–1839. issn: 1573756X. https://doi.org/10.1007/s10618-017-0513-2 (cit. on pp. 6, 7)

Andrienko G et al. (2019) Constructing Spaces and Times for Tactical Analysis in Football. In: IEEE Transactions on Visualization and Computer Graphics 27.4, pp. 2280–2297. https://doi.org/10.1109/TVCG.2019.2952129. https://ieeexplore.ieee.org/document/8894420 (cit. on p. 1)

Anzer G, Bauer P (2021) A Goal Scoring Probability Model based on Synchronized Positional and Event Data. In: Frontiers in Sports and Active Learning (Special Issue: Using Artificial Intelligence to Enhance Sport Performance) 3.0, pp. 1–18. https://doi.org/10.3389/fspor.2021.624475. (cit. on pp. 2, 3, 9, 13)

Anzer G, Bauer P, Brefeld U (2021) The origins of goals in the German Bundesliga. J Sports Sci. https://doi.org/10.1080/02640414.2021.1943981

Arbués SA et al. (2020) Using player.s body-orientation to model pass feasibility in soccer. https://arxiv.org/abs/2004.07209

Asai T et al. (2007) Fundamental aerodynamics of the soccer ball. In: Sports Engineering 10.2, pp. 101–109. issn:1369-7072. https://doi.org/10.1007/bf02844207 (cit. on pp. 2, 3)

Battaglia PW et al. (2018) Relational inductive biases, deep learning, and graph networks (cit. on p. 2)

Bauer P, Anzer G (2021) Data-driven detection of counterpressing in professional football a supervised machine learning task based on synchronized positional and event data with expert-based feature extraction. Data Mining and Knowledge Discovery. https://doi.org/10.1007/s10618-021-00763-7

Bradley PS et al. (2013) The effect of high and low percentage ball possession on physical and technical profiles in English FA Premier League soccer matches. In: Journal of Sports Sciences 31.12, pp. 1261–1270. issn: 02640414. https://doi.org/10.1080/02640414.2013.786185 (cit. on p. 1)

Bransen L, Haaren JV (2019) Measuring football players' on-the-ball contributions from passes during games. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11330 LNAI, pp. 3.15. issn: 16113349. https://doi.org/10.1007/978-3-030-17274-91 (cit. on p. 1)

Brefeld U, Lasek J, Mair S (2019) Probabilistic movement models and zones of control. In: Machine Learning 108.1, pp. 127.147. issn: 15730565. https://doi.org/10.1007/s10994-018-5725-1. (cit. on pp. 2, 4, 9)

Brier GW (1950) Verification of Forecasts Expressed in Terms of Probability. In: Monthly Weather Review 78.1, pp. 1.3. issn:0027-0644. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2 (cit. on p. 7)

Brooks J, Matthew K, John G (2016). Developing a data-driven player ranking in soccer using predictive model weights. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 49.55. https://doi.org/10.1145/2939672.2939695 (cit. on p. 1)

Chawla S et al. (2017) Classification of passes in football matches using spatiotemporal data. In: ACM Transactions on Spatial Algorithms and Systems 3.2. issn: 23740361. https://doi.org/10.1145/3105576 (cit. on p. 2)

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17, pp. 785.794. https://doi.org/10.1145/2939672.2939785 (cit. on p. 6)

Fernandez J, Bornn L (2018) Wide Open Spaces : A statistical technique for measuring space creation in professional soccer. In: MIT Sloan Sports Analytics Conference, Boston (USA), pp. 1.19 (cit. on pp. 2, 4, 10)

Fernández J, Bornn L, Cervone D (2020) A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. https://arxiv.org/abs/2011.09426

Francisco PA et al. (2020) Seeing in to the future : using self-propelled particle models to aid player decision-making in soccer. In: MIT Sloan Sports Analytics Conference. Boston (USA). pp. 1–23 (cit. on p. 2)

Goes F et al. (2021) A risk-reward assessment of passing decisions: comparison between positional roles using tracking data from professional men's soccer. In: Science and Medicine in Football 00.00, pp. 1.9. issn: 2473-3938. https://doi.org/10.1080/24733938.2021.1944660.(cit. on p. 2)

Goes FR et al. (2019) Not Every Pass Can Be an Assist: A Data-Driven Model to Measure Pass Effectiveness in Professional Soccer Matches. In: Big Data 7.1, pp. 57.70. issn: 2167647X. https://doi.org/10.1089/big.2018.0067. (cit. on pp. 2, 10)

Gómez JLI et al (2019) Landscapes of passing opportunities in Football . where they are and for how long are available ? In: Barça sports analytics summit February, pp. 1.14 (cit. on p. 2)

Hubáček O, Šourek G, Železný F (2018) Deep learning from spatial relations for soccer pass prediction. In: CEUR Workshop Proceedings 2284, pp. 162.169. issn: 16130073. https://dtai.cs.kuleuven.be/events/MLSA18/papers/hubacek_mlsa18.pdf (cit. on pp. 2, 6)

Król M et al. (2017) Pass Completion Rate and Match Outcome at the World Cup in Brazil in 2014. In: Polish Journal of Sport and Tourism 24.1, pp. 30.34. issn: 2082-8799. https://doi.org/10.1515/pjst-2017-0004 (cit. on p. 1)

Li H, Zhang Z (2019) Predicting the receivers of football passes. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11330 LNAI, pp. 167.177. issn: 16113349. https://doi.org/10.1007/978-3-030-17274-9_15 (cit. on pp. 2, 9)

Linke D, Link D, Lames M (2018) Validation of electronic performance and tracking systems EPTS under field conditions. In: PLoS ONE 13.7, pp. 1.20. issn: 19326203. https://doi.org/10.1371/journal.pone.0199519 (cit. on p. 3)

Linke D, Link D, Lames M (2020) Football-specific validity of TRACAB's optical video tracking systems. In: PLoS ONE 15.3, pp. 1.17. issn: 19326203. https://doi.org/10.1371/journal.pone.0230179 (cit. on p. 3)

Lundberg SM, Lee SI (2017) Consistent feature attribution for tree ensembles. https://arxiv.org/abs/1706.06060

Mchale IG, Lukasz S (2014) A mixed effects model for identifying goal scoring ability of footballers. In: Journal of the Royal Statistical Society. Series A: Statistics in Society 177.2, pp. 397.417. issn: 09641998. https://doi.org/10.1111/rssa.12015 (cit. on p. 6)

McHale IG, Relton SD (2018) Identifying key players in soccer teams using network analysis and pass difficulty. In: European Journal of Operational Research 268.1, pp. 339.347. issn: 03772217. https://doi.org/10.1016/j.ejor.2018.01.018 (cit. on p. 1)

Nazareth JL (2004) An Optimization Primer. In: An Optimization Primer, pp. 2.5. https://doi.org/10.1007/978-1-4684-9388-7 (cit. on p. 6)

Oggiano L, Satran L (2010) Aerodynamics of modern soccer balls. In: Procedia Engineering 2.2, pp. 2473. 2479. issn: 18777058. https://doi.org/10.1016/j.proeng.2010.04.018. (cit. on pp. 2, 3)

Pappalardo L et al. (2019) A public data set of spatio-temporal match events in soccer competitions. In: Scientific Data 6.1, pp. 1.15. issn: 20524463. https://doi.org/10.1038/s41597-019-0247-7.(cit. on p. 3)

Pettersen SA et al. (2014) Soccer video and player position dataset. In: Proceedings of the 5th ACM Multimedia Systems Conference, MMSys 2014 (Singapore, March 2014), pp. 18.23. https://doi.org/10.1145/2557642.2563677 (cit. on p. 3)

Power P et al (2017) Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Part F1296:1605–1613. https://doi.org/10.1145/3097983.3098051 (cit. on pp. 1, 2, 6, 9, 10)

Reep C, Benjamin B (1968) Skill and Chance in Association Football Author. In: Journal of the Royal Statistical Society 131.4, pp. 581–585. issn: 14698005. https://www.jstor.org/stable/2343726?seq=1 (cit. on p. 1)

Rein R, Raabe D, Memmert D (2017) Which pass is better? Novel approaches to assess passing effectiveness in elite soccer. In: Human Movement Science 55.July, pp. 172–181. issn: 18727646. https://doi.org/10.1016/j.humov.2017.07.010. (cit. on p. 2)

Spearman W et al. (2017) Physics-Based Modeling of Pass Probabilities in Soccer. In: MIT Sloan Sports Analytics Conference, Boston (USA), pp. 1–14. https://www.researchgate.net/profile/William-Spearman/publication/315166647_Physics-Based_Modeling_of_Pass_Probabilities_in_Soccer/links/58cbfca2aca272335513b33c/Physics-Based-Modeling-of-Pass-Probabilities-in-Soccer.pdf (cit. on pp. 2–4, 6, 9, 10)

Stein M et al. (2017) How to Make Sense of Team Sport Data: From Acquisition to Data Modeling and Research Aspects. In: Data 2.1, p. 2. issn: 2306-5729. https://doi.org/10.3390/data2010002 (cit. on p. 1)

Steiner S et al. (2019). Outplaying opponents–a differential perspective on passes using position data. In: German Journal of Exercise and Sport Research 49.2, pp. 140–149. issn: 25093150. https://doi.org/10.1007/s12662-019-00579-0 (cit. on pp. 2, 7, 10)

Stöckl M et al. (2021) Making Offensive Play Predictable - Using a Graph Convolutional Network to Understand Defensive Performance in Soccer. In: MIT Sloan Sports Analytics Conference, Boston (USA), pp. 1–19 (cit. on pp. 2, 9)

Szczepański L, Mchale I (2016) Beyond completion rate: Evaluating the passing ability of footballers. In: Journal of the Royal Statistical Society. Series A: Statistics in Society 179.2, pp. 513–533. issn: 1467985X. https://doi.org/10.1111/rssa.12115 (cit. on p. 1)

Vercruyssen V, Raedt LD, Davis J (2016) Qualitative spatial reasoning for soccer pass prediction. In: CEUR Workshop Proceedings 1842. issn: 16130073 (cit. on pp. 2, 9)

Williams LRT (2000) Coincidence timing of a soccer pass: Effects of stimulus velocity and movement distance. In: Perceptual and Motor Skills 91.1, pp. 39–52. issn:00315125. https://doi.org/10.2466/pms.2000.91.1.39 (cit. on p. 1)

C Appendix—Study III: The Origins of Goals in the German Bundesliga

In the following, we present Anzer, Bauer, and Brefeld (2021), an accepted Manuscript of an article published by Taylor Francis in Journal of Sport Science on July 25th 2021, available online: http://www.tandfonline.com/10.1080/02640414.2021.1943981

The Origins of Goals in the German Bundesliga

Gabriel Anzer[1,2], Pascal Bauer[2,3] and Ulf Brefeld[4]

[1]Sportec Solutions AG, subsidiary of the Deutsche Fußball Liga (DFL)

[2]Institute of Sports Science, University of Tübingen

[3]DFB Akademie, Deutscher Fußball-Bund e.V. (DFB)

[4]Leuphana University of Lüneburg, Machine Learning Group

Abstract

We propose to analyze the origin of goals in professional football (soccer) in a purely data-driven approach. Based on positional and event data of 3,457 goals from two seasons German Bundesliga and 2nd Bundesliga (2018/20219 and 2019/2020), we devise a rich set of 37 features that can be extracted automatically and propose a hierarchical clustering approach to identify group structures. The results consist of 50 interpretable clusters revealing insights into scoring patterns. The hierarchical clustering found 8 alone standing clusters (penalties, direct free kicks, kick and rush, one-two's, assisted by header, assisted by throw-in) and 9 categories (e.g. corners) combining more granular patterns (e.g. 5 subcategories of corner-goals). We provide a thorough discussion of the clustering and show its relevance for practical applications in opponent analysis, player scouting and for long-term investigations. All stages of this work have been supported by professional analysts from clubs and federation.

Keywords Sports analytics • Professional football (soccer) • Hierarchical clustering • Tactical analysis

1 Introduction

In the 1960s, Charles Reep began to manually annotate games of Swinden Town FC (Reep et al., 1968). Though, some of his data-driven conclusions were later questioned (Witts, 2019) , his primitive collection of detailed game data constituted the birth of football analytics: Today, most international leagues not only collect manually annotated events from their matches systematically, but also use device- or camera-based tracking systems in addition. Compared to event logs focusing on ball-actions (e.g. passes, shots, fouls; often referred to as event data), tracking systems allow to record the positions of all 22 players and the ball for an entire match (often referred to as positional or tracking data).

Several studies aim to group goals—the most important quantified metric in football—into predefined categories. For example, González-Ródenas et al. (2019) differentiate between open-play and set-pieces to categorize 380 goals from of the UEFA Champions League 2016/2017 season. They observe that 75.9% of all goals occur from open-play, and only 24.1% are scored from set-pieces. A similar study confirms these numbers on 101 goals taken from the World Cup in 2010 (Njororai, 2013). However, the expressiveness of manually crafted categories is naturally limited as only a relatively small amount of data can be processed by hand. Focusing on more detailed features of the goal origin, rather than high-level groups (e.g. whether the shooter was under pressure), Mitrotasios et al. (2012) investigated factors associated with goal scoring in 76 matches of the European Championship in 2012. Besides finding a similar dispersion of goal origins (27.6% after set-pieces; 72.4% from open-play), they show that in more than 50% of the cases the goal-scorer took his shot without any pressure. Plummer (2013) analyze goal scoring patterns of a lower English league, pointing out stark differences in the origin of goals between non-professional leagues. In general, set-pieces are often a decisive factor for winning a game, particularly when teams are equally strong (Szwarc, 2007; Göral, 2019). Especially for corner-kicks, there exist several studies examining how they lead to goals: (Taylor et al., 2005; Carling et al., 2006; Armatas et al., 2007; Schmicker, 2013; Pulling et al., 2013; Pulling, 2015; Casal et al., 2015; Fernández-Hermógenes et al., 2017; Casal et al., 2017). While these papers are based on manually annotated data, Power et al. (2017) offer an approach using tracking data. They report a scoring efficiency of 2.1% after corner kicks for the English Premiere League 2016/2017 season. They also found that scoring with the second ball touch after a corner, is even more likely than converting with the first touch.

Whereas the usage of manually recorded data, acquired for the sole purpose of a single investigation, is a common practice in sport sciences, the potential of automatically acquired positional data, as well as off-the-shelf available event data, has not been fully exploited—particularly when it comes to clustering goals. Hobbs et al. (2018) aim to detect counterattack situations automatically, based on positional and event data, and derive that it is the most efficient strategy for scoring goals. Sarmento et al. (2014) propose mixed methods to analyze attacking patterns of 36 games of different European top teams. They also focused on counterattacks and combined quantitative analyses with expert knowledge to discover team philosophies, showing that the combination proved to be very beneficial. Several studies highlighted the relevance of this interplay between sports and computer science (Rein et al., 2016; Herold et al., 2019; Andrienko et al., 2019; Goes et al., 2020; Marcelino et al., 2020)

Note that a similar approach has been successfully deployed in basketball. Reich et al. (2006) investigate so-called *shot charts*, that visualize the location and outcome of every shot, in professional basketball matches. Similar visualizations are enriched by spatial clustering techniques in (López et al., 2013). An approach taking different shot characteristics into consideration is presented in Erčulj et al. (2015). Although these analyses often focus on the location of shots, it led to significant changes in team strategies and player's shooting decisions (López et al., 2013; Reich et al., 2006). Simply focusing on shot locations of goals does of course not translate to football with its complex attacking plays, its low scoring nature, different shot types (e.g. header, volley) and the additional role of a goalkeeper.

Since only about 1% of all ball possession phases are completed with a goal (Pollard et al., 1997; Tenga et al., 2010), many studies thus extend the focus to all shots (Fernando et al., 2015) or on proxies such as carrying the ball into dangerous zones (Njororai, 2013; Merlin et al., 2020) in order to quantify offensive success on larger sample sizes. Although these approaches may be biased towards successful teams that use their chances more effectively (Castellano et al., 2012; Delgado-Bordonau et al., 2013; Dufour et al., 2017), they allowed studies to evaluate processes (i.e. an attacking-play) more granular than just by considering pure results. Ruiz et al. (2015) analyze the efficiency of shots taken based on the distance and angle to the goal, Schulze et al. (2018) consider also the set-up of the opposing team during the shot to improve the expressiveness of the investigation. On the basis of these ideas, a lot of expected goal models exist that aim to quantify scoring probabilities (Lucey et al., 2014; Rathke, 2017; Ruiz et al., 2017; Robberechts et al., 2020; Anzer et al., 2021).

Consequently, compared to other team sports, goals in football are rare events and there is a trivial probability of observing the same goal twice as every goal is sui generis due to the complexity of the game (Siegle et al., 2013; Salmon et al., 2020). Nonetheless, teams come up with dedicated match plans to increase the probability of scoring and winning the game. Coaches and video-analysts devise attacking patterns that ought to exploit weaknesses of the opposing team and result in the creation of chances. Since these patterns are not random, there must be structure in the creation of goals. In that respect, the categorization of goals plays an important role in the daily business of professional football clubs. Clubs typically employ several match-analysts whose role includes to regularly examine scored and conceded goals, particularly before facing an opponent. Since viewing the video footage is a tedious task and even experts may disagree on categories (Chawla et al., 2017), it is the objective of this paper to both automatize and objectify the categorization of goals and support the respective match-analysis departments: Being able to cluster goals by their origin allows for an unbiased analysis that provides unseen patterns and discloses trends.

In this paper, we follow a data-driven approach to leverage such data to identify the underlying structure of the origin of goals in professional football. The contribution of this paper is as follows: First, we propose a rich set of expert features that can be computed from aligned positional and event data to formally represent goals as instances in a vector space. Second, we deploy a hierarchical clustering (Murtagh et al., 2017) to group $3,457$ goals from two seasons of the German Bundesliga and 2^{nd} Bundesliga into meaningful and interpretable clusters and provide a thorough analysis of the results. Compared to the literature, our analysis is on a much larger scale, provides rich feature representations, and follows a purely data-driven approach that renders manual categorization or the definition of rules unnecessary. All quantitative results have been evaluated qualitatively by professional match-analysts.

2 Methods

2.1 Data

The German Bundesliga and 2^{nd} Bundesliga collects tracking and event data for all their league matches. The former is captured by optical tracking systems while the latter consists of manual annotations. *Tracking data* is recorded automatically using camera-based systems. Optical tracking systems are installed in every stadium and capture the positions of players, referees and the ball at 25 frames per second. The quality of

the tracking data acquired by Chyronhego's TRACAB system[1] is evaluated on a regular basis and presents sufficient accuracy (Taberner et al., 2020; Linke et al., 2020). However, there remain many events on the pitch that currently cannot be captured automatically. The *event data* is therefore collected manually. Trained operators annotate about 3,000 basic events per match categorized into different event classes. There are 30 top-level event classes including passes, crosses, fouls, etc. as well as about another 100 sub-attributes describing the events even in greater detail. The definition of each event follows the official match data-catalogue designed by German Bundesliga.[2] For further processing, the tracking and event data are synchronized so that the timestamps of the events are aligned to the right frames in the tracking data as described in Anzer et al. (2021).

We focus in our analysis on 3,457 goals scored in the Bundesliga and 2nd Bundesliga in the 2017/2018 and 2018/2019 seasons and excluded the 85 own goals due to their often random nature. Every goal is described by the raw data of all 22 players and the ball in 25Hz as well as all annotated events during the ball possession phase leading to the goal. We also extracted 8.167 shots of the season 2018/2019 (containing 953 goals). The shots are used for an efficiency analysis of each cluster as described later.

2.2 Mapping goals into feature space

We extract a rich feature set from the synchronized data to turn goals into machine-readable quantities encoding episodes that end with a successful shot at goal. We mirror the pitch in both dimensions, so that all goals are scored on the same side of the field. Later on, this transformation remedies the clustering from having to differentiate between left or right wings.

Besides the location and set-up of the shot itself, football experts (i.e. coaches, match-analysts, ...) are explicitly interested in the complete ball possession phase prior to the goal. However, the fluent invasive character of football implicates a lot of vagueness in terms of a consistent definition of an *attacking play* (Merlin et al., 2020). Particularly very short ball possession phases of defending players during an attacking play should not be considered as a separate ball possession phase. To establish an appropriate definition, we reviewed video footage of critical scenes together with experts. Finally, we define the start of such an episode as either a dead-ball situation (e.g. throw-in, goal-kick, etc.) or a turnover by the opposing team lasting at least six seconds.

Together with experts—match-analysts with a minimum of five years experience in professional football teams[3]—we define in total 37 features describing the evolution of a goal, from the origin to its finish. The features are described in detail in the Appendix A. To provide an accurate representation of what leads to a goal, the features make full use of the synchronization of the positional data with the manual collected event data. In total we settled on features describing the shot itself (location, type, goalkeeper positioning, pressure on the goal scorer, ...), its assist (location, assist type, ...) and features describing the entire ball possession phase leading up to the goal The latter features include the location and type of the initial gain of the ball, the number of passes, meters dribbled and bypassed opponents. As a measure of chaos, the number of opponent touches during the ball possession phase is also counted. Next to prominent scores like expected goals (xG)[4], describing the probability of a shot being converted, we include several categorical expert features, such as whether a chance is a sitter, originates from a counterattack, etc. Categorical features are one-hot encoded in the final representation.

More sophisticated metrics describing the ball possession phase, the assist or the shot itself, present in the literature were also used. To quantify the average pressure, for instance, we implemented the approach taken by Andrienko et al. (2017). Additionally, the compactness of both teams is a decisive factor to differentiate transition situations and counterattacks from other open-play situations. We therefore added the *stretch-index* based on the definition in Santos et al. (2018) at the beginning and at the end of the ball possession phase. Finally, the number of successfully played passes within an attacking play is complemented with a *packing value*—describing the number of outplayed opponents per pass as in Steiner et al. (2019)—to include a notion of the degree of ball control the offensive team had prior to scoring the goal. All features were discussed, consolidated and steadily improved during workshops and based on several steps of evaluation.[5]

[1] `https://chyronhego.com/products/sports-tracking/tracab-optical-tracking/`, accessed 06/20/2020

[2] `https://www.bundesliga.com/en/news/Bundesliga/noblmd-dfl-subsidiary-sportcast-setting-up-company-for-official-match-data.jsp`, accessed 02/02/2020

[3] We provide more information on the experts in the acknowledgements.

[4] The xG-value used is calculated as defined in Anzer et al. (2021).

[5] A video showing some of the features is available at `https://bit.ly/3sa3phw`.

2.3 Clustering the goals

To accomplish practical needs, it is our primary objective to automatically assign goals to interpretable categories. We refrained from collecting labeled data from match-analysts for two reasons: On one hand, categories of goals differ per club, coach and the respective match philosophy, and we prefer to compute an objective structure that can be augmented in the daily practice irrespectively of the club, analyst or philosophy. On the other hand, time constraints do allow match-analysts to review only a small amount of data and the categorization of goals is naturally on a rather high level. Manual inspection of only a few goals per opponent does not allow for detecting the variety of clusters that a purely data-driven approach is able to produce at large-scales. A data-driven clustering allows us to reveal and discuss the hidden structure of goals with our experts. To the best of our knowledge this is the first purely data-driven approach to clustering goals on synchronized positional and event data and clearly unmatched in terms of scale.

Agglomerative hierarchical clustering (HCA) provides a conceptually simple framework to compute interpretable clusterings. HCA works bottom-up by (i) initializing every instance as a singleton cluster, and (ii) iteratively combining the two most similar clusters, (iii) until only one cluster remains that contains all instances. The resulting structure is a cluster tree called dendrogram (Murtagh et al., 2017). Different instantiations of HCA arise by different ways to merge clusters in step (ii). For instance, single-link merges the two clusters containing the two most similar elements (Sibson, 1973). Hence, single-link often leads to chain-like structures as only one element of the cluster needs to be similar to one of the other while all other instances may be very dissimilar. The other extreme is called max- or complete-link and focuses on the most different elements when merging clusters (Defays, 2015). Max-link leads therefore to more balanced clusters (Brian, 2011). We do not want to put such a strong prior on the solution and instead leverage a compromise called average-link that merges clusters that are closest on average (Sokal, 1958). Average-link is often used in bio- and health-related domains, for instance to infer phylogenetic tress (Felsenstein, 1996), and serves our needs very well. However, instead of commonly used Euclidean distances, we deploy cosine distance to meet the characteristics of the data. Recall that numeric elements of the feature representation encode variables like packing or pass distance. Consider two similar goals, where one has almost twice the packing score and almost twice the pass distance than the other. Using Euclidean distance, the two goals would turn out very different. However, the angle between the two vectors in feature space is small and, hence, the cosine implements the intuition that longer passes may also result in higher packing scores. In sum, similarity of cluster X' and X is computed by

$$\mathrm{sim}(X, X') = \frac{1}{\mid X \mid\mid X' \mid} \sum_{x \in X} \sum_{x' \in X'} \frac{x^T x'}{\|x\| \|x'\|} \tag{1}$$

An extensive model selection optimizes the pre-processing pipeline as well as additional parameters like the number of clusters. The final solution maximized the silhouette measure (Rousseeuw, 1987) and consists of a z-transformation and a subsequent mapping onto the 20 most informative dimensions corresponding to the largest eigenvalues identified in a principal component analysis (PCA) (Wold et al., 1987) before the data is fed into the hierarchical clustering using 50 clusters.

The resulting dendrogram is shown in Figure 1. Starting at the root, the hierarchy differentiates primarily between the assist type before splitting further into goals arising individualized features per branch. The tree is evaluated together with professional match-analysts of national teams and a Bundesliga club to analyze its possible use for practice. Together with the experts, we went through 2D visualizations of the goals and corresponding video footage, in order to derive a better grasp of the clustering. To reduce workload, primarily the 2D visualizations were used by the experts to assign names and descriptions to all nodes in the dendrogram. These characterizations were evaluated on random samples of video footage manually to verify the solution; in total more than 800 goals were viewed.

After finalizing the contextual description of the clustering, the experts agreed on a simplified version of the tree they would use in their match-analysis. This simpler version essentially merges small clusters with close neighbours. We indicate merged clusters by the same colors in Figure 1 and provide a thorough discussion in the remainder.

3 Results

In this section, we discuss the induced grouping by the dendrogram in Figure 1 and highlight interesting features of the data-driven solution. In the remainder, we differentiate between goals from open-play, set-pieces, type of assist, type of shot, and dedicated special goals. Representative goals for each cluster can be found in the Appendix D. To assess conversion rates per cluster, we classified 8,167 shots from 2018/2019 season into the clustering by assigning every goal to the most similar cluster using the distance metrics

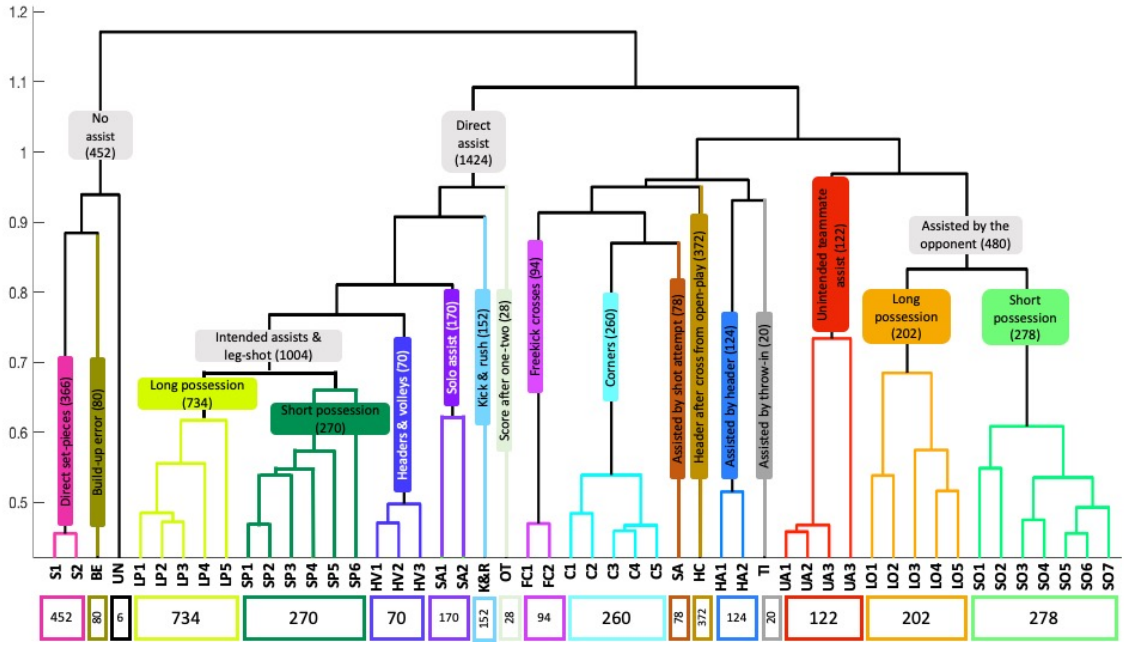

Figure 1: The resulting dendrogram with contextual annotations. Numbers show the amount of goals in the respective branches.

suggested in the previous chapter. Table 5 in the Appendix C provides additional details on conversion rates per cluster.

3.1 Open-play

A straight forward classification of goals is to differentiate between goals that originate from open-play and from set-pieces. With in total 2,231 goals, in-play goals constitute the most frequent type of goals in our data, with 64.0% of all goals being placed in one of the corresponding clusters. Focusing on the former, open-play ball possession phases leading to a goal contain 3.6 passes and last 12.8 seconds on average. Clusters containing goals from open-play are spread throughout the clustering; Figure 2 shows two-dimensional visualizations of exemplary goals for the largest clusters.

The majority of all in-play goals are contained in the *light green* and *dark green* clusters and add up to a total of 1,424 goals. The goals can be distinguished by an intended assist from a teammate, without an opponent touching the ball between assist and shot. The individual clusters further differentiate nuances of the goal's origin. For example, **LP1** and **LP2** in the light green cluster represent prototypical goals from build-up to a finish. Goals in **LP2** however, are typically the greater chances as 98.0% are labeled as sitters and their goals per shot ratio is 42.0% **LP2** compared to **LP1**'s of 20.0%.

The clustering allows to further dive into the resulting groups and show fine granular differences that are usually only identified by manual expert inspections. As an example, Figure 3 shows that the clustering differentiates between different strategies to regain the ball during an opponent's possession. Coaches and teams develop complex patterns that involve coordinated actions by many players and we easily identify goals after successful *counterpressing* (**SP1**), *midfield-block pressing* (**SP2**) and *high-block pressing* (**SP5**). Figure 3 shows heat maps of the shot location (top), the assist location (center), and the start of the ball possession phase (bottom). **SP1**, for example, contains 128 goals scored after regaining the ball in the opponents half, preferably close to the sideline.

Interestingly, about 40.0% of all shots in **SP5** lead to a goal compared to only 5.0% for **SP1** and 2.0% for **SP2**. The numbers support that excellent goal opportunities are created by a very high pressing. By contrast, **SP2** turns out the most inefficient cluster in terms of goal conversion rate.

In-game crosses that are directly converted into goals are contained in **HV3** and **HC**. Both clusters encode crucial goal-scoring patterns. In **HV3**, for example, the ball is gained in the own half and after a save build-up phase crossed from just inside the box and converted directly with a header (typically labeled as a sitter). **HC** distinguishes itself by broader areas where the ball has been won, particularly including the wings in the opponent's half and crosses in this cluster are predominantly played from outside the box. Figure 4 visualizes the differences using heat maps. Note that **HC** contains more than 10% of all goals

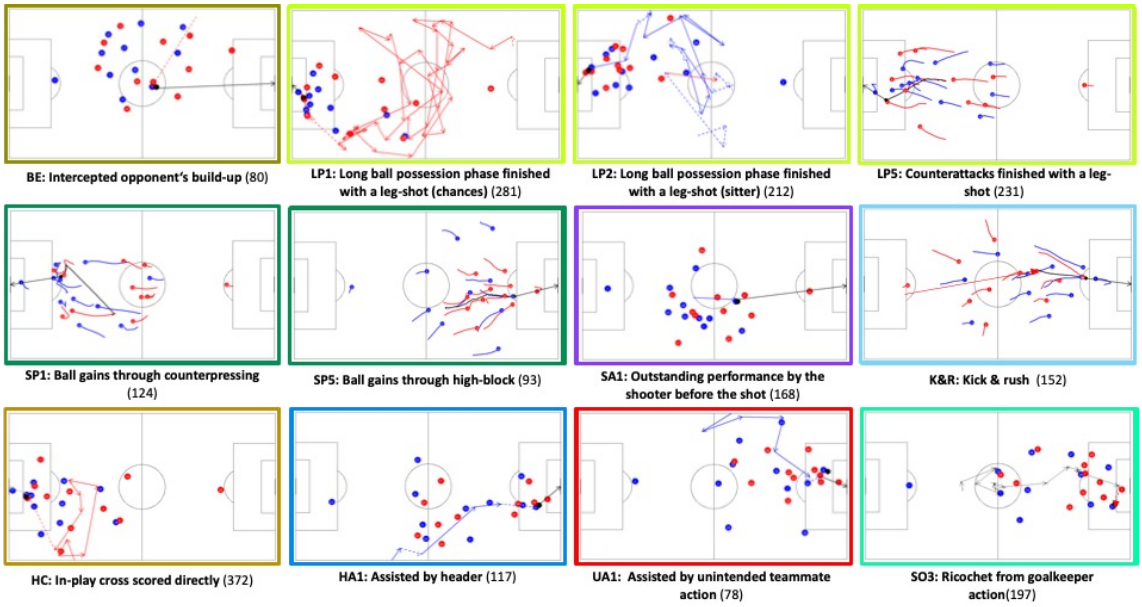

Figure 2: Exemplary in-play goals and their clusters. Arrows show the path of the ball leading up to the successful shot, positions of players at the time of shot are indicated in blue and red, respectively.

and is by far the largest cluster in the tree.

3.2 Set-Pieces

Roughly a third (in total 1,227) of all converted ball possession phases in our data begin with a set-piece in the opponent's half. Hence, match-analysts dedicate a significant amount of their time to identify opponent's strategies and tricks for all sorts of set-pieces. Figure 5 shows exemplary goals from clusters representing corners and crosses. A total of 7.1% of all goals originate in corners and contained in the cyan clusters **C1**–**C5**. Cluster **C1** contains all goals where the ball is touched by at least one opponent before it is received by the scorer; these situations often end-up in rather uncontrolled *ping-pong* situations in the box. Cluster **C2** encodes *flick-ons*, where a target player is positioned at the closest post who slightly deflects the ball before it can be converted. This cluster is complemented with **HA1** that contains header flick-ons. Goals in **C2** show very high xG values with an average of 60.0% and all were rated as *sitters* by the experts.

Set-pieces played as crosses into the box follow a similar idea as corner kicks but turn out to be less effective. In total we count 330 freekick-crosses in cluster **S1** and **S2** in the data but only 4.7% of them were converted to goals. The clustering distinguishes between three scenarios: Taking the freekick-cross directly (**FC1**, 97 goals), scoring after a resulting ping-pong-situation (**FC2**, 30 goals), and scoring the rebound of a freekick-cross in a spectacular way (**SO2**, 12 goals).

The most straight-forward way of turning set-pieces into goals is through penalties (272 goals, **S1**) and direct freekicks (94 goals, **S2**). Together, the two clusters account for 13.1% of all scored goals in the two seasons. Unsurprisingly, penalties are the most efficient way of scoring. Even without taking deflected penalties into consideration, 91 penalties in Bundesliga season 2018/2019 lead to 74 direct goals which corresponds to a conversion rate of 81.3%. If the goalkeeper initially parries a penalty, but the rebound is then converted, the goal is not considered as a penalty goal and therefore part of a different cluster **SO3** and described in the remainder.

In total 7.2% of all direct attempted freekicks from Bundesliga 18/19 season lead to a goal. Figure 6 shows the shot locations of all directly scored freekicks in **S2**. Throw-in crosses very rarely lead to goals (20 goals in total), which are contained in **TI**.

3.3 Assists

The dendrogram in Figure 1 differentiates between types of assists. Clusters **S1** and **S2** as described in the previous section, stem from directly scored set-pieces and trivially do not contain assists. Similarly, **BE** represents goals where a pass of the defending team was intercepted and converted by the scorer. Figure 7

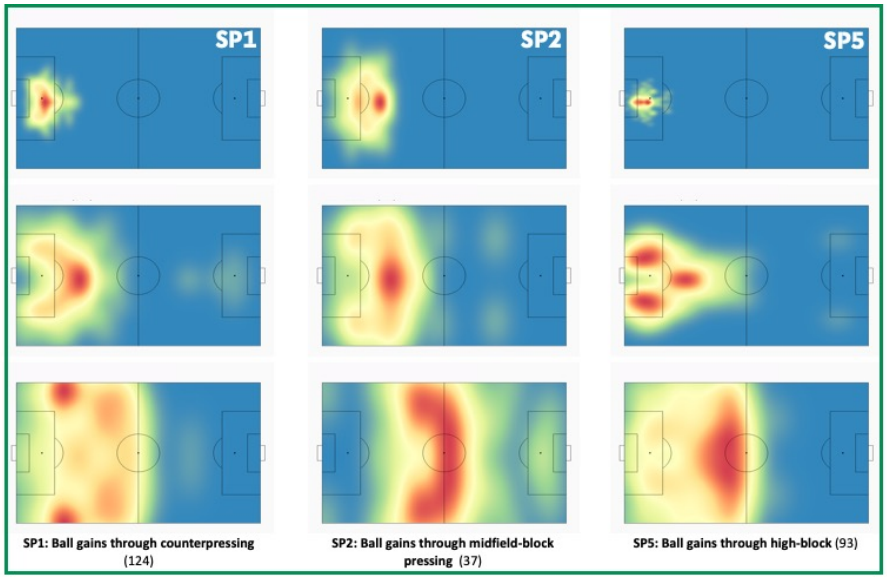

Figure 3: Goals originating from strategic ball gains in **SP1**, **SP2**, and **SP5**. The figure shows heat maps for shot (top row), assist location (center row) and start of the ball possession phase (bottom row). In total 270 goals (7.8% of all) were scored this way.

(top left) shows a heat map of shot locations in **BE**. While the majority of goals in that cluster are scored from within the box, there are clearly visible outliers indicating long-ranged shots at goal. On average, cluster **BE** is characterized by fatal build-up errors that allow shots from large distances to be converted due to mispositioning by the goalkeeper.

The clustering further differentiates between types of assists such as an intentionally played final pass that clearly aims to assist the scorer. This very large group containing 1,949 goals is further divided by the clustering into goals from open-play with an intended assist (dendrogram **LP1**–**LP5**, **SP1**–**SP6**, **HV1**–**HV3** and **OT**) and assists in form of crosses. The latter contains directly converted goals by corners (**C3**), freekick-crosses (**FC1**), and open-play crosses (**HC**) as well as goals arising only after several opposing ball touches; these ping-pong situations are again separated into corners (**C1**), freekick-crosses (**FC2**), and open-play crosses (**LO3** and **LO5**). Moreover, there are spectacular rebound-volleys where unsuccessful clearances are scored at large distances (**C1** and **SO2**, see below). Cluster **HA1** contains header assists by flick-ons after crosses and long balls from the own half.

Unintentional assists may arise from regular passes that are completed with outstanding maneuvers of the scorer and can be found in **SA2** and **SA1**. The contextual analysis of these clusters showed two different kinds of situations: Either the scorer takes a surprise shot, often at large distance or from difficult angles (two plots on the right side), or dribbles past several opponents before taking a shot.

Many unintentional assists are simply random and contained in the clusters **UA1**–**UA4**. In contrast to assists by opponents (**LO1**–**SO7**), these random assists come in fact from a teammate but without the direct intention to create a shot. The experts consider goals in this group to be lucky events. Nevertheless, fortune picks its favorites: in our data, the luckiest teams in every league and season scored about twice as many random goals as the unluckiest ones. Cluster **SA** contains shot attempts that are deflected by team members. Positioning players in the line of shot turns out to be very efficient: almost half of the situations are converted into goals

The last group in this section constitutes *indirect assists* from opponents. We already discussed intercepted build-ups in **BE**, however, compared to **BE**, indirect assists in **LO1**–**SO7** primarily stem from uncontrolled and random opponent actions. Additional indirect assists are also contained in the ping-pong clusters **C1** (corners) and **FC2** (freekick-crosses). Figure 8 visualizes exemplary goals induced by indirect assists. For example, Cluster **SO3** contains all goals where the opponent's goalkeeper failed to save a shot and accidentally assisted the scorer. With 197 goals this cluster contains surprisingly many goals, albeit, our experts do not consider all these situations as mistakes by the goalkeeper. Although 'flaws' of the goalkeepers are often a decisive factor in top leagues, a characteristic trait of excellent strikers is their sixth sense for these *poacher goals*.

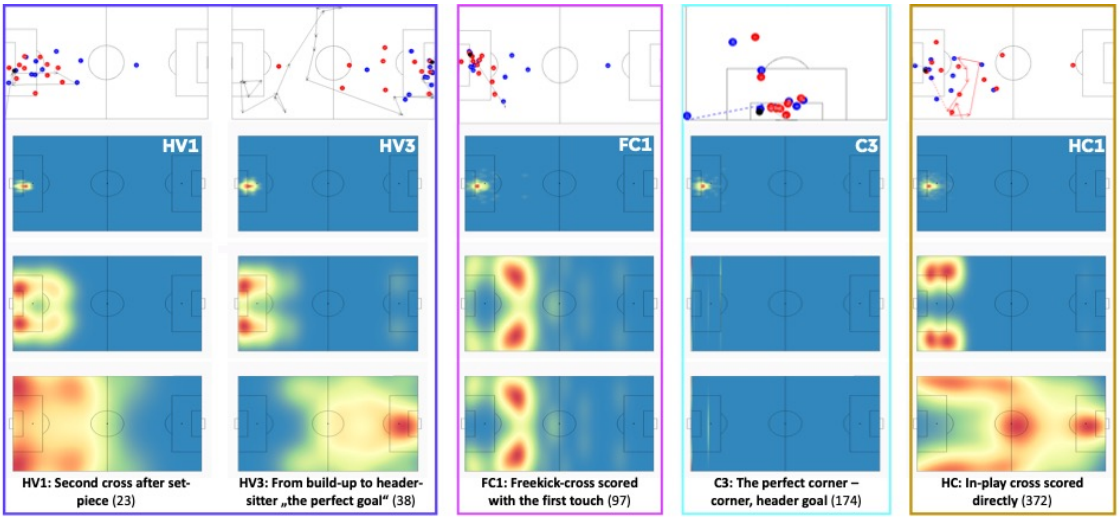

Figure 4: Visualizations of goals scored by headers: example goal (top row), scorer position (2nd row), assist location (3nd row), begin attacking phase (bottom row).

3.4 Shots

Possibly the most important part of a goal is the shot itself. We differentiate between leg-shots, volleys, and headers. From our 3,457 goals, 83.0% are scored by a non-volley leg-shot. The remaining 17.0% are either headers or volleys and exemplified in Figure 4. Surprisingly, more than the half of these goals originate from open-play phases (**HV3** and **HC**). For instance, Clusters **SO2** and **SO1**, displayed in Figure 8, contain lovely volley rebounds.

Headers are the predominant way to score after freekicks (74.0%) and corners (87.0%) but play only a minor role in ping-pong situations (**C1** and **FC2**). Cluster **HV2** for example contains spectacular headers, some of which are also highlighted as triangles in Figure 6.

Cluster **HV1** and **HV3** are efficient ways of scoring with conversion rates of 26.7% and 31.0%, respectively. As mentioned above, **HV3** encapsulates a blueprint worth striving for. Cluster **HV1** shows another constellation: Either a freekick or a corner is cleared by the opponent followed by a second cross into the box that is then converted in a goal. From the overall 32,406 shots, 5,612 headers and volleys led to 616 goals (11.0%) which is slightly more efficient than leg-shots (10.7%).

3.5 Patterns

Many clusters encode strategic patterns or tricks and by discovering the next opponent's strategies one can increase the likelihood of winning. Some of these strategies can be seen in Cluster **SA** where strikers cross the line of the shot (likely) on purpose as well as in **HA1** with header flick-ons. From the perspective of a goalkeeper it is crucial to know the locations of freekicks, direct shots as well as crosses into the box. Figure 6 thus shows the locations of successful long-distance shots depending on the cluster.

The most basic tactical pattern in football is a *one-two* and encoded by cluster **OT**. Figure 6 visualizes a nice example of this pattern.

Cluster **K&R** represents the *kick-and-rush* strategy. Goals in this cluster are characterized by a long-distance pass to the scorer. These passes bridge on average 48.47 m, and are often difficult to control.

Finally, a cluster containing special *corner-tricks* is **C5**. Clearly, knowing whether the next opponents have some corner-tricks in their portfolio is an important piece of information for every coaching staff.

4 Discussion

Analyzing the origin of goals is often limited to small sample sizes due to manual annotation, nevertheless, studies breaking down scoring patterns are common in sport-science literature (Reep et al., 1968; Njororai, 2013; Mitrotasios et al., 2012). Exploiting the availability of positional and event data can present a change in paradigm for pattern analysis in football. The automated analysis based on 3,457 goals, allows us to put results from recent literature on a sound base: With 64.0% goals scored from open-play, we present a lower number than previous literature (e.g. Njororai (2013) 75.86% of 145 from several competitions; Mitrotasios et al. (2012) 72.4% of 76 goals European championship). Njororai (2013) claimed that history

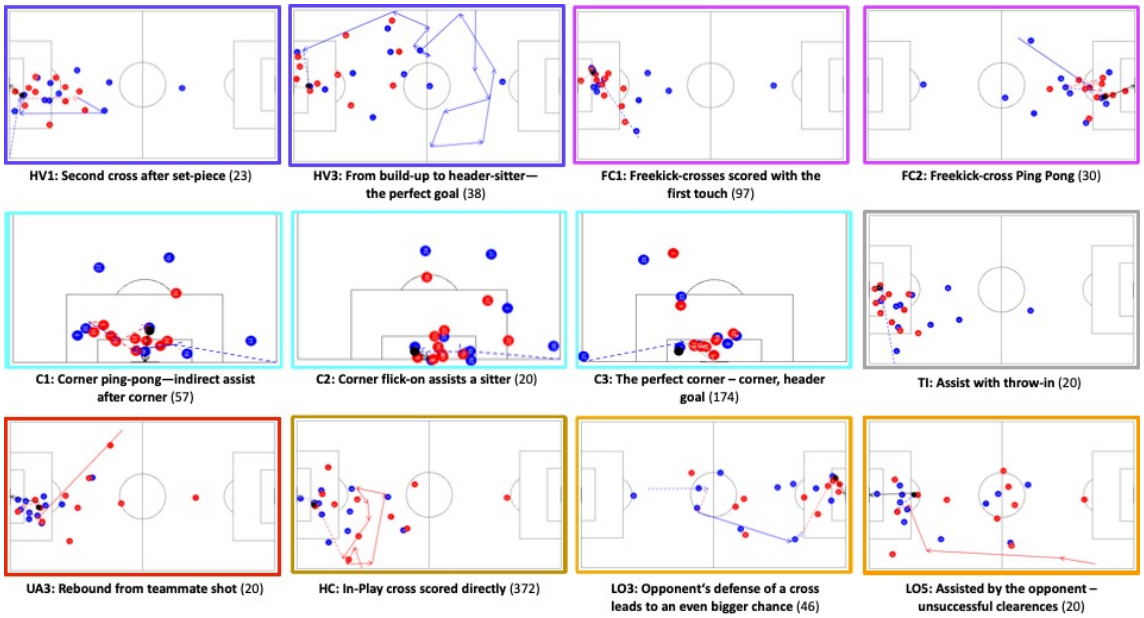

Figure 5: Exemplary visualizations of all goals occurring by corners and crosses. Both goals from set-pieces and from open-play are included representing a total amount of 490 goals (14.17%).

showed a trend towards more open-play goals, which cannot be confirmed by our data-set. Goals occurring from open-play follow and attacking phase with 3.6 passes on average and the average conversion rate of shots is 11.67% roughly in line with the original findings from Reep et al., 1968, and later confirmed by Collet (2013); Sarmento et al. (2014); Vogelbein et al. (2014); González-Ródenas et al. (2019). Another insight regarding set-pieces, is the lower header-rate of freekicks (74.0%) compared to corners (87.0%), which can be explained by the additional space behind the offside line, increasing the likelihood of creating enough separation to finish the cross with the foot. However, the definition which goal still counts as a converted set piece or when a possession phase starts, varies across the literature, making a comparison between the results difficult—data-driven studies like the one presented here could overcome this issue by using consistent definitions, without the need for manual annotations.

Nevertheless, the key benefit of our approach is not the ability to conduct a large-scale descriptive analysis, but rather to use a hierarchical clustering in order to identify patterns in the origin of goals automatically and, consequently, to derive meaningful insights for football practitioners from these patterns. The efficiency of fast ball regains followed by a successful offensive action has been investigated in several studies (Reep et al., 1968; Hobbs et al., 2018; Vogelbein et al., 2014). Our clustering detects that strategy as a pattern represented in its own cluster in **SP1** (3.7% of all goals). Another useful insight are particularly high conversion rates of ball-gains after high-blocks (**SP5**, 40.0%), especially in comparison with ball gains after counterpressing (**SP1**, 5.0%) and after mid-blocks (**SP2**, 2.0%). This finding regarding the efficiency of counterpressing is in line with Bauer et al. (2021). In the latter case when the ball is won, the defense is typically quite well organized with many players behind the ball, often leading to long-range shots. Compared to the usual categorization into corners, direct freekicks, freekick-crosses and penalties, our approach allows for a much granular view of set-pieces and discloses hidden insights. Cluster **HA1** and **C2**, contain several flick-on goals after corner kicks and confirm the relevance of this sub-category of corner goals presented in Power et al., 2017. After set-pieces (**SO2**) and after open-play **SO1** a significant amount of goals were scored through volley rebounds. The high total amount of 2% of all goals, even surprised the experts. Training shot techniques is a crucial part in professional football, and our analysis can help to identify the right shooting situations to focus on.

In the following we describe four exemplary use cases of how the insights can support analysts and coaches in their everyday business:

Use case 1—Automatize and objectify the match-analysts weekly processes: Nowadays, spending vast amounts of time and resources to perform pre-match-analyses of the next opponents and on the post-match-analyses of the own performance has become an integral part of professional football. In a well established process, match-analysts spend hours observing video footage of their upcoming opponent to figure out what to expect. One of the most crucial questions they need to answer is: how does the

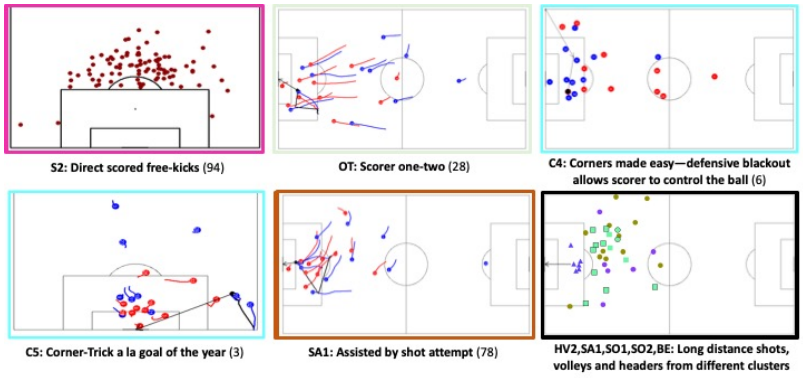

Figure 6: Selected special goals. Top left: shot chart for **S2**. Bottom right: aggregation of extraordinary shots (circles), volleys (squares), and headers (triangles) from different clusters: shots are plotted as circles, volleys as quarters and headers as triangles - all in the respective cluster-color.

opponent score and concede goals? Typically, time constraints allow only to examine the last few goals from the opponent which are then classified into one of few categories. These categories vary from club to club, but due to the sample size, the analysis is coarse and expressivity is limited. By contrast, our fully automated and purely data-driven approach processes arbitrarily long periods and as many goals as desired and provides detailed clusters that allow for fine-grained analyses. Throw-in crosses, for example, are rare events and thereafter hard to scout for each upcoming opponent. But some teams actively practice throw-in crosses[6,7] and our clustering automatically discloses whether an opponent uses them effectively. Our analysis shows that almost half of the goals after long throw-ins are scored by only three teams in our data-set (Union Berlin, Dynamo Dresden, MSV Duisburg). Our clustering also reveals teams with a distinct counterpressing strategy (RB Leipzig scored twice as often with **SP1** as the runner-up), teams with dedicated cornertricks (Arminia Bielefeld with several goals in **C5**), and especially successful teams after kick and rush plays **K&R** (TSG Hoffenheim, Bayer 04 Leverkusen and Fortuna Düsseldorf).

Use case 2—Scouting players: Scouting prospective players who will quickly adapt to a teams' playing-style or identifying a (near) equal substitute for a leaving or injured player is key to running a professional club (**Radicchi2016a**; Pappalardo, 2019). While there already exist many different approaches using event and/or tracking data, aiming to objectively evaluate players for scouting purposes like expected goals (Anzer et al., 2021), space-control (Fernandez et al., 2018) or expected possession values (Spearman, 2018; Fernández et al., 2019), these typically only quantify a player's output. By looking at patterns instead of the pure outcome, our approach presents a possibility to identify players that not only produce a high output, but do it in a way that fits a team stylistically.. Figure 9 shows the footprints of the two famous strikers (Robert Lewandowski and Timo Werner) where line widths are proportional to the number of scored goals in the respective branch of the tree. To evaluate whether one could substitute another, we let the data speak and compare their scoring footprints. Since both are strikers, their performance is measured to a high degree by the number of goals scored per match and the data-driven footprints reveal whether they score their goals in similar fashions. Another non-trivial aspect in scouting players is to identify promising talents. If, for instance, a technically skilled talent is needed, an aggregated view on clusters **SA1**, **SA2** and **BE** is helpful as they solely contain goals that require technically skilled players to score. Another data-driven approach to quantify fingerprints of players is presented in Marcelino et al. (2020). They analyze the directional correlations between players' movements and find that players whose movement correlates more strongly with their teammates' tend to have higher market values. Following Marcelino et al., 2020 in future studies one could evaluate the connection between players' footprints and their market value.

Use case 3—Long term team analysis: Analyzing the dendrogram allows to join clusters that encode semantically similar goals. As an example consider clusters **SP1** and **LO2**. The former contains classical counterattacks where the ball is gained and quickly carried forward with determined passes. The latter, located at a very different branch of the dendrogram, contains similar situations but the (unintentional)

[6] `https://www.bbc.com/sport/football/46312234`, accessed 06/28/2020

[7] `https://trainingground.guru/articles/leeds-hire-set-piece-specialist-gianni-vio`, accessed 06/28/2020

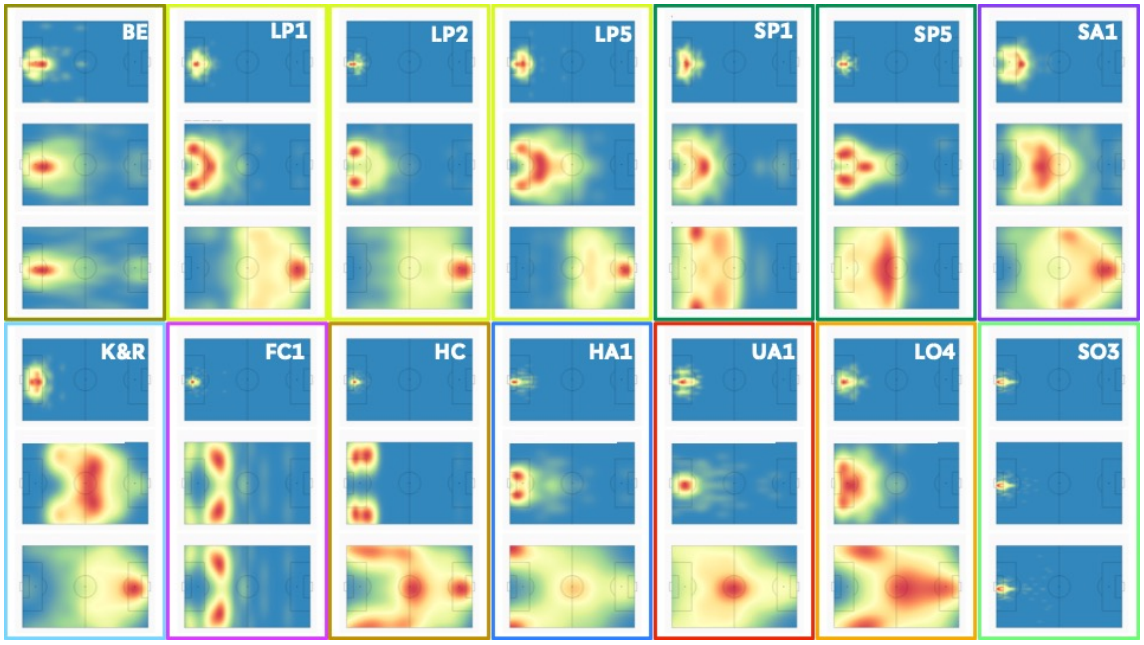

Figure 7: The largest 14 in-play clusters shown by three heatmaps each: start of the ball possession phase (bottom), the assist location (middle) such as the shot location (top).

assist comes from the opponent. Merging the clusters allows to reason about counterattacks in general. A very traditional category found by our clustering are one-two's **OT**. While this pattern is a basic element of football, its scarcity leading up to goals (0.8% of all goals) meant, it has not been investigated by any scientific study. While one-two's are very effective against men-oriented defensive structures, their relevance in today's top leagues seems to shrink significantly. However, are able to detect teams or players using this strategy more frequently.

Use case 4—Scouting coaches: Moreover, the clustering allows to shed light on many very different aspects of teams such as the effects of replacing head coaches. While selecting the right coach is a crucial decision for any club, doing so while making use of positional or event data to support this decision has not been addressed in literature. Just like players, coaches leave their own footprint in the dendrogram. Analyzing this footprint can be a massive support when identifying head-coaches with a playing style suiting their potential new team. Several studies investigated the effect coaching changes had on team results (Kattuman et al., 2019; Besters et al., 2016), but our method aims to show before a possible change how a coach would fit stylistically.

Professional football is highly affected by competitional pressure and emotions. Having an objective and unbiased view on a team's performance is indispensable for long-term success. By following a purely data-driven approach, our contribution allows for such an unbiased view on the origin of goals. In order to overcome biases towards established patterns, we present an exploratory way of analyzing goal scoring patterns. In future research, the gained insights can be build upon to train supervised machine learning models that automatically classify goals into pre-defined classes depending on individual club philosophies. The possibility of (partially) automating regular tasks (e.g. weekly opponent analysis) not only allows to save time but also to put the human focus on more sophisticated analyses and leave the easy tasks to number crunching machines. Compared to human analyses, the proposed clustering offers a finer granularity and, hence, provides a deeper understanding of the origin of goals.

The resulting clustering tree was analyzed and sanity checked by professional match-analysts from national teams and Bundesliga clubs. The interdisciplinary cooperation with domain experts was of utmost importance to the project to bridge the gap between computer and sports science and practice (see also Goes et al., 2020; Herold et al., 2019; Rein et al., 2016). Combining expert opinions with statistical evaluations (i.e. the Silhouette value of the clustering) turned out to be very beneficial for determining the number of clusters.

When naming the clusters and discussing the ideal cluster number, in most cases the experts immediately agreed and in the few remaining cases after a brief discussion a conses was found. Nevertheless, a more systematic evaluation would be desirable for future studies. Since many of the categorical features are

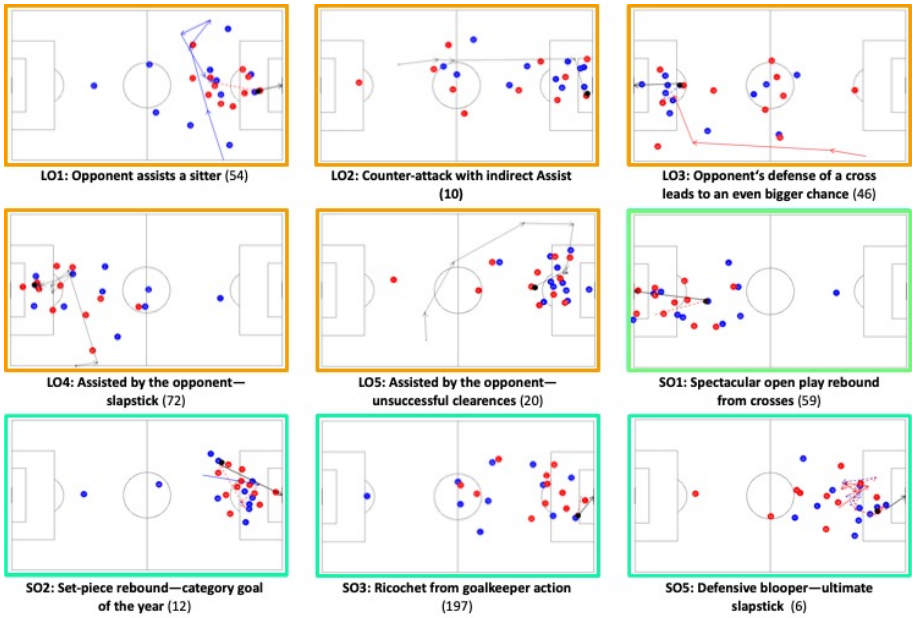

Figure 8: Visualizations of goals assisted by the opponent. The 278 goals in this category present 7.8% of all goals.

derived from manually annotated event data, further studies could also analyze the inter-labeler reliability of the features. Furthermore, a general limitation of an unsupervised learning task is, that the resulting clusters are not guaranteed to to make the distinctions a human would make. While, we used experts opinions to guide us to find the right clustering, and the results satisfied their expectations, it could be of future interest, to investigate how closely this unsupervised clustering matches an experts clustering.

Besides the reliability of the event data, an improvement of the tracking data quality (e.g. through limb tracking), could open avenues for even more granular analysis of goals. And, while the data set used for this study is already one of the largest in the literature, increasing the number of considered goals would certainly further increase the usefulness of this work (e.g. by identifying very rare types of goals, like direct corner kick goals). As mentioned earlier, we are excluding own-goals from this analysis, but investigating how they originate, and what they have in common with "typical" goals could be another area to explore further in the future.

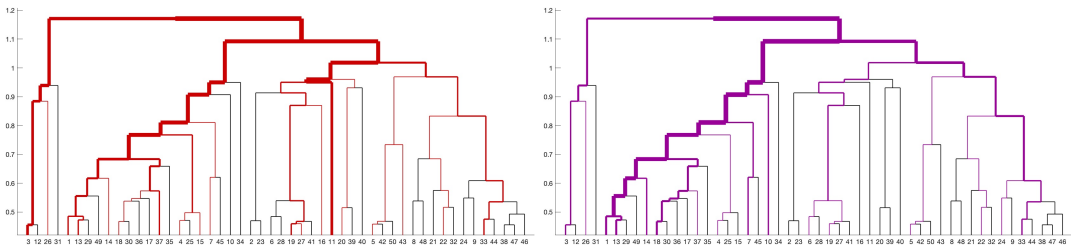

Figure 9: Footprints of Robert Lewandowski (left) and Timo Werner (right) in the dendrogram.

5 Conclusions

We studied the origin of goals in the German Bundesliga and 2nd Bundesliga. We proposed a rich set of features that can be extracted from synchronized tracking and event data. The feature representations of the goals were then processed by an agglomerative clustering algorithm. Using two entire seasons of data, we showed that the clustering allowed for fine grained differentiations and non-obvious insights that are approved by professional match-analysts working for national teams and Bundesliga clubs. Our approach can support professionals in their daily work and renders manual inspection of large amounts of video footage unnecessary. Moreover, the proposed clustering can objectify pivotal decision making and offers quantitative solutions to traditionally qualitative domains like scouting players or coaches or analyzing the

next opponent.

6 Acknowledgements

This work would not have been possible without the perspective of professional match-analysts from world class teams who helped us to define relevant features and spend much time evaluating (intermediate) results. We would cordially like to thank Dr. Stephan Nopp and Christofer Clemens (head match-analysts of the German mens National team), Jannis Scheibe (head match-analyst of the German U21 mens National team) as well as Sebastian Geißler (former match-analyst of Borussia Mönchengladbach). Additionally, the authors would like to thank Dr. Hendrik Weber and Deutsche Fußball Liga

Disclosure Statement The authors report no conflict of interest.

Ethics and Data Sharing By informing all participating players, all tracking is compliant to the general data protection regulation (GDPR)[8]. An ethics approval for wider research program using the respective data is authorized by the ethics committee of the Faculty of Economics and Social Sciences at the University of Tübingen. In order to respect the player's and club's sensitive information, the data cannot be shared public.

Additional Material (Confidential) We provide a video with representative goals for each cluster.[9]

References

Andrienko, Gennady et al. (2017). "Visual analysis of pressure in football". In: *Data Mining and Knowledge Discovery* 31.6, pp. 1793–1839. ISSN: 1573756X. DOI: 10.1007/s10618-017-0513-2 (cit. on pp. 3, 18, 19).

Andrienko, Gennady et al. (2019). "Constructing Spaces and Times for Tactical Analysis in Football". In: *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1. ISSN: 1077-2626. DOI: 10.1109/tvcg.2019.2952129. URL: https://ieeexplore.ieee.org/document/8894420/ (cit. on p. 2).

Anzer, Gabriel & Pascal Bauer (2021). "A Goal Scoring Probability Model based on Synchronized Positional and Event Data". In: *Frontiers in Sports and Active Learning (Special Issue: Using Artificial Intelligence to Enhance Sport Performance)* 3.0, pp. 1–18. DOI: 10.3389/fspor.2021.624475. URL: https://www.frontiersin.org/articles/10.3389/fspor.2021.624475/full (cit. on pp. 2, 3, 10, 18, 19).

Armatas, Vasilios, Athanasios Yiannakos, & Dimitris Hatzimanouil (2007). "Record and evaluation of set-plays in european football championship in Portugal 2004". In: *Inquiries in Sport and Physical Education* (cit. on p. 1).

Bauer, Pascal & Gabriel Anzer (2021). "Data-driven detection of counterpressing in professional football—A supervised machine learning task based on synchronized positional and event data with expert-based feature extraction". In: *Data Mining and Knowledge Discovery* 35.5, pp. 2009–2049. ISSN: 1573-756X. DOI: 10.1007/s10618-021-00763-7. URL: https://link.springer.com/article/10.1007/s10618-021-00763-7 (cit. on p. 9).

Besters, Lucas M., Jan C. van Ours, & Martin A. van Tuijl (2016). "Effectiveness of In-Season Manager Changes in English Premier League Football". In: *Economist (Netherlands)* 164.3, pp. 335–356. ISSN: 15729982. DOI: 10.1007/s10645-016-9277-0 (cit. on p. 11).

Brian, S (2011). *Cluster analysis Brian S. Everitt ... [et al.]* John Wiley & Sons, XII, 330 p. ill. ISBN: 978-0-470-74991-3 (cit. on p. 4).

Carling, Christopher, A. Mark Williams, & Thomas Reilly (2006). "Handbook of Soccer Match Analysis: A Systematic Approach to Improving Performance". In: *Journal of Sports Science & Medicine*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3818670/ (cit. on p. 1).

[8] https://gdpr-info.eu/

[9] https://bit.ly/2NAXQcW.

Casal, C. A. et al. (2017). "Influence of match status on corner kicks tactics in elite soccer". In: *Revista Internacional de Medicina y Ciencias de la Actividad Fisica y del Deporte*. ISSN: 1577-0354. DOI: 10.15366/rimcafd2017.68.009. URL: http://cdeporte.rediris.es/revista/revista68/artinfluencia851e.pdf (cit. on p. 1).

Casal, Claudio A. et al. (2015). "Analysis of corner kick success in elite football". In: *International Journal of Performance Analysis in Sport* 15.2, pp. 430–451. ISSN: 14748185. DOI: 10.1080/24748668.2015.11868805 (cit. on p. 1).

Castellano, Julen, David Casamichana, & Carlos Lago (2012). "Accepted for printing in". In: *Journal of Human Kinetics* 31, pp. 139–147. DOI: 10.2478/v10078-012-0015-7. URL: http://fifa.com/worldcup/index.html (cit. on p. 2).

Chawla, Sanjay et al. (2017). "Classification of passes in football matches using spatiotemporal data". In: *ACM Transactions on Spatial Algorithms and Systems* 3.2. ISSN: 23740361. DOI: 10.1145/3105576 (cit. on p. 2).

Collet, Christian (2013). "The possession game? A comparative analysis of ball retention and team success in European and international football, 2007-2010". In: *Journal of Sports Sciences* 31.2, pp. 123–136. ISSN: 02640414. DOI: 10.1080/02640414.2012.727455 (cit. on p. 9).

Defays, D. (2015). "An efficient algorithm for a complete link method". In: URL: https://academic.oup.com/comjnl/article-abstract/20/4/364/393966 (cit. on p. 4).

Delgado-Bordonau, Juan Luis et al. (2013). "Offensive and defensive team performance: Relation to successful and unsuccessful participation in the 2010 Soccer World Cup". In: *Journal of Human Sport and Exercise* 8.4, pp. 894–904. ISSN: 19885202. DOI: 10.4100/jhse.2013.84.02 (cit. on p. 2).

Dufour, Michel, John Phillips, & Viviane Ernwein (2017). "What makes the difference? Analysis of the 2014 World Cup". In: *Journal of Human Sport and Exercise* 12.3, pp. 616–629. ISSN: 19885202. DOI: 10.14198/jhse.2017.123.06 (cit. on p. 2).

Erčulj, Frane & Erik Štrumbelj (2015). "Basketball shot types and shot success in different levels of competitive basketball". In: *PLoS ONE* 10.6, pp. 1–14. ISSN: 19326203. DOI: 10.1371/journal.pone.0128885 (cit. on p. 2).

Felsenstein, Joseph (1996). "[24] Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods". In: *Methods in Enzymology* 266, pp. 418–427. ISSN: 00766879. DOI: 10.1016/s0076-6879(96)66026-1 (cit. on p. 4).

Fernández-Hermógenes, Daniel, Oleguer Camerino, & Antonio García De Alcaraz (2017). "Set-piece offensive plays in soccer". In: ISSN: 2014-0983. DOI: 10.5672/apunts.2014-0983.es.(2017/3).129.06. URL: https://core.ac.uk/download/pdf/132357632.pdf (cit. on p. 1).

Fernandez, Javier & Luke Bornn (2018). "Wide Open Spaces : A statistical technique for measuring space creation in professional soccer". In: *MIT Sloan Sports Analytics Conference*, pp. 1–19 (cit. on p. 10).

Fernández, Javier, Luke Bornn, & Dan Cervone (2019). "Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer". In: *MIT Sloan Sports Analytics Conference, Boston (USA)*, pp. 1–18. URL: https://lukebornn.com/sloan_epv_curve.mp4 (cit. on p. 10).

Fernando, T et al. (2015). "Discovering Methods of Scoring in Soccer Using Tracking Data". In: *KDD Workshop on Large-Scale Sports Analytics*, pp. 1–4. URL: https://large-scale-sports-analytics.org/Large-Scale-Sports-Analytics/Submissions2015_files/paperID19-Tharindu.pdf (cit. on p. 2).

Goes, F R et al. (2020). "Unlocking the Potential of Big Data to Support Tactical Performance Analysis in Professional Soccer: A Systematic Review". In: *European Journal of Sport Science* 0.0, pp. 1–16. ISSN: 1746-1391. DOI: 10.1080/17461391.2020.1747552. URL: https://doi.org/10.1080/17461391.2020.1747552 (cit. on pp. 2, 11).

González-Ródenas, Joaquin et al. (2019). "Technical, tactical and spatial indicators related to goal scoring in European elite soccer". In: *Journal of Human Sport and Exercise*. ISSN: 1988-5202. DOI: 10.14198/jhse.2020.151.17 (cit. on pp. 1, 9).

Göral, Kemal (2019). "The importance of set-pieces in soccer : Russia 2018 FIFA World Cup analysis Futbolda duran t opların ö nemi : Rusya 2018 FIFA Dünya Kupasının analizi". In: 16.3. DOI: 10.14687/jhs.v16i3.5758 (cit. on p. 1).

Herold, Mat et al. (2019). "Machine learning in men's professional football: Current applications and future directions for improving attacking play". In: *International Journal of Sports Science & Coaching*, p. 1747954119879350. ISSN: 1747-9541. DOI: 10.1177/1747954119879350. URL: https://doi.org/10.1177/1747954119879350 (cit. on pp. 2, 11).

Hobbs, Jennifer et al. (2018). "Quantifying the Value of Transitions in Soccer via Spatiotemporal Trajectory Clustering". In: pp. 1–11 (cit. on pp. 2, 9).

Kattuman, Paul, Christoph Loch, & Charlotte Kurchian (2019). "Management succession and success in a professional soccer team". In: *PLoS ONE* 14.3, pp. 1–20. ISSN: 19326203. DOI: 10.1371/journal.pone.0212634 (cit. on p. 11).

Linke, Daniel, Daniel Link, & Martin Lames (2020). "Football-specific validity of TRACAB's optical video tracking systems". In: *PLoS ONE* 15.3, pp. 1–17. ISSN: 19326203. DOI: 10.1371/journal.pone.0230179. URL: http://dx.doi.org/10.1371/journal.pone.0230179 (cit. on p. 3).

López, F. A., J. A. Martínez, & M. Ruiz (2013). "Spatial pattern analysis of shot attempts in basketball; The case of L.A. Lakers". In: *Revista Internacional de Medicina y Ciencias de la Actividad Fisica y del Deporte* 13.51, pp. 585–613. ISSN: 1577-0354 (cit. on p. 2).

Lucey, Patrick et al. (2014). ""Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data". In: *Proc. 8th Annual MIT Sloan Sports Analytics Conference*, pp. 1–9. URL: http://www.sloansportsconference.com/?p=15790 (cit. on p. 2).

Marcelino, Rui et al. (2020). "Collective movement analysis reveals coordination tactics of team players in football matches". In: *Chaos, Solitons and Fractals* 138, p. 109831. ISSN: 09600779. DOI: 10.1016/j.chaos.2020.109831. URL: https://doi.org/10.1016/j.chaos.2020.109831 (cit. on pp. 2, 10).

Merlin, Murilo et al. (2020). "Exploring the determinants of success in different clusters of ball possession sequences in soccer". In: *Research in Sports Medicine* 28.3, pp. 339–350. ISSN: 15438635. DOI: 10.1080/15438627.2020.1716228 (cit. on pp. 2, 3).

Mitrotasios, Michalis & Vasilis Armatas (2012). "Analysis of goal scoring patterns in the 2012 European Football Championship". In: *The Sport Journal* 50, pp. 1–9. ISSN: 15439518. URL: http://thesportjournal.org/article/analysis-of-goal-scoring-patterns-in-the-2012-european-football-championship/ (cit. on pp. 1, 8).

Murtagh, Fionn & Pedro Contreras (2017). *Algorithms for Hierarchical Clustering: An Overview, II*. Tech. rep. (cit. on pp. 2, 4).

Njororai, W. W.S. (2013). "Analysis of goals scored in the 2010 world cup soccer tournament held in South Africa". In: *Journal of Physical Education and Sport*. ISSN: 22478051. DOI: 10.7752/jpes.2013.01002. URL: https://scholarworks.uttyler.edu/cgi/viewcontent.cgi?referer=https://scholar.google.de/scholar?hl=de&as_sdt=0%2C5&q=Analysis+of+goals+scored+in+the+2010+world+cup+soccer+tournament+held+in+South+Africa&btnG=&httpsredir=1&article=1008&context=hkdept_fac (cit. on pp. 1, 2, 8).

Pappalardo, Luca (2019). "Explainable Injury Forecasting in Soccer via Multivariate Time Series and Convolutional Neural Networks". In: *Barça sports analytics summit, Barcelona (Spain)*, pp. 1–15. URL: https://static.capabiliaserver.com/frontend/clients/barca/wp_prod/wp-content/uploads/2020/01/c6658839-paper-format-luca-pappalardo-1.pdf (cit. on p. 10).

Plummer, B T (2013). "Analysis of Attacking Possessions Leading to a Goal Attempt, and Goal Scoring Patterns within Men's Elite Soccer". In: *Journal of Sports Science* (cit. on p. 1).

Pollard, Richard & Charles Reep (1997). "Measuring the effectiveness of playing strategies at soccer". In: *Journal of the Royal Statistical Society Series D: The Statistician*. ISSN: 00390526. DOI: 10.1111/1467-9884.00108 (cit. on p. 2).

Power, Paul et al. (2017). "Not all passes are created equal: Objectively measuring the riskt and reward of passes in soccer from tracking data". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1605–1613. DOI: 10.1145/3097983.3098051. URL: http://doi.acm.org/10.1145/3097983.3098051 (cit. on pp. 1, 9).

Pulling, Craig (2015). "Long corner kicks in the English premier league: Deliveries into the goal area and critical area". In: *Kinesiology* 47.2, pp. 193–201. ISSN: 13311441 (cit. on p. 1).

Pulling, Craig, Matthew Robins, & Thomas Rixon (2013). "Defending corner kicks: Analysis from the English premier league". In: *International Journal of Performance Analysis in Sport*. ISSN: 14748185. DOI: 10.1080/24748668.2013.11868637 (cit. on p. 1).

Rathke, Alex (2017). "An examination of expected goals and shot efficiency in soccer". In: *Journal of Human Sport and Exercise* 12.Proc2. ISSN: 1988-5202. DOI: 10.14198/jhse.2017.12.proc2.05. URL: http://www.redalyc.org/articulo.oa?id=301052437005 (cit. on p. 2).

Reep, Charles & B. Benjamin (1968). "Skill and Chance in Association Football". In: *Journal of the Royal Statistical Society. Series A (General)*. ISSN: 00359238. DOI: 10.2307/2343726 (cit. on pp. 1, 8, 9).

Reich, Brian J. et al. (2006). "A spatial analysis of basketball shot chart data". In: *American Statistician* 60.1, pp. 3–12. ISSN: 00031305. DOI: 10.1198/000313006X90305 (cit. on p. 2).

Rein, Robert & Daniel Memmert (2016). "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science". In: *SpringerPlus* 5.1. ISSN: 21931801. DOI: 10.1186/s40064-016-3108-2 (cit. on pp. 2, 11).

Robberechts, Pieter & Jesse Davis (2020). "How data availability affects the ability to learn good xG models". In: *Communications in Computer and Information Science, Springer, Cham* 1324, pp. 17–27. ISSN: 18650937. DOI: 10.1007/978-3-030-64912-8 (cit. on p. 2).

Rousseeuw, Peter J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20.C, pp. 53–65. ISSN: 03770427. DOI: 10.1016/0377-0427(87)90125-7 (cit. on p. 4).

Ruiz, H. et al. (2015). "Measuring scoring efficiency through goal expectancy estimation". In: *23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015 - Proceedings* April, pp. 149–154 (cit. on p. 2).

Ruiz, Hector et al. (2017). ""The Leicester City Fairytale?": Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 EPL Seasons". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1991–2000. DOI: 10.1145/3097983.3098121. URL: http://doi.acm.org/10.1145/3097983.3098121%0Ahttp://dl.acm.org/citation.cfm?doid=3097983.3098121 (cit. on p. 2).

Salmon, Paul M. & Scott McLean (2020). "Complexity in the beautiful game: implications for football research and practice". In: *Science and Medicine in Football* 4.2, pp. 162–167. ISSN: 24734446. DOI: 10.1080/24733938.2019.1699247 (cit. on p. 2).

Santos, Alejandro Benito et al. (2018). "Data-driven visual performance analysis in soccer: An exploratory prototype". In: *Frontiers in Psychology* 9.DEC. ISSN: 16641078. DOI: 10.3389/fpsyg.2018.02416 (cit. on pp. 3, 18).

Sarmento, Hugo et al. (2014). "Patterns of play in the counterattack of elite football teams - A mixed method approach". In: *International Journal of Performance Analysis in Sport*. ISSN: 14748185. DOI: 10.1080/24748668.2014.11868731 (cit. on pp. 2, 9).

Schmicker, Robert H. (2013). "An application of satscan to evaluate the spatial distribution of corner kick goals in major league soccer". In: *International Journal of Computer Science in Sport*. ISSN: 16844769. URL: https://pdfs.semanticscholar.org/a36c/694c79c3d38d19baf9d01a3677834289b340.pdf (cit. on p. 1).

Schulze, Emiel et al. (2018). "Effects of positional variables on shooting outcome in elite football". In: *Science and Medicine in Football* 2.2, pp. 93–100. ISSN: 24734446. DOI: 10.1080/24733938.2017.1383628. URL: https://doi.org/10.1080/24733938.2017.1383628 (cit. on p. 2).

Sibson, R. (1973). "SLINK: An optimally efficient algorithm for the single-link cluster method". In: *The Computer Journal* 16.1, pp. 30–34. ISSN: 0010-4620. DOI: 10.1093/comjnl/16.1.30 (cit. on p. 4).

Siegle, Malte & Martin Lames (2013). "Modeling soccer by means of relative phase". In: *Journal of Systems Science and Complexity* 26.1, pp. 14–20. ISSN: 15597067. DOI: 10.1007/s11424-013-2283-2 (cit. on p. 2).

Sokal, C.D. Michener (1958). *A statistical method for evaluating systematic relationships*. URL: https://archive.org/details/cbarchive_33927_astatisticalmethodforevaluatin1902/page/n2/mode/2uphttp://archive.org/details/cbarchive_33927_astatisticalmethodforevaluatin1902 (cit. on p. 4).

Spearman, William (2018). "Beyond Expected Goals". In: *MIT Sloan Sports Analytics Conference, Boston (USA)*, pp. 1–17. URL: https://www.researchgate.net/publication/327139841 (cit. on p. 10).

Steiner, Silvan et al. (2019). "Outplaying opponents—a differential perspective on passes using position data". In: *German Journal of Exercise and Sport Research* February. ISSN: 2509-3142. DOI: 10.1007/s12662-019-00579-0 (cit. on p. 3).

Szwarc, Andrzej (2007). "Efficacy of successful and unsuccessful soccer teams taking part in finals of Champions League". In: *Research Yearbook* 13.2, pp. 221–225. URL: http://journals.indexcopernicus.com/abstracted.php?icid=838944 (cit. on p. 1).

Taberner, Matt et al. (2020). "Interchangeability of position tracking technologies; can we merge the data?" In: *Science and Medicine in Football* 4.1, pp. 76–81. ISSN: 24734446. DOI: 10.1080/24733938.2019.1634279. URL: https://doi.org/10.1080/24733938.2019.1634279 (cit. on p. 3).

Taylor, Joseph B, Nic James, & Stephen D Mellalieu (2005). "Notational Analysis of Corner Kicks in English Premier League Soccer". In: *Science and Football V*. URL: https://books.google.de/books?hl=de&lr=&id=nyFr-2uwPGoC&oi=fnd&pg=PA229&dq=Notational+Analysis+of+Corner+Kicks+in+English+Premier+League+Soccer&ots=DYs2PvFfk_&sig=W-i76h0Zxw_omd2Ty18Nu-giF6w#v=onepage&q=NotationalAnalysisofCornerKicksinEnglishPremi (cit. on p. 1).

Tenga, Albin et al. (2010). "Effect of playing tactics on goal scoring in norwegian professional soccer". In: *Journal of Sports Sciences* 28.3, pp. 237–244. ISSN: 1466447X. DOI: 10.1080/02640410903502774 (cit. on p. 2).

Vogelbein, Martin, Stephan Nopp, & Anita Hökelmann (2014). "Defensive transition in soccer - are prompt possession regains a measure of success? A quantitative analysis of German Fußball-Bundesliga 2010/2011". In: *Journal of Sports Sciences*. ISSN: 1466447X. DOI: 10.1080/02640414.2013.879671 (cit. on p. 9).

Witts, James (2019). "Training secrets of the world's greatest footballers : how science is transforming the modern game". In: *Bloomsbury Publishing PLC*. URL: https://www.bookdepository.com/Training-Secrets-Worlds-Greatest-Footballers-James-Witts/9781472948458 (cit. on p. 1).

Wold, Svante, Kim Esbensen, & Paul Geladi (1987). "Principal component analysis". In: *Chemometrics and Intelligent Laboratory Systems* 2.1-3, pp. 37–52. ISSN: 01697439. DOI: 10.1016/0169-7439(87)80084-9 (cit. on p. 4).

Appendix

Appendix A (Tables 1, 2) detail the features that are extracted from positional and event data for the clustering in detail. Appendix B (Tables 3, 4) provide details on goal scoring and receiving patterns on a club level and may be of interest to analysts of the respective teams. Similarly, Appendix C (Table 5) shows the conversion rates of every cluster. Finally, Appendix D (Table 6) contains representative goals for selected clusters. Interested analysts may use these goals to evaluate the clustering on their own.

A Appendix

Table 1: Features describing the ball possession phase prior to a goal.

Feature	Value	Description
Start Action	Categorical	Describes the start of the ball possession phase in pre-defined abstraction levels *(Own Half, Offensive Ball Gain, Throw in, corner kick, Free kick, Penalty)*.
Build-up	Categorical	Describing the build-upleading up to the goal *(crossOpenPlay, pass open play, free kick, Penalty, corner kick, throw in, loss Of Possession)*.
Location of ball possession start	Numeric	x- and y-coordinate of a shot. The synchronized location from positional and event data as described in Anzer et al., 2021 is used.
Set-up Origin	Categorical	Describes where the build-up play for the shot at goal starts *(inside, outside)*.
Duration ball possession phase	Numeric	Length of the ball possession phase measured in [s]. The start of a ball possession phase is either a dead ball, or an open-play turnover. Interruptions where the opposing team gains possession of the ball for less than six consecutive seconds do not end an possession phase.
Number of passes	Numeric	Number of completed passes during ball possession phase.
Number of opposing touches	Numeric	Number of opposing touches during ball possession phase.
Bypassed players	Numeric	Bypassed players is defined as the positive difference between the number of players that are closer to their own goal than the ball at the time of the shot and when the ball possession started.
Meters dribbled	Numeric	Meters dribbled during ball possession phase. This feature is calculated as the sum of all the euclidean distances between starting and end location of each player's possessions.
Meters passed	Numeric	Meters passed during ball possession phase.
Average passing pressure	Numeric	Average amount of pressure passing players received during the ball possession phase at the moment they played a pass according Andrienko et al., 2017.
Average receiving pressure	Numeric	Average amount of pressure pass receiving players received during the ball possession phase. at the moment they received a pass according Andrienko et al., 2017.
Counterattack	Categorical	Describes whether the build-up was a counterattack. Counterattacks are defined in the official manually collected event data as attacks during which a team gains ball control in its own half, immediately starts a quick counterattack and takes a shot within at maximum 14 seconds.
Number of opposing touches	Numerical	Counts the amount of uncontrolled touches the opposing team had during a possession.
Maximum vertical pass length	Numeric	The longest vertical distance of any pass within the possession.
Maximum horizontal pass length	Numeric	The longest horizontal distance of any pass within the possession.
Maximum pass length	Numeric	The distance of the longest pass within the possession.
Compactness Ballgain	Numeric	Compactness of the attacking team (Santos et al., 2018) at the beginning of the ball possession phase.
Compactness Shot	Numeric	Compactness of the attacking team (Santos et al., 2018) at the time of the shot

Table 2: Features describing the assist and shot setup.

Feature	Value	Description
Shot location	Numeric	X and Y coordinate of a shot
Type of shot	Categorical	Describing the body part used for the shot (head or leg)
Chance evaluation	Categorical	Classifying the quality of a chance (chance, sitter)
Taker Ball-control	Categorical	Ball-control type describes the type of control the shot taker had prior to scoring. It includes the following categories: • "Direct" a shot with the first touch, unless the shot is considered a volley. • "Volley" a shot with the first touch and the ball did not touch the ground previously. • "Control - shot" a shot followed after a single touch to control the ball. • "Distance covered $<$ 10m" a shot following a short dribble (less than 10 meters). • "Distance covered $>$ 10m" a shot following a longer dribble (more than 10 meters). • "Set piece taker" a direct set-piece shot.
Setup	Categorical	Describes how the person taking the shot was set up (header, long pass from open play, other pass from open play, one two, cross from open play, shot, free kick; corner kick, throw in, teammate action, rebound woodwork).
xG	Numeric	The "Expected Goal" (xG) value of a shot according to Anzer et al., 2021.
Distance to goal	Numeric	Distance in meters between the location of the shot and the center of the opposing goal.
Goal angle	Numeric	Angle in radians between the location of the shot and the two posts of the opposing goal.
Speed of player taking the shot	Numeric	The speed in [km/h] the player attempting the shot was travelling at the time of the shot.
Pressure on the player taking the shot	Numeric	The amount of pressure the player attempting the shot was under at the time of the shot according to Andrienko et al. (2017).
Defenders in the line of the shot	Numeric	The number of defenders in the line of the shot, defined as the triangle between the shot location and the two goal posts.
Distance of the goalkeeper to the goal	Numeric	The distance in meters the goalkeeper between the goalkeeper and the center of the goal at the time of the shot.
Goalkeeper in the line of the shot	Numeric	Describes whether the goalkeeper is in the line of the shot or not.
Solo	Categorical	Solo indicates that a remarkable individual contribution (=solo) by the goalscorer lead to the successful shot.
Assist location	Numeric	X and Y coordinate of the assist
After free kick	Categorical	Indicates whether the goal followed a freekick.
Assist type	Categorical	Describing whether it was a direct, indirect assist or not assisted
Assist action	Categorical	Describing the assist action (e.g. "long pass")

B Appendix

Table 3: Scored goals per team.

Scored goals	Total	Penalties	Direct Free-Kicks	Intercepted Build-Up	Intended Assists with leg-finishes	Intended Assists with leg-finishes	Headers and Volleys	Solo Assist	Kick & Rush	One-Two	Free-kick-Crosses	Corners	Assisted by shot attempt	Header after Cross from open-play	Assisted with head	Assisted with Throw-In	Unintended Teammate Assist	Assisted by the opponent	Assisted by the opponent
	3,457	272	94	80	734	270	70	170	152	28	127	260	76	372	124	20	122	202	278
1. FSV Mainz 05	81	7	0	1	11	8	3	5	5	0	3	4	2	12	4	1	4	4	7
Borussia Dortmund	142	7	3	3	43	18	2	4	3	4	5	7	5	10	3	0	2	14	9
Bayer 04 Leverkusen	126	7	2	2	34	10	3	6	8	4	3	9	7	14	4	0	0	6	7
FC Bayern München	176	9	1	4	39	17	10	12	5	2	2	18	4	18	2	1	9	9	14
SV Werder Bremen	91	4	2	1	17	8	2	10	4	0	2	11	0	10	2	0	5	3	10
FC Augsburg	89	9	4	2	17	4	1	3	4	0	3	11	4	15	3	0	1	4	4
FC Schalke 04	86	17	2	1	15	5	0	5	5	2	5	6	5	3	3	0	1	5	6
Borussia Mönchengladbach	99	10	1	2	27	8	1	6	3	2	2	11	3	6	1	0	2	6	8
Eintracht Frankfurt	103	6	1	2	30	8	2	3	4	0	3	6	1	10	5	0	7	6	9
Hamburger SV	71	5	4	4	13	3	1	4	1	0	5	7	2	3	1	1	1	6	10
TSG 1899 Hoffenheim	133	10	3	3	31	7	2	8	8	2	4	7	4	13	2	1	7	10	11
VfL Wolfsburg	94	6	4	2	24	3	3	3	2	0	4	5	2	13	4	0	4	6	9
Sport-Club Freiburg	75	11	3	4	7	6	0	4	0	1	4	8	0	12	6	0	0	3	6
Hertha BSC	91	8	2	1	17	6	1	4	5	1	3	8	2	10	3	1	4	6	9
VfB Stuttgart	67	2	0	2	11	7	1	1	5	0	4	7	2	9	2	1	1	4	8
RB Leipzig	117	9	3	3	35	17	1	7	5	2	0	5	1	7	2	0	4	6	10
Fortuna Düsseldorf	105	9	1	2	19	9	1	9	9	1	4	6	0	15	3	1	2	5	9
1. FC Köln	115	8	2	4	21	4	5	2	5	0	7	11	2	22	6	1	2	7	6
1. FC Nürnberg	85	3	1	1	17	6	3	2	3	1	4	12	1	9	2	0	0	3	4
Hannover 96	74	7	2	0	15	3	0	3	4	0	3	8	2	9	4	0	0	1	8
MSV Duisburg	88	10	7	2	16	2	3	4	3	1	5	6	0	10	2	2	2	2	11
SG Dynamo Dresden	81	10	2	2	21	4	1	3	3	2	1	8	1	4	3	3	3	3	7
FC Ingolstadt 04	89	10	5	4	14	9	2	5	1	0	5	4	3	5	5	0	0	6	7
1. FC Heidenheim 1846	101	6	7	1	21	4	4	4	4	0	3	7	0	16	5	0	0	7	7
1. FC Union Berlin	99	8	1	2	16	9	2	5	1	0	3	9	2	10	4	3	5	9	10
Holstein Kiel	121	13	3	1	32	9	2	3	3	0	5	7	4	14	6	0	0	2	6	11
Eintracht Braunschweig	34	0	0	0	12	5	0	1	3	0	0	1	1	1	2	0	0	4	0	4
SSV Jahn Regensburg	103	7	4	3	19	7	1	7	5	2	5	5	3	12	5	0	0	4	8	6
DSC Arminia Bielefeld	99	8	5	5	16	10	0	5	3	0	6	5	1	15	7	1	1	4	5	3
SV Darmstadt 98	81	9	4	0	7	2	2	5	4	0	4	4	1	13	7	0	0	4	5	10
FC St. Pauli	81	6	2	2	16	8	2	3	5	0	4	4	2	8	1	1	1	4	7	6
FC Erzgebirge Aue	76	6	1	1	18	4	1	9	5	0	1	6	1	7	1	0	0	4	6	5
SpVgg Greuther Fürth	71	4	3	3	16	6	2	3	5	0	3	8	3	8	2	1	0	1	3
VfL Bochum 1848	84	6	1	1	18	8	1	3	5	0	3	3	3	9	3	0	0	6	7	7
SC Paderborn 07	73	3	4	4	23	9	3	4	2	0	2	4	0	4	0	0	0	3	3	5
SV Sandhausen	79	6	2	3	14	6	2	3	8	0	3	7	0	9	6	1	1	1	1	7
1. FC Magdeburg	34	2	2	0	8	5	0	2	2	1	2	2	1	1	2	0	1	2	1
1. FC Kaiserslautern	37	4	0	2	4	6	0	0	2	0	2	3	1	6	1	0	2	1	3

Table 4: Received goals per team.

Received goals	Total	Penalties	Direct Free-Kicks	Intercepted Build-Up	Intended Assists with leg-finishes	Intended Assists with leg-finishes	Headers and Volleys	Solo Assist	Kick & Rush	One-Two	Free-kick-Crosses	Corners	Assisted by shot attempt	Header after Cross from open-play	Assisted with head	Assisted with Throw-In	Unintended Teammate Assist	Assisted by the opponent	Assisted by the opponent
	3,457	272	94	80	734	270	70	170	152	28	127	260	76	372	124	20	122	202	278
Hannover 96	124	8	2	2	32	12	5	5	6	2	5	15	0	12	2	1	2	7	6
TSG 1899 Hoffenheim	96	10	2	2	20	8	2	3	5	1	5	7	2	9	4	0	2	7	7
VfL Wolfsburg	95	10	4	3	23	13	1	7	3	1	1	3	4	9	0	0	0	4	5
Borussia Mönchengladbach	93	7	0	2	28	6	1	6	2	1	4	5	1	9	5	0	0	4	7
1. FC Nürnberg	102	8	1	2	24	4	2	4	7	2	3	5	3	10	5	0	3	9	10
1. FSV Mainz 05	108	8	2	1	27	9	1	7	3	2	4	12	4	14	2	0	3	5	4
Borussia Dortmund	90	7	2	4	16	8	2	2	5	2	5	7	1	9	1	0	3	7	9
1. FC Köln	112	9	1	2	24	9	2	7	2	0	8	8	1	12	4	1	5	8	9
FC Schalke 04	89	8	1	0	17	6	2	5	4	0	3	5	2	11	4	1	2	4	14
Sport-Club Freiburg	113	11	1	3	24	7	2	8	6	0	4	11	5	10	4	0	3	9	5
Bayer 04 Leverkusen	92	9	4	2	19	8	4	4	5	2	1	7	4	12	1	2	2	4	2
Hamburger SV	94	4	2	4	19	10	0	6	4	0	5	6	4	5	8	0	3	7	7
VfB Stuttgart	102	6	4	2	18	11	1	7	2	2	5	8	0	13	1	0	7	6	9
SV Werder Bremen	84	6	2	3	16	5	2	7	6	0	2	7	0	9	5	0	3	4	7
Eintracht Frankfurt	88	3	3	1	26	8	3	6	3	2	3	9	3	8	2	0	1	4	3
FC Bayern München	55	6	2	1	12	5	1	1	3	0	0	3	2	7	0	1	3	5	3
FC St. Pauli	101	8	1	1	17	4	2	7	4	1	5	8	2	9	7	2	3	6	14
1. FC Kaiserslautern	50	4	2	0	9	3	0	3	5	0	1	5	1	4	3	0	6	1	3
FC Ingolstadt 04	94	15	5	2	20	4	1	4	3	1	4	9	3	10	2	2	2	5	2
SC Paderborn 07	50	2	0	0	17	3	1	1	3	0	2	4	2	5	1	0	2	4	3
SG Dynamo Dresden	97	6	5	5	21	7	0	3	4	0	3	6	1	14	2	0	3	7	10
Fortuna Düsseldorf	108	11	3	0	18	6	7	5	5	0	3	5	3	9	3	0	5	10	15
MSV Duisburg	117	10	5	3	24	11	2	8	2	2	2	10	1	10	9	2	4	8	4
VfL Bochum 1848	84	10	2	5	22	6	1	2	3	0	3	6	2	6	5	0	3	0	8
1. FC Union Berlin	76	9	2	2	13	6	0	4	1	0	4	6	1	11	0	2	4	5	6
SpVgg Greuther Fürth	98	6	2	1	17	12	3	7	2	0	6	6	1	16	5	0	4	2	8
Eintracht Braunschweig	41	2	3	1	7	3	0	1	5	0	1	3	0	7	1	0	2	2	3
FC Erzgebirge Aue	94	4	2	2	20	9	4	4	7	0	3	6	3	10	5	0	5	5	5
Hertha BSC	100	10	1	3	23	9	2	2	8	0	2	8	4	11	5	0	2	2	8
FC Augsburg	115	3	4	2	20	8	2	5	5	3	4	12	3	15	7	1	2	3	16
SSV Jahn Regensburg	104	7	4	1	13	9	3	1	11	0	5	7	2	16	3	2	6	5	9
SV Sandhausen	82	5	2	6	16	7	1	3	3	1	3	5	4	10	3	1	2	5	5
DSC Arminia Bielefeld	94	6	3	2	25	6	2	6	4	0	1	5	2	6	4	1	5	6	10
SV Darmstadt 98	91	8	2	3	25	6	2	6	3	1	2	5	2	11	3	0	3	3	6
RB Leipzig	80	10	1	1	14	7	1	2	2	0	3	7	2	9	3	0	2	5	11
1. FC Heidenheim 1846	95	6	4	0	19	4	4	4	3	0	2	9	0	13	4	0	3	10	10
1. FC Magdeburg	50	2	5	2	10	0	0	0	2	1	4	2	1	8	1	1	2	5	4
Holstein Kiel	93	8	3	4	19	11	1	7	1	1	6	8	0	3	0	0	2	8	11

C Appendix

Table 5: Conversion rate of goals per shot. Values above (> 20%) and below (< 5%) average are indicated by green and red arrows, respectively.

Description	Efficiency	Cluster	Total	Goals	% Goals	Av. xG (%)	xG (abs)
All goals	11,67%	All	8167	953	11,67%	11,38%	928,18
Penalties	81,32 %	S1	91	74	81,32%	76,00%	69,16
Direct freekicks	7,24%	S2	304	22	7,24%	4,54%	13,80
Intercepted build-up	11,66 %	BE	163	19	11,66%	8,43%	13,75
Inteded assist with leg-finishes (long possession)	16,07%	LP1	803	82	10,21%	9,48%	76,16
		LP2	116	49	42,24%	44,91%	52,10
		LP3	17	8	47,06%	47,03%	7,99
		LP4	21	2	9,52%	8,42%	1,77
		LP5	325	65	20,00%	18,47%	60,04
Intended assists with leg finishes (short possession)	5,22%	SP1	673	34	5,05%	7,03%	47,29
		SP2	817	16	1,96%	2,91%	23,74
		SP3	40	6	15,00%	16,95%	6,78
		SP4	45	1	2,22%	13,90%	6,25
		SP5	76	31	40,79%	38,93%	29,59
		SP6	72	2	2,78%	7,31%	5,27
Headers and volleys	16,52%	HV1	15	4	26,67%	41,45%	6,22
		HV2	58	2	3,45%	7,40%	4,29
		HV3	42	13	30,95%	20,65%	8,67
Solo Assist	14,79%	SA1	281	41	14,59%	6,53%	18,34
		SA2	3	1	33,33%	14,85%	0,45
Kick & rush	15,87%	K&R	315	50	15,87%	11,46%	36,10
One-two	15,00%	OT	80	12	15,00%	10,23%	8,18
Free-kick crosses	6,67%	FC1	257	17	6,61%	9,96%	25,59
		FC2	73	5	6,85%	8,29%	6,05
Corners	8,12%	C1	242	14	5,79%	5,13%	12,42
		C2	18	8	44,44%	45,14%	8,13
		C3	574	45	7,84%	9,27%	53,19
		C4	17	3	17,65%	8,27%	1,41
		C5	36	2	5,56%	3,13%	1,13
Assisted by shot attempt	48,28%	SA	58	28	48,28%	34,81%	20,19
Headers from open play	13,39%	HC	784	105	13,39%	14,60%	114,46
Assisted with head	8,27%	HA1	229	18	7,86%	12,78%	29,28
		HA2	37	4	10,81%	8,44%	3,12
Assisted with throw-in	2,11%	TI	95	2	2,11%	6,33%	6,01
Unintended teammate assist	9,13%	UA1	195	17	8,72%	9,25%	18,04
		UA2	3	1	33,33%	15,77%	0,47
		UA4	43	4	9,30%	7,56%	3,25
Assisted by the opponent (long possession)	11,25%	LO1	16	12	75,00%	51,41%	8,23
		LO2	11	4	36,36%	42,78%	4,71
		LO3	199	18	9,05%	7,12%	14,17
		LO4	258	23	8,91%	7,02%	18,10
		LO5	76	6	7,89%	5,41%	4,11
Assisted by the opponent (short possession)	14,29%	SO1	322	15	4,66%	5,28%	17,01
		SO2	36	1	2,78%	2,36%	0,85
		SO3	170	59	34,71%	31,97%	54,36
		SO4	25	4	16,00%	13,63%	3,41
		SO7	21	3	14,29%	21,77%	4,57

D Appendix

Table 6: Exemplary goals for selected clusters.

Cluster	Season(League)	Pairing	Scoring Team	Goal Scorer	Assist	Minute
S2	2018/2019(1)	FC Augsburg:Hannover 96	Augsburg	Schmid	NaN	0
S2	2018/2019(2)	SpVgg Greuther Fürth:FC Erzgebirge Aue	Aue	Hochscheidt	Krüger	0
S2	2017/2018(2)	Fortuna Düsseldorf:SpVgg Greuther Fürth	Fürth	Wittek	Narey	0
S2	2018/2019(1)	Sport-Club Freiburg:FC Augsburg	Freiburg	Grifo	Grifo	0
BE	2018/2019(1)	Sport-Club Freiburg:Borussia Mönchengladbach	Freiburg	Höler	Sommer	90
BE	2018/2019(1)	Eintracht Frankfurt:Sport-Club Freiburg	Frankfurt	Jovic	Jovic	45
LP1	2017/2018(1)	TSG 1899 Hoffenheim:Borussia Dortmund	Dortmund	Reus	Guerreiro	58
LP1	2017/2018(1)	FC Schalke 04:FC Bayern München	Bayern	Vidal Pardo	Rodríguez Rubio	75
LP2	2017/2018(1)	Borussia Mönchengladbach:Hamburger SV	M'gladbach	Hazard	Caetano de Araújo	9
LP5	2018/2019(1)	Sport-Club Freiburg:Borussia Mönchengladbach	Freiburg	Waldschmidt	Haberer	57
LP5	2017/2018(1)	TSG 1899 Hoffenheim:Bayer 04 Leverkusen	Leverkusen	Alario	Bailey Butler	70
SP1	2017/2018(2)	MSV Duisburg:Fortuna Düsseldorf	Düsseldorf	Hennings	Fink	40
SP1	2018/2019(2)	SpVgg Greuther Fürth:Holstein Kiel	Fürth	Green	Dona Atanga	90
SP5	2017/2018(1)	1. FSV Mainz 05:Sport-Club Freiburg	Mainz	De Blasis	Quaison	79
SP5	2017/2018(1)	TSG 1899 Hoffenheim:Hannover 96	Hoffenheim	Kramaric	Gnabry	16
HV1	2018/2019(2)	FC St. Pauli:SSV Jahn Regensburg	St. Pauli	Flum	Carstens	52
HV1	2017/2018(1)	RB Leipzig:Hannover 96	Leipzig	Werner	Forsberg	85
HV2	2017/2018(2)	MSV Duisburg:SSV Jahn Regensburg	Duisburg	Nauber	Tashchy	52
HV2	2018/2019(1)	Fortuna Düsseldorf 1895 e.V.:FC Augsburg	Augsburg	Hahn	Richter	76
HV2	2018/2019(2)	1. FC Union Berlin:1. FC Heidenheim 1846	Union Berlin	Gikiewicz	Andersson	90
HV2	2017/2018(1)	FC Augsburg:TSG 1899 Hoffenheim	Hoffenheim	Kramaric	Hübner	30
HV2	2018/2019(1)	Borussia Mönchengladbach:SV Werder Bremen	Bremen	Klaassen	Osako	79
HV2	2017/2018(1)	Hertha BSC:FC Bayern München	Bayern	Hummels	Boateng	10
HV2	2017/2018(2)	FC Erzgebirge Aue:1. FC Nürnberg	Aue	Köpke	Tiffert	77
HV2	2017/2018(2)	Fortuna Düsseldorf:1. FC Heidenheim 1846	Heidenheim	Verhoek	Schnatterer	83
HV3	2017/2018(1)	FC Bayern München:1. FSV Mainz 05	Bayern	Lewandowski	Kimmich	77
HV3	2018/2019(1)	Borussia Dortmund:FC Bayern München	Bayern	Lewandowski	Kimmich	52
HV3	2017/2018(1)	RB Leipzig:FC Bayern München	Bayern	Wagner	Rodríguez Rubio	12
SA1	2018/2019(1)	FC Bayern München:Eintracht Frankfurt	Bayern	Ribéry	Kimmich	72
SA1	2017/2018(1)	TSG 1899 Hoffenheim:1. FC Köln	Hoffenheim	Gnabry	Grillitsch	47
SA1	2017/2018(1)	TSG 1899 Hoffenheim:RB Leipzig	Hoffenheim	Gnabry	Amiri	62
K&R	2017/2018(2)	1. FC Nürnberg:FC St. Pauli	St. Pauli	Sobota	Himmelmann	63
K&R	2017/2018(1)	Sport-Club Freiburg:1. FSV Mainz 05	Mainz	Berggreen	Brosinski	90
OT	2018/2019(1)	Eintracht Frankfurt:FC Bayern München	Bayern	Ribéry	Kimmich	79
OT	2018/2019(1)	1. FC Nürnberg:Hertha BSC	Berlin	Ibisevic	Selke	15
FC1	2017/2018(1)	Borussia Dortmund:Eintracht Frankfurt	Frankfurt	Jovic	de Guzmán	75
FC1	2017/2018(1)	Eintracht Frankfurt:1. FC Köln	Köln	Terodde	Risse	74
FC2	2017/2018(1)	FC Augsburg:Eintracht Frankfurt	Augsburg	Koo	Baier	19
FC2	2017/2018(1)	TSG 1899 Hoffenheim:1. FSV Mainz 05	Hoffenheim	Kramaric	Uth	67
C1	2017/2018(2)	SSV Jahn Regensburg:1. FC Heidenheim 1846	Regensburg	George	Lais	34
C1	2018/2019(1)	FC Augsburg:Eintracht Frankfurt	Augsburg	Córdova Lezama	da Silva	90
C3	2017/2018(1)	Hamburger SV:Eintracht Frankfurt	Hamburg	Papadopoulos	Hunt	9
C3	2018/2019(1)	Hertha BSC:Eintracht Frankfurt	Berlin	Grujic	Plattenhardt	40
C4	2017/2018(1)	Bayer 04 Leverkusen:VfL Wolfsburg	Leverkusen	Bender	Retsos	29
C4	2018/2019(1)	FC Bayern München:Borussia Mönchengladbach	M'gladbach	Herrmann	Kramer	88
C5	2017/2018(2)	DSC Arminia Bielefeld:VfL Bochum 1848	Bielefeld	Kerschbaumer	Staude	35
C5	2018/2019(2)	DSC Arminia Bielefeld:1. FC Heidenheim 1846	Bielefeld	Schütz	Hartherz	33
C5	2018/2019(1)	Bayer 04 Leverkusen:TSG 1899 Hoffenheim	Hoffenheim	Nelson	Grifo	19
SA	2018/2019(1)	Bayer 04 Leverkusen:Eintracht Frankfurt	Leverkusen	Brandt	Aránguiz Sandoval	13
HC	2018/2019(2)	1. FC Heidenheim 1846:SV Sandhausen	Sandhausen	Wooten	Diekmeier	69
HC	2018/2019(1)	Fortuna Düsseldorf 1895 e.V.:Eintracht Frankfurt	Frankfurt	Mendes Paciencia	de Guzmán	48
TI	2017/2018(1)	Hamburger SV:FC Schalke 04	Hamburg	Kostic	dos Santos Justino De Melo	17
TI	2017/2018(1)	Bayer 04 Leverkusen:1. FC Köln	Köln	Guirassy	Sørensen	23
TI	2018/2019(2)	SG Dynamo Dresden:MSV Duisburg	Dresden	Röser	Heise	39
UA1	2017/2018(2)	SSV Jahn Regensburg:FC Erzgebirge Aue	Aue	Köpke	Riese	57
UA1	2017/2018(1)	Eintracht Frankfurt:SV Werder Bremen	Frankfurt	Rebic	Willems	17
UA1	2017/2018(2)	MSV Duisburg:Fortuna Düsseldorf	Duisburg	Tashchy	Stoppelkamp	90
UA3	2018/2019(2)	SSV Jahn Regensburg:SG Dynamo Dresden	Dresden	Dumic	Koné	52
UA3	2017/2018(1)	FC Bayern München:FC Augsburg	Bayern	Vidal Pardo	Süle	31
LO1	2017/2018(1)	VfB Stuttgart:Eintracht Frankfurt	Stuttgart	Thommy	Ginczek	13
LO1	2018/2019(1)	1. FSV Mainz 05:Borussia Dortmund	Mainz	Quaison	Hack	70
LO2	2017/2018(1)	Borussia Dortmund:FC Augsburg	Dortmund	Reus	Schürrle	16
LO2	2018/2019(2)	FC Ingolstadt 04:Holstein Kiel	Ingolstadt	Lezcano Farina	Kutschke	13
LO3	2017/2018(1)	FC Bayern München:Eintracht Frankfurt	Frankfurt	Haller	Vieira da Costa	78
LO3	2017/2018(1)	FC Bayern München:Hannover 96	Bayern	Coman	Müller	67
LO4	2017/2018(2)	1. FC Kaiserslautern:SV Sandhausen	Sandhausen	Förster	Linsmayer	78
LO4	2017/2018(1)	Eintracht Frankfurt:FC Schalke 04	Schalke	Aparecido Rodrigues	Embolo	90
LO5	2017/2018(2)	FC St. Pauli:FC Ingolstadt 04	Ingolstadt	Träsch	Pledl	33
LO5	2018/2019(2)	Holstein Kiel:FC Erzgebirge Aue	Aue	Hochscheidt	Iyoha	26
SO1	2017/2018(2)	MSV Duisburg:VfL Bochum 1848	Duisburg	Tashchy	Bomheuer	7
SO1	2017/2018(1)	VfL Wolfsburg:Borussia Mönchengladbach	Wolfsburg	Akoi Fara Guilavogui	Gómez García	71
SO1	2017/2018(2)	1. FC Heidenheim 1846:FC St. Pauli	Heidenheim	Thiel	Schnatterer	16
SO1	2018/2019(2)	1. FC Union Berlin:MSV Duisburg	Duisburg	Oliveira Souza	Iljutcenko	77
SO2	2017/2018(2)	SV Darmstadt 98:SG Dynamo Dresden	Dresden	Konrad	Berko	80
SO2	2017/2018(2)	MSV Duisburg:DSC Arminia Bielefeld	Duisburg	Wolze	Stoppelkamp	72
SO2	2018/2019(2)	MSV Duisburg:SC Paderborn 07	Duisburg	Tashchy	Wolze	63
SO2	2017/2018(2)	Holstein Kiel:SV Sandhausen	Sandhausen	Klingmann	Höler	35
SO2	2018/2019(1)	Hertha BSC:TSG 1899 Hoffenheim	Berlin	Lazaro	Plattenhardt	87
SO2	2017/2018(1)	Hertha BSC:Borussia Mönchengladbach	M'gladbach	Caetano de Araújo	Wendt	20
SO2	2017/2018(2)	SG Dynamo Dresden:SV Sandhausen	Sandhausen	Paqarada	Daghfous	25
SO2	2018/2019(2)	SC Paderborn 07:1. FC Köln	Paderborn	Pröger	Michel	86
SO2	2017/2018(2)	FC Erzgebirge Aue:MSV Duisburg	Aue	Nazarov	Tiffert	83
SO2	2017/2018(1)	Hannover 96:FC Augsburg	Hannover	Sané	Klaus	37
SO3	2018/2019(1)	FC Augsburg:Bayer 04 Leverkusen	Leverkusen	Tah	Brandt	60
SO3	2017/2018(1)	FC Bayern München:Sport-Club Freiburg	Bayern	Coman	Robben	42
SO5	2017/2018(2)	Fortuna Düsseldorf:1. FC Heidenheim 1846	Düsseldorf	Raman	Hennings	90
SO5	2018/2019(1)	FC Augsburg:Hannover 96	Hannover	Weydandt	Maina	8
SO5	2018/2019(1)	Sport-Club Freiburg:1. FSV Mainz 05	Mainz	Onisiwo	Niakhaté	75
SO5	2017/2018(2)	1. FC Nürnberg:1. FC Heidenheim 1846	1. FC Nürnberg	Stefaniak	Ishak	38

D Appendix—Study IV: Putting Team Formations in Association Football into Context

Noname manuscript No.
(will be inserted by the editor)

Putting Team Formations in Association Football into Context

Gabriel Anzer[1,2], **Pascal Bauer** [2,3], **Laurie Shaw**[4]

[1]**Sportec Solutions AG, subsidiary of the Deutsche Fußball Liga (DFL), Munich, Germany**
[2]**Institute of Sports Science, University of Tübingen, Tübingen, Germany**
[3]**DFB-Akademie, Deutscher Fußball-Bund e.V. (DFB), Frankfurt, Germany**
[4] **Department of Statistics, Harvard University, Boston, USA.**

Received: date / Accepted: date

Abstract Choosing the right formation is one of the coach's most important decisions in football. Teams change formation dynamically throughout matches to achieve their immediate objective: to retain possession, progress the ball up-field and create (or prevent) goal-scoring opportunities. In this work we identify the unique formations used by teams in distinct phases of play in a large sample of tracking data. This we achieve in two steps: first, we trained a convolutional neural network to decompose each game into non-overlapping segments and classify these segments into phases with an average F_1-score of 0.76. We then measure and contextualize unique formations used in each distinct phase of play. While conventional discussion tends to reduce team formations over an entire match to a single three-digit code (e.g. 4-4-2; 4 defender, 4 midfielder, 2 striker), we provide an objective representation of teams formations per phase of play. Using the most frequently occurring phases of play, mid-block, we identify and contextualise six unique formations. A long-term analysis in the German Bundesliga allows us to quantify the efficiency of each formation, and also to present a helpful scouting tool to identify how well a coach's preferred playing style is suited to a potential club.

Keywords Football, sports analytics, human-in-the-loop machine learning.

1 Introduction

The great Dutch football player Johan Cruyff famously observed that, on average, each player is in possession of the ball for only 3 of the 90 minutes during a football match.[1] He expanded on this observation by stating *"... so, the most important thing is: what do you do during those 87 minutes when you do not have the ball? That is what determines whether you are a good player or not."*[2] The implication is that a player can significantly influence the game through their positioning and movement on the field, even when they do not directly interact with the ball (Brefeld et al. 2019; Fernandez et al. 2018).

The movement of players in a football match represent a high-dimensional spatio-temporal configuration. Various approaches aimed to embed teams' behavior in higher-level problems. Balague et al. (2013) focuses on coordination of motion within a team by modelling a team's movement as collective behavior in a complex system. Indeed, synchronicity of movements is investigated in football in specific situations (Goes et al. 2020b; Sarmento et al. 2018). Several studies described football matches more concrete as a multi-agent systems (Beetz et al. 2006; Fujii 2021) highlighting the intelligence of interactions between the agents (players). Analysing movement patterns in spatio-temporal data, especially the detection of repeating, collective patterns is not only researched in invasion sports (Gudmundsson et al. 2017a), but also in traffic management, surveillance and security or in the military and battlefield domain (Gudmundsson et al. 2017b). Key challenges in spatio-temporal pattern detection are: (a) Using the interaction of movement for dimensional reduction (Balague et al.

P. Bauer
E-mail: pascal.bauer@dfb.de

Gabriel Anzer
E-mail: gabriel.anzer@herthabsc.de

[1] Link et al. (2017) showed that it is even less with large differences between playing positions: central forwards (0:49 ± 0:43 min), central defenders (1:38 ± 1:09 min), central midfielders (1:27 ± 1:08 min) and, surprisingly, the longest for goalkeepers (1:38 ± 0:58 min).

[2] https://wheecorea.com/johan-cruyff-football-my-philosophy/25-johan-cruyff-quotes/, accessed 02/07/2021

2013), (b) finding appropriate similarity metrics for related, but never identical trajectories of multiple entities (Vilar et al. 2013), and (c) project multi-agents in a permutation-invariant space (Yeh et al. 2019).

The literature differentiates between tactics (decisions made during a match as a consequence of the dynamic interaction in a match) and strategy (decisions made before the match) (Gréhaigne et al. 1999). However, these concepts are often hard to distinguish (Rein et al. 2016). Coming from a more general understanding of team formations (Wang et al. 2015), Budak et al. (2019) highlighted the problem of optimizing the team composition (i.e. which players should be on the pitch) before the season, before the match and during the match stage as a relevant problem in team sports. According to this definition, several approaches presented evidence-based strategies to optimize this composition of players (Boon et al. 2003). However, this neglects the players actual interaction on the pitch (i.e. tactics), what is in the focus of our investigation and will further be declared as the *(playing) formation*.

One potential reason for this high-level consideration is the lack of available data quantifying what happens on the pitch. For the longest time one could not objectively measure a team's playing formation, since the only available data describing football matches was so-called *event data*. Dating as far back as 1968 when Charles Reep started manually collecting events such as shots or passes (Reep et al. 1968), this event data, which is still being manually collected today, describes all ball actions and the players involved (Pappalardo et al. 2019a; Stein et al. 2017). Although event data allowed for ground-breaking discoveries in football tactics (Xu 2021; Pantzalis et al. 2020; Decroos et al. 2019; Danisik et al. 2018; Decroos et al. 2018; Pappalardo et al. 2019b; Cintia et al. 2015; Haaren et al. 2013), it does not include any information about the positioning of all other players. Now, with recent developments in computer vision technologies (Thinh et al. 2019; Baysal et al. 2016; Teoldo et al. 2009) it has become possible to capture exactly that: optical tracking systems are able to record centimeter-accurate positions of all players at every moment of a match (hereafter referred to as *positional* or *tracking data*). This development unlocked huge potentials for professional football (Anzer et al. 2022; Anzer et al. 2021a; Araújo et al. 2021; Wang et al. 2020; Goes et al. 2020a; Andrienko et al. 2019; Rein et al. 2016; Herold et al. 2019).

The first approaches in football analysed formations assuming that teams play with a fixed formation across the whole match, describing them simply as playing with a 4-4-2 (4 defenders, 4 midfielders and 2 forwards), 5-3-2, 4-3-3, or one of approximately ten other formations that are commonly referenced (Wilson 2009). Differences in physical requirements for similar player-roles in different formation (e.g. a central defender in a 4-4-2 versus a 5-4-1) were analysed (Vilamitjana et al. 2021; Tierney et al. 2016; Carling 2011; Bradley et al. 2011). However, breaking a team's formation down to three digits in a complex sport like football is a gross over-simplification (Müller-Budack et al. 2019).

Driven by the increasing availability of tracking data, analysing team formations has been a research issue in several sports (Gudmundsson et al. 2017b). Initiated by a pioneering work in 1999 (Intille et al. 1999), unique formations were derived at the moment a play starts using positional data in American football (Atmosukarto et al. 2013). Hochstedler et al. (2017) build on the static formation detection in American football by classifying the routes of chosen player during the plays. In basketball, event data has been used to investigate established player roles (Bianchi et al. 2017). Lucey et al. (2013) published a quantitative analyses of team formations in field-hockey using tracking data, which was transferred to football (Wei et al. 2013) and incrementally extended Bialkowski et al. (2014a), Bialkowski et al. (2015), and Bialkowski et al. (2016). They describe formations as a "*a coarse spatial structure which the players maintain over the course of the match*" and which assigns each player at every time of the match a unique role. Bialkowski et al. (2015) further define a role as a players position relative to the other roles. They describe a role-identification methodology for measuring formations, iteratively refining estimates of the average spatial positions (and deviations from those positions) of ten unique outfield roles throughout a match. Applying a clustering algorithm on tracking data for a season of a 20-team professional league, Bialkowski et al. (2014a) identified six unique formation types: 4-4-2, 3-4-3, 4-4-1-1 and 4-1-4-1 are all visible in their results. Variations in formations between game-states (i.e. offensive, defensive) were first explored in Bialkowski et al. (2016). Using a more supervised approach, Müller-Budack et al. (2019). annotated twelve typical formations (split between offense and defense) and addressed the formation problem as a classification task. Narizuka et al. (2019) derived unique formations of 45 Japanese J1 league using a Delaunay method combined with hierarchical clustering.

Ric et al. (2021) and Shaw et al. (2019) presented a data-driven technique for measuring and classifying team formations as a function of game-state (offensive, defensive, transition), analysing the offensive and defensive configurations of each team separately and dynamically detecting major tactical changes during the course of a match. Defensive and offensive formations were measured separately by aggregating together consecutive periods of possession of the ball for each team into two-minute windows of in-play data. Splitting up formations into different game-states, i.e. excluding fuzzy transition situations, presented a major improvement of formation analysis, however, they stated that further sub-game-states should be considered in future work to achieve even more granularity (Ric et al. 2021).

While these pioneering studies have provided methods for measuring team formations and demonstrated observations of the coherent structures formed by teams as they move around the field (and validated by football

experts), they do not fully account for the changing objectives of a football team as a match evolves, influencing team formations drastically (Andrienko et al. 2019; Gudmundsson et al. 2017b; Shaw et al. 2019; Lucey et al. 2013; Bialkowski et al. 2016). Several studies pointed out, that football consist of repetitive movement patterns, that can be recognized by experts (Sampaio et al. 2012). We define a *tactical pattern* as a recurring, collective behaviour conducted by a team or a sub-group of a team in a specific situation of a match, that can be clearly identified by experts (Rein et al. 2016; Kempe et al. 2015; Wang et al. 2015; Grunz et al. 2012). Whereas the detection of tactical patterns has been a relevant issue in basketball (Kempe et al. 2015; Chen et al. 2014; Perse et al. 2006), handball (Pfeiffer et al. 2015), American football (Hochstedler et al. 2017; Stracuzzi et al. 2011; Li et al. 2010; Siddiquie et al. 2009), and Australian rules football (Alexander et al. 2019), often only patterns conducted by subgroups of players are analysed. The complexity of a football match requires so called *team tactics* in which the whole team is involved (Rein et al. 2016). Some exemplary patterns like counterattacks (Fassmeyer et al. 2021; Hobbs et al. 2018), ball regain strategies (Vogelbein et al. 2014), i.e. counterpressing (Bauer et al. 2021) or general offensive strategies (Decroos et al. 2018; Kempe et al. 2014; Grunz et al. 2012; Borrie et al. 2002; Montoliu et al. 2015; Fernando et al. 2015) have been addressed in literature and classified as sub-categories of game-states (e.g. counterattacks and counterpressing as a subgroup of transitions in Bauer et al. (2021) and Hobbs et al. (2018)). For such well established tactical patterns, which unavoidably occur in every match, practitioners often use the term *(tactical) phases of play*[3] (although no scientific definition established) or *(tactical) game-phases* (Lucey et al. 2014).

The consequence of this is that the results are not observations of a single distinct formation of a team, but a mixture (or 'superposition') of the different formations used in different phases of play (Shaw et al. 2019; Müller-Budack et al. 2019). This paper resolves this problem by using a convolution neural network (CNN) to classify a football match over time into distinct phases of play, before measuring the formations used by either team in each distinct phase. There are therefore two parts to our approach:

(1) A phases of play detection CNN, with architecture specifically designed for the purpose, was trained using labeled tracking data from 97 matches in the German Bundesliga based on phases of play classifications provided by professional analysts. Our classification scheme is described in Section 3.
(2) Within each match, periods of play classified to the same phases of play (from the perspective of one team) are then aggregated to obtain precise measurements of the formations used. This is described in Section 4.

We apply the phases of play classifier and formation measurement tools to tracking data obtained for 2,142 matches in the German Bundesliga over seven seasons, identifying the unique formations used in each phase of play across our sample. This combination of a phase of play detection and formation detection fully automates the process of identifying the distinct formation configurations used by teams during a game, revealing the specific instructions that managers gave their team. This research was conducted in close collaboration with professional match analysts from German Bundesliga clubs and the German national teams, who have provided human validation of our methodology and results. This project therefore combines machine learning and human experience aiming to obtain results that are insightful, meaningful and of practical use to coaches, managers and scouts.

As a side-product of a practical relevant process automatization for match analysis departments, we outline two clear use-cases of our work in Sec. 5. We are the first to quantify the strengths and weaknesses of a specific formation when pitted against another, providing the foundation for evidence-based advice for managers seeking the most effective counter to an opponent's strategy during specific phases of the game (Sec. 5.1). Second, we assess the tactical preferences of individual managers, highlighting how our tools can be used to find managers that would provide continuity to a team's existing playing style (Sec. 5.2). Style-matching is a crucial element of managerial recruitment, helping to prevent a large turnover of players as a manager seeks to impose a new playing style on a new team.

2 Positional Data

The German Bundesliga collects consistent positional data on a league-wide level, making this data available to every team. Positional data, often also referred to as tracking or movement data (Stein et al. 2017), contains measurements of the positions of all players, referees and the ball, sampled at a frequency of 25 Hz. These data are gathered by an optical tracking system that captures high resolution video footage from different camera perspectives.

In this paper, we make use of positional data from seven seasons of the German Bundesliga, from 2013/2014 until 2019/2020: a total of 2,142 matches and nearly half a billion frames are acquired by Chryronhego's TRACAB system.[4] Validating the quality of such tracking data presents somehow an ill-posed problem due to

[3] An exemplary explanation of the definition can be found here: https://www.statsperform.com/resource/phases-of-play-an-introduction/.

[4] https://chyronhego.com/wp-content/uploads/2019/01/TRACAB-PI-sheet.pdf (accessed 02/05/2021).

missing ground truth positions. Even though, several studies evaluated the accuracy of the underlying data used in this study (Redwood-Brown et al. 2012; Linke et al. 2018; Linke et al. 2020; Taberner et al. 2020), and found an average diversion of less then 10 cm for player positioning compared to an accurate measurement system. Pettersen et al. (2014) presents a publicly available set of positional data, which can be used for reproduction.[5]

3 Phases of Play Classification

3.1 Defining Phases of Play

The primary goal in football is to score more goals than the respective opponent. Consequently, the two major objectives are scoring goals and preventing the opponent from doing so (Kempe et al. 2014). However, given specific situations those goals are often only implicitly followed, while sub-tasks (e.g. (re)gaining possession of the ball), are predominant in certain situations. The concept of phases of play derives from the idea that any moment of a match can be categorized based on the immediate intentions of each team, e.g. in defense, teams always have to balance between the two most relevant objectives of regaining the ball (preferably in a good position to perform an attack) and purely prevent the opponent from scoring. At the simplest level, a match can be divided into the phases of *offense* and *defense* for each team (António et al. 2014), i.e., periods in and out of possession of the ball. At a more granular level, professional analysts involved in our project classified the progressive stages of attacking and defense into distinct phases.[6] Fig. 1 provides an example of the phases of play classification scheme developed by German Bundesliga analysts (see Acknowledgements). In this scheme, open-play during a match revolves between periods of offense, transition to defense, defense and transition to offense, with set-pieces providing a separate category (which could also be broken further down into offensive and defensive set-pieces as well as different categories like corner kicks, throw-ins, freekicks, etc.).

Offensive play is divided into two phases: *build-up*, where the objective is to breach the opponent's first defensive line, and *attacking-play*, where the first line of defenders has been outplayed and the main objective is to create a goal-scoring opportunity. In defense, professional analysts differentiate between aggressive attempts to reclaim possession near the opponent's goal (*high-block*), a default defensive stance as the opponent progresses the ball up the field (*midfield-block* or *mid-block*) and a very compact defensive stance near to a team's own goal, where the sole objective is to prevent the opponent from scoring (*low-block*). These defensive phases were also explored in Anzer et al. (2021b) and Power et al. (2017).

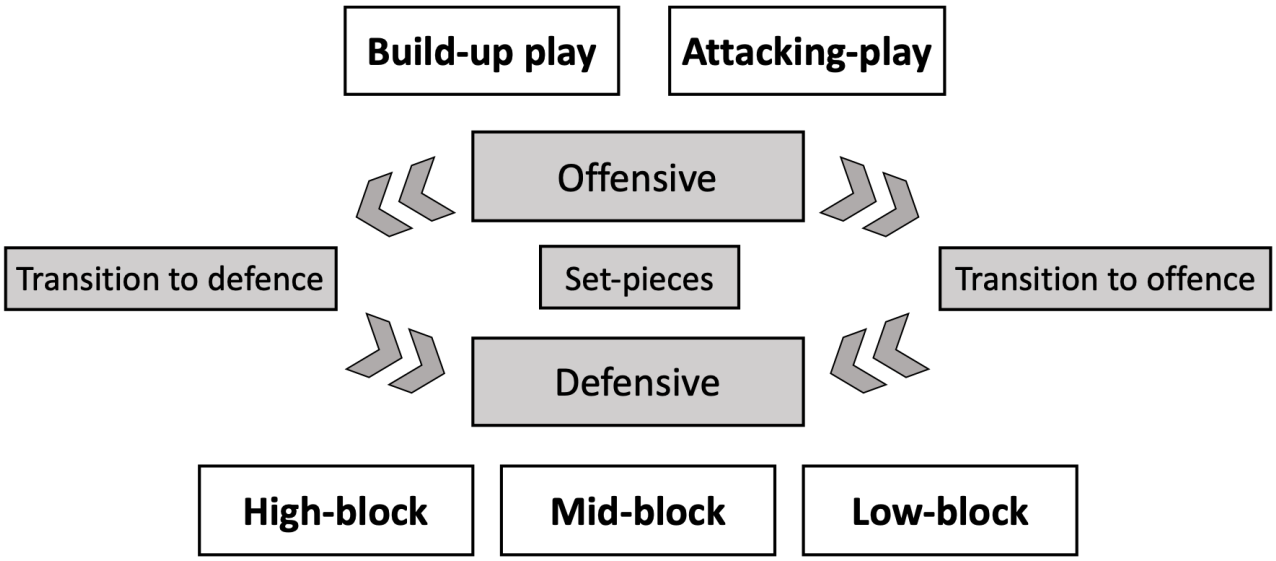

Fig. 1: Overview of tactical phases of play considered.

Fig. 2 shows the phases of play break-down of a two-minute sequence of play during the Nations League match between the German men's national team and Spain in September 2020. The central plot shows the distance between the German team centroid (the average position of the outfield players) and their own goal from the 36th to 38th minutes of the game. The highlighted regions indicate the phases of play classifications,

[5] Other (non-scientific) open-source positional data sets can be accessed from Skillcorner (https://github.com/SkillCorner/opendata) or Metrica sports (https://github.com/metrica-sports/sample-data).

[6] See also: https://www.statsperform.com/resource/phases-of-play-an-introduction/.

from the perspective of the German team, as determined by professional German match analysts. Freeze frames from the footage are shown at four different instants.

The passage of play starts with a Spanish goalkick. Germany confronted this situation by attempting to force a turnover near to the Spanish goal with a high-block. Over the first 30 seconds of play, the Spanish team played through the high-block, forcing Germany to retreat, first into a mid-block and then to a low-block to defend their own goal. Germany regained possession after a shot saved by Manuel Neuer (Germany's goalkeeper) and immediately initiated a build-up phase of possession. A long pass towards Leroy Sané on the right side of the field briefly brought Germany into the attacking-play phase. However, Spain rapidly won the ball back, after which Germany transitioned into a defensive mid-block and then a low-block as Spain advanced again.

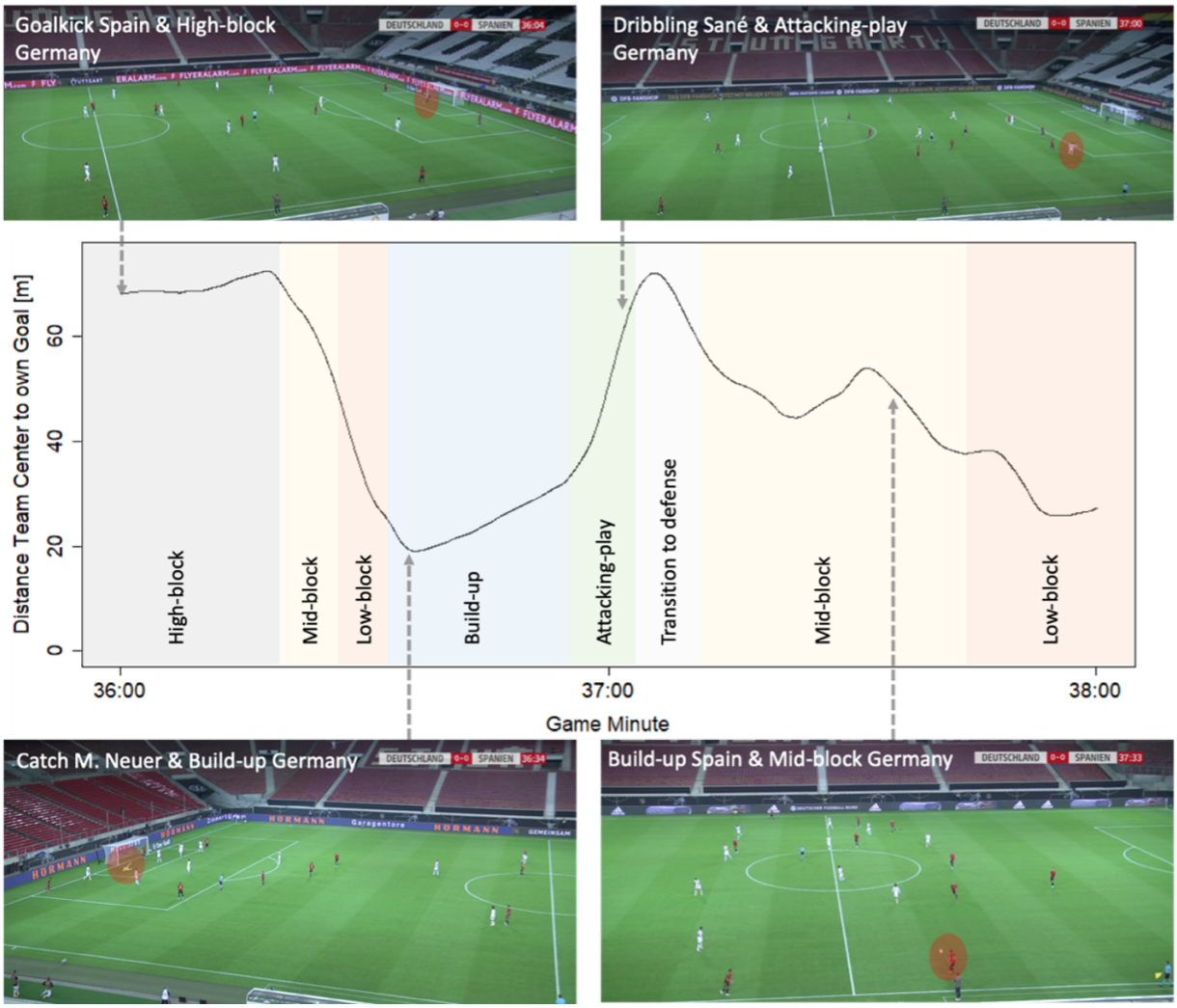

Fig. 2: Team behaviour per phase of play by the reference of Germany against Spain (3rd of September 2020, venue: Stuttgart, result: 1:1). The highlighted areas (red) in the video-footage mark the current ball action.

Match analysts spend a substantial proportion of their time manually breaking down and classifying matches into tactical phases by watching video footage. There are very few methods published in the literature that attempt to automate this process. Those that do focus on finding a single specific transition phases, such as counterattacking (Fassmeyer et al. 2021; Decroos et al. 2018; Hobbs et al. 2018) or counterpressing (Bauer et al. 2021), but none attempt to classifying entire games. We now describe our methodology for achieving this.

3.2 Automated detection of Phases of Play

The phases of play definitions shown in Fig. 1 were established in collaboration with professional match analysts from Bundesliga teams (see Acknowledgements). These definitions were then adopted by professional match analysts to annotate 97 Bundesliga matches from the 2018/2019 season. Using the expert-labelled matches as a training set, we explored two different machine learning approaches for automated classification of phases of play using optical tracking data.

Table 1: Rules for baseline model formation detection.

Phase of Play	Rule
Offensive	The first 6 seconds after a team gains ball possession are classified as transition to offense. The remaining time during a ball possession are classified as the offensive phase.
Build-up	Any moment during the offensive phase, when the ball is within its own third or the mid third of the pitch is classified as build-up.
Attacking-play	Any moment during the offensive phase, when the ball is within the opponents third is classified as attacking-play.
Defensive	The first 6 seconds after a team looses ball possession are classified as transition to defense. The remaining time during a ball possession are classified as the defensive phase.
Low-Block	Any moment during the defensive phase, when the defending team's center (of the outfield players) is at most 20 meters from its own goal-line, is classified as low-block.
Mid-block	Any moment during the defensive phase, when the defending team's center (of the outfield players) is between 20 meters and 60 meters from its own goal-line, is classified as mid-block.
High-block	Any moment during the defensive phase, when the defending team's center (of the outfield players) is at further than 60 meters from its own goal-line, is classified as high-block.

Table 2: Outcome of the phases of play detection CNN.

Tactical Phase of Play	Low-block	Mid-block	High-block	Build-up	Attacking-play
Labeled phases	1 h 57 min	23 h 30 min	1 h 53 min	27 h 37 min	4 h 53 min
Average duration	9.1 s	19.0 s	13.3 s	18.6 s	8.1 s
F_1-score	0.37	0.80	0.29	0.83	0.54
Baseline model F_1-score	0.18	0.75	0.26	0.76	0.39
Inter-labeller reliability (avg. F_1-score)	0.38	0.78	0.24	0.79	0.45

The first approach is a rule-based baseline model, as described in Table 1; the results of the prediction of the rule-based approach (compared to the inter-labeller accordance) are shown in Table 2.

The second approach makes use of convolutional neural networks (CNN), which enables us to model spatio-temporal football data in a high dimensional, permutation-invariant space (see also Dick et al. (2019), Zheng et al. (2016), and Wang et al. (2016)), using the raw positional data as input instead of requiring a costly step of feature engineering (as conducted in Bauer et al. (2021) to detect counterpressing as another example of a tactical pattern). For the CNN's the positional data is mapped to 2-D images. Further details regarding the network architecture can be found in the Appendix A.

On a frame-by-frame level, the CNN predicts the phases of play in our test set with a weighted average F_1 score of 0.76, which is basically limited by the pairwise inter-labeller reliability of 85% (weighted F_1-score 0.72) and exceeds the accuracy of the baseline model (0.69). On further examination, we found that the mis-classified frames mainly occurred near the start and end points of each phase of play.

Table 2 shows some basic statistics for the training data, including the F_1-score—the harmonic mean of recall and precision (see also Goutte et al. (2005))—for each phase of play. By taking both false positives and false negatives into consideration, the F_1-score (calculated for each class individually) presents a very stable evaluation metric for our purpose. Mid-block and build-up are clearly the dominant phases, making up 39% and 47% of the phases shown in Table 2. They are also the phases with the longest duration, lasting an average of 19.0 seconds (mid-block) and 18.6 seconds (build-up). As the mid-block is the standard opponent response to the build-up phase, it is not surprising that the average durations are similar in length. These phases also have the highest classification accuracy for our CNN, with both having F_1-scores exceeding 0.8. The next most regular phase is attacking-play, making up 7% of the training data. Low-block (3%) and high-block (3%) are the least frequently occurring phases.

The trained model was applied on seven full seasons of German Bundesliga (2013/2014-2019/2020). Much of the following analysis focuses on the two most frequent phases: build-up and mid-block.

4 Formation Detection

4.1 Phase-dependent formations

Although positional data has been used in recent literature to quantify team-formations (Shaw et al. 2019; Müller-Budack et al. 2019; Bialkowski et al. 2016; Bialkowski et al. 2014b; Bialkowski et al. 2015; Wei et al. 2013), they aggregate player positions over the entire match ignoring tactical changes during the match. In the following we motivate the relevance of a more granular contemplation.

Fig. 3 shows the different formations employed across each of the five phases of play for one team during a Bundesliga match.The dots indicate the average position of each player in the formation; the ellipses provide an estimate of how far players tend to move from their average positions (the team is playing from left to right), visualized through their 80% confidence region. The lower three images show the formations in the three

defensive phases: low-block (left), mid-block (center) and high-block (right); the top images show the formation in the two offensive phases: build-up (left) and offense (right).

The figure clearly indicates that team formations do not only depend on which team is in possession of the ball, it is also heavily influenced by the tactical patterns teams are applying in different situations on the pitch, e.g. whether the team is currently building up in their own half or attacking in the last third of the pitch. Also, in defensive phases of play, Fig 3 (lower row) shows significant differences depending on the teams defending strategy (high-/mid-/low-block).

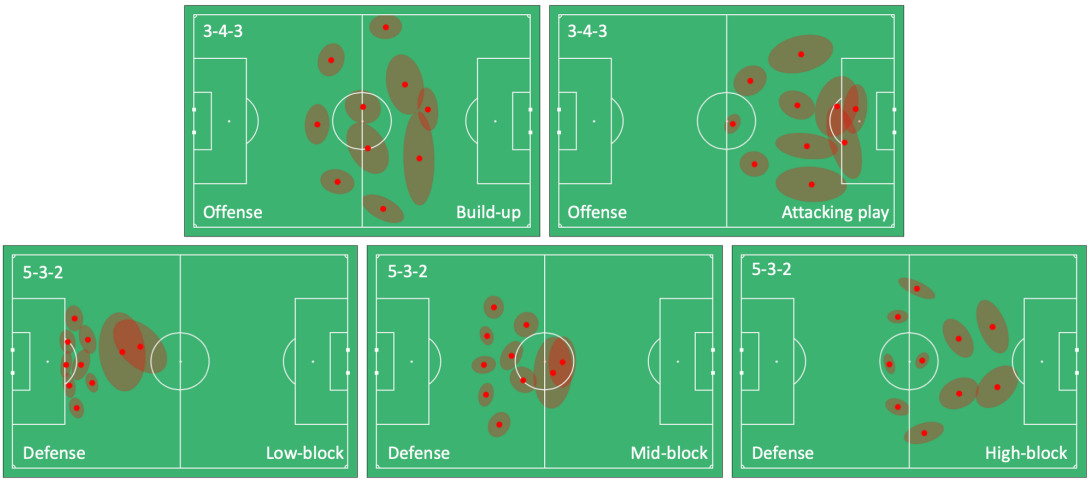

Fig. 3: Average player positions of a team per tactical phase of play during one match. The ellipses provide an estimate of how far the player would tend to move from their average position during each phase of play. The considered team plays from left to right. Player's positions are collected by optical tracking systems at 25 Hz (positional data).

4.2 Measuring Formation in Distinct Phases of Play

A major objective of this work is to identify the distinct formations used by teams during different phases of play during their matches. We focus specifically on the three defensive phases (high-block, mid-block and low-block) and two offensive phases (build-up and attacking-play) shown in Fig. 1. Transitions and set-pieces are ignored: by definition, teams do not have a clear spatial structure during transitions, while positioning during set-pieces are extremely dependent on the position of the ball (Casal et al. 2015). Furthermore, as it takes some time for a team to change from one formation to another—for example, they cannot instantly shift from a high-block to a mid-block—we ignore the first three seconds of any continuous sequence of play that was classified to a single phase of play; if the duration of the entire sequence is less than three seconds, we discard it from our sample. In our case, the range of observations encompasses all frames classified to the same phase of play. At least 60 seconds of (aggregated) data are required to obtain a precise measure of a formation: if the total amount of time spent by a team in any given phase does not meet this criterium, we do not measure a formation for that phase.

Our method for measuring formations proceeds as follows. For each team, we aggregate together all the tracking data frames classified to a particular phase during the match and use them to measure the formation of the team in that phase. This is achieved using the methodology of Shaw et al. (2019), who introduced a geometric approach to measuring formations, calculating the vectors between each pair of teammates at a given instant during a match and averaging these over a range of observations (frames) to gain a clear measure of the team formation: each player's position is calculated relative to the position of his nearest teammate. This process starts with the player in the centre of the team (specifically, the player with the lowest average distance to their third-nearest neighbour), stepping from player to player until the entire team formation is mapped out. This method is founded on the intuition that players orient themselves relative to their nearest teammates to retain the relative positioning required by the team's formation.

A coach may, of course, make a major tactical change during a match, changing their team's formations across all phases of play. To avoid mixing two different formation strategies within a match, we search for major tactical changes in formation by looking at each player's average position relative to their teammates over a rolling time window. If the relative positions change for more then ten meters (based on a three minute rolling average), we start a new set of formation observations; more details are given in appendix B. At least one major change in formation of either team is found in 43% of matches—taking this factor into consideration presents

Table 3: Included formation observations from seven years of the German Bundesliga (2013/2014 until 2018/2019)

Tactical Phase	Low-block	Mid-block	High-block	Build-up	Attacking-play
Formation Observations	1,212	5,200	638	4,867	3,164

a major improvement compared to prior work. In these games there are therefore two (or more) formation measures for each phase of play.

From the 2,142 matches, we exclude 345 matches that did not end with 22 players on the pitch (e.g. due to injuries or expulsions) resulting in a final sample of 1,803 matches. The final number of formation observations in each phase of play are given in Table 3. As discussed above, there was not always sufficient data to measure a formation in all phases of play during a match for both teams. Therefore, there are fewer observations in the least frequent phases, the low-block and high-block (furthermore, not all teams employ a high-block for tactical reasons). There are observations of the mid-block, build-up and attacking-play for almost all teams in every match in our sample (and, on occasion, more if a team made a major tactical change during the match).

4.3 Formation Classification

To study how a specific team plays over multiple matches, we must reduce the size of our formation dataset by identifying the unique formations within each phase of play over our entire sample of matches and classifying individual observations into these unique formations. The pioneering football coach, Marcelo Bielsa, has previously claimed that there are not more than ten formations[7] in common use in professional football—our methods enable us to explore this claim directly. Classifying formations allows us to quantify the strengths and weaknesses of a given formation when pitted against another (Section 5.1), and study the preferred formations used by individual Bundesliga coaches (Section 5.2).

To identify unique formation types, we apply agglomerative hierarchical clustering to the formation observations within each phase of play, using the Wasserstein metric to quantify formation similarity and the Ward metric (Ward et al. 1963) as the linkage criterion, as described in Shaw et al. (2019). The square of the Wasserstein distance is calculated according Olkin et al. (1982):

$$W(\mu_1,\mu_2)^2 = \|m_1 - m_2\|^2 + \text{trace}\left(C_1 + C_2 - 2\left(\sqrt{C_2}C_1\sqrt{C_2}\right)^{1/2}\right),$$

whereby $\mu_i = N(m_i, C_i)$ are bivariate normal distributions, m is the mean and C_i is the covariance matrix. To solve the player-assignment problem of two formations the Hungarian algorithm is used (Kuhn 1955). Hierarchical clustering does not automatically identify the number of unique formations. Therefore, for each phase of play, we varied the number of clusters from 3 to 15, creating a visual representation of the aggregated formations within each cluster before consulting with professional match analysts to determine the true number of unique formations within each phase of play. The final number of clusters was determined during several discussions with expert video analysts, using quantitative metrics (i.e. Silhouette values) to achieve an alignment among the involved experts. For different number of clusters, we plotted the cluster centroid formations (focusing on regions with good Silhouette values). For clusters of interest, we inspected the full set of detected formations to the analysts. Based on these observations, taking the Silhouette values into consideration, we decided on the number of clusters for each playing phase liaising with the experts. Once the final number of unique formations per phases of play was determined, the match analysts named each formation with a typical declaration (e.g. 4-4-2).

Fig. 4 shows the unique formations identified in the most frequently observed defensive phases of play: the midfield-block. Results for all the most-frequently observed in-possession phases of play, build-up, are provided in Appendix C. All the formations shown were familiar to the match analysts that inspected them. Indeed, the analyst's input was important in distinguishing the 4-2-3-1 formation from the 4-4-2: while the two appear similar in the figure, inspection of the individual observations that comprised each cluster indicated that the outside midfielders in the 4-2-3-1 (top-left plot) formed part of a triplet of attacking midfielders rather than two conventional wingers, as in the case of the 4-4-2 (top-center).

Formations #1−#4 in Fig. 4 are all variants of a player configuration that uses four defenders as a foundation and are distinguished by differences in the structure of the midfield and attacking players. Formation #3 sacrifices a forward for a central defensive midfielder, while formation #4 is a narrow 'Christmas tree' formation[8] (see

[7] Marco Bielsa's explanation of those ten formations can be found https://www.youtube.com/watch?v=qXt3rKnfbz8 (accessed 12/06/2020).

[8] The term Christmas tree formation—associated with a 4-3-2-1—has established in the football community (see https://thefalse9.com/2017/08/football-tactics-beginners-christmas-tree-formation.html, accessed 12/12/2020).

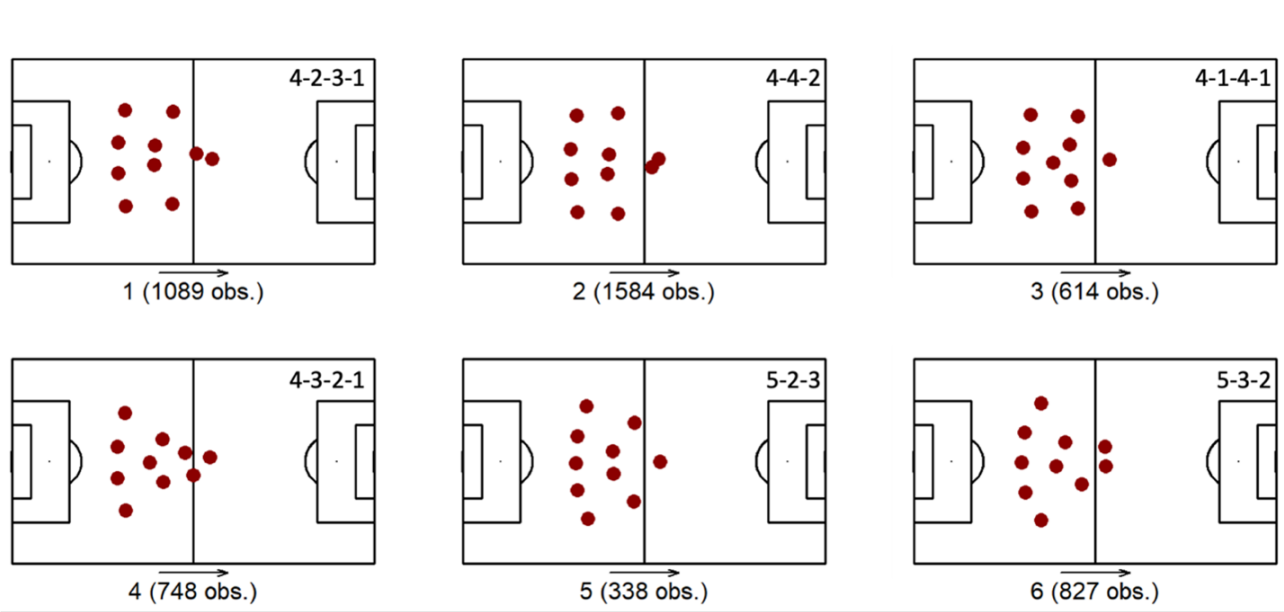

Fig. 4: Outcome of the clustering for mid-block including the number of observations (obs.) of our sample.

also: Janetzko et al. (2015)) with three defensive midfielders, two attacking midfielders and a lone forward. The remaining two formations show variants of player configurations with five defensive players.

5 Practical Applications

The primary aim of this paper is to describe our methodology for automating the process of formation detection per phase of play. In this section we highlight two practical applications of our methods that are enabled by our approach.

5.1 Formation versus Formation

A very common question in tactical discussion is: what is the most effective way to counter a particular formation (Wilson 2009)? This is a challenging question as it requires a large sample of formation observations as well as a contextualised formation detection per game-phase to attempt a quantitative answer. With over 13,081 formation observations measured over a sample of 1,803 Bundesliga games, we have a sufficient sample size to attempt a comparison of the relative performance of different formation options.

The most frequently observed offensive phase of play is the build-up; the most frequently observed formation in the build-up phase is the 2-4-3-1 (2 central defenders, 4 midfielders, 3 attacking midfielders and one forward), hereafter referred to as a 'two-defender' build-up. As the most frequently observed defensive phase of play is the mid-block, we attempt to quantify the performance of different mid-block formations in our data set when defending against a team using a two-defender build-up. Since goals are rare events in football[9] and not all shots have an equal chances to score a goal, the concept of expected goals (xG) is often used as a more granular proxy for the offensive contribution of a team (Anzer et al. 2021a).[10] XG values are only taken into consideration in periods of the match, where no formation change (see Appendix B) was detected. For such periods, xG values created from all phases of play were taken into consideration, since our experts claim that the formation in the basic phases of play (mid-block and build-up) has a latent influence on almost all situations.

The top row of Fig. 5 shows the strongest and weakest mid-block options. A 4-2-3-1 concedes, on average, 1.32 (SE: ±0.03; SD: ±0.81) xG[11] per match against the two-defender build-up, while the 5-2-3 (a five-defender formation) concedes 1.59±0.06 xG per match. The unconditional scoring rate of the two-defender build-up formation is 1.41±0.02 xG per match; the 4-2-3-1 therefore appears to significantly reduce the attacking threat of the two-defender build-up, while the 5-2-3 is the least effective counter-formation. The difference between the two amounts to 0.27 xG per game, or nearly nine goals over a 34-game season.

[9] For the given data set of seven seasons German Bundesliga, 3.1 goals were scored in average per match.

[10] The xG value of a shot denotes the a priori probability of a shot being converted to a goal, hence its value ranges from [0, 1]. The probability is estimated using both tracking and event data and applying a machine learning model, that was trained on more than 100.000 shots. A detailed description of the xG-model used can be found in Anzer et al. (2021a).

[11] Errors quoted are the standard error on the mean.

An ongoing discussion in the football tactics community is whether a build-up with two or three central defenders is more effective (Wilson 2009).[12] In the lower row of Fig. 5 we repeat the exercise for the 3-1-4-2 build-up formation, which utilizes three, rather than two, players at the back. The base scoring rate of the three-defender build-up is 1.36±0.03 xG per game, slightly below the two-defender build-up formation. This drops to just 1.17±0.08 xG per game when facing a 4-2-3-1 mid-block formation (lower-left)—the most effective counter-formation—and increases to 1.45±0.08 xG per game against a 4-1-4-1 (lower-right, the weakest mid-block formation against a 3-1-4-2). The conclusion is that the three-defender build-up formation appears to be more easily countered than the two-defender formation while showing less of an up-side benefit against other formations. Building up with two defenders is significantly more popular amongst Bundesliga teams than building with three defenders; our results indicate that the latter does indeed appear to be a weaker option.

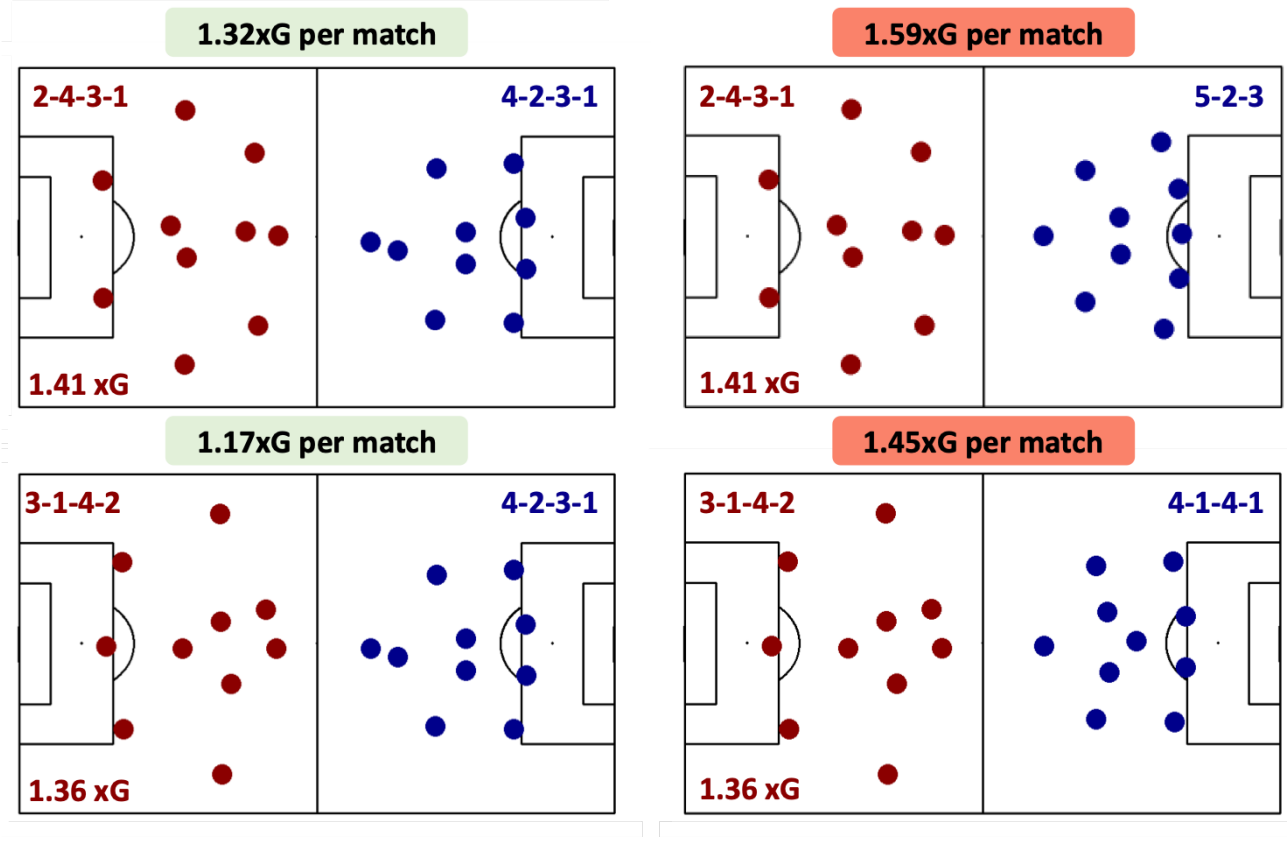

Fig. 5: Effectiveness of defensive formations (blue) against two (upper) and three (lower) player build-up (red).

Of course, even with a sample-size 1,803 matches, there are several potentially confounding factors, most notable if there is a preference for stronger (or weaker) teams to use a particular formation, although an initial inspection showed that every mid-block formation was used by at least 21 distinct teams once or more across the seven seasons. Future work (as described in the discussion) should investigate these confounding factors in significantly more detail.

5.2 Scouting the Tactical Preferences of Coaches

A major task that clubs must answer when seeking to fill a managerial vacancy is to ascertain the tactical preferences of the candidates and determine whether each represents continuity in the team's existing tactical style or a significant departure. While some clubs may specifically seek a completely new style of play, there are considerable risks associated with this. Most notably, a new tactical system will require different players, creating turnover in the playing style as the new manager implements their preferred tactical systems and sells the players that they do not require. Our methods allow a characterization of the types of formations that coaches prefer to use, which is often a clear indication of their overall strategic preferences.

[12] An exemplary blog-article can be found here https://thefalsefullback.de/2019/12/23/the-advantages-of-the-build-up-with-a-back-three/.

Individual teams demonstrate a preference for certain formations. Fig. 6 compares the frequency with which a selection of Bundesliga clubs, have utilized different formation options in the mid-block phase (radar-charts). Whereas Eintracht Frankfurt tends to play in a modern 5-3-2, Bayern Munich prefers the (somewhat similar) 4-2-3-1 or 4-1-4-1 systems. Another difference is that Bayern's formation in the build-up phase is rather traditional, utilizing two central defenders, whereas Eintracht Frankfurt more regularly builds up with three central defenders, which aligns with their significantly preferred 5-3-2 mid-block formation.

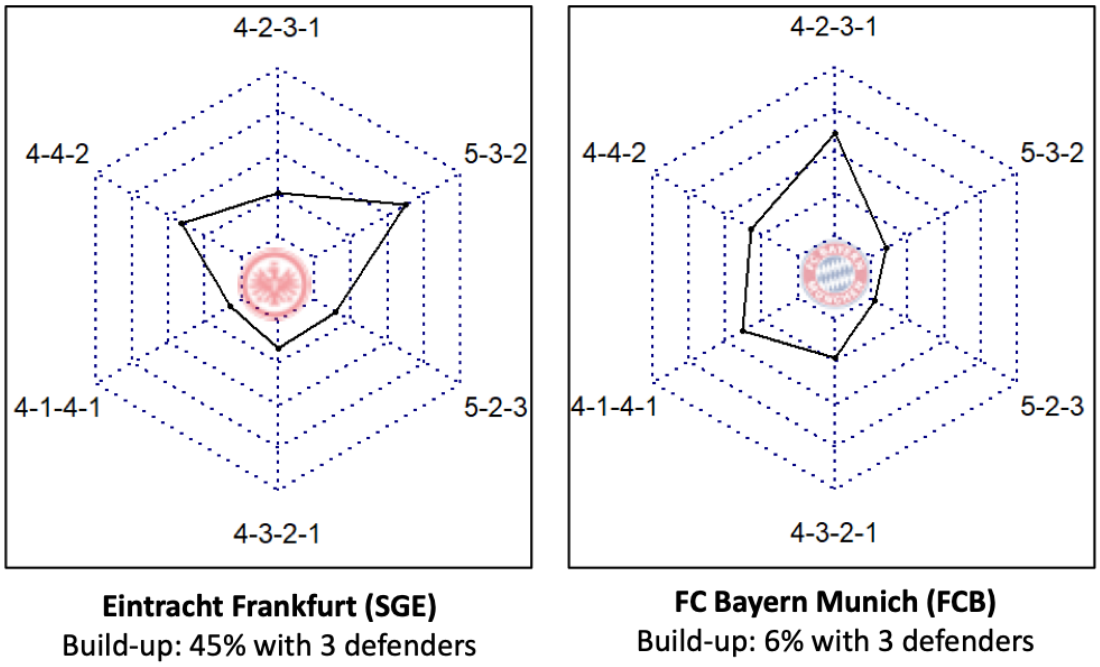

Fig. 6: Formations used by selected German Bundesliga clubs in the mid-block phase.

This visualization shows how different teams' preferences can be over a long period of seven seasons. These formation-profiles may often be determined by the key players of each team, some of whom may be particularly suited to one formation type. Bayern Munich's success in the past few seasons has been greatly influenced by the central axis consisting of Jérôme Boateng, Robert Lewandowski and individually strong wingers like Frank Ribéry, Arjen Robben, Kingsley Coman or Serge Gnabry. Our match analysts agreed the formations most frequently utilized by Bayern's coaches over the previous seven years—a 4-2-3-1 or a 4-1-4-1—are the most suitable formations for the players that were at the club.

Fig. 7 demonstrates the tactical preferences of four Bayern-coaches in the mid-block phase over this period. Guardiola, Heynckes and Flick all maintained a similar strategic approach, and all three had successful tenures. Only Niko Kovac is generally perceived to have been a failure. One reason, often referenced in the media, is that he was unwilling to part with the 5-3-2 build-up formation—with which he experienced success at his previous club, Eintracht Frankfurt—instead of adapting his style of play to exploit the full potential of the players at Bayern. The appointment of Niko Kovac did not represent continuity in Bayern's playing style.

A valuable use-case of our methods is in the search for future managers with a similar playing style (at least in terms of formations) to the existing approach at the hiring club. Fig. 8 shows a short-list of coaches that could be touted as potential successors of Hansi Flick—head coach at FC Bayern from 2019 until 2021. By comparing the coaches' formation profiles (black)[13] with that of FC Bayern (red) a similarity metric (top left in Fig. 8) can be calculated. Although Julian Nagelsmann (currently head coach at FC Bayern Munich) is often considered to be one of the biggest German coaching talents, his preferred formations diverge significantly from Bayern's existing style, resulting in a similarity score of only 44%. Jürgen Klopp and Thomas Tuchel represent intermediate fits (72% and 73%), but Ralph Hasenhüttel, currently head coach of FC Southampton, is the best fit for FC Bayern in our managerial database, with a similarity score of 81%. Again, the choice of a coach relies on various factors, not solely on formations played in one or two phases of play (as displayed here). However, our approach provides evidence for one key component, which can drastically help club's management to take informed decisions.

[13] Note that only data from the respective coaches' time in the German Bundesliga (2013/2014-2018/2019) are used for this analysis.

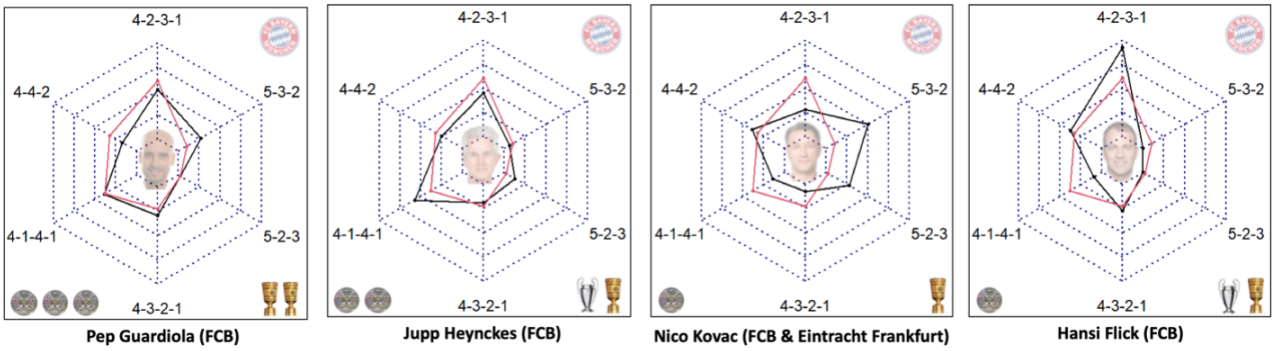

Fig. 7: FC Bayern Munich coaches by their formation (black) in comparison to the overall Bayern profile (red). The data from all coaches and FC Bayern are aggregated over the seasons 2013/2014 to 2019/2020. The trophies (Bundesliga championship, DFB-Cup and UEFA Champions-League) that each coach earned at his time at FC Bayern are displayed.

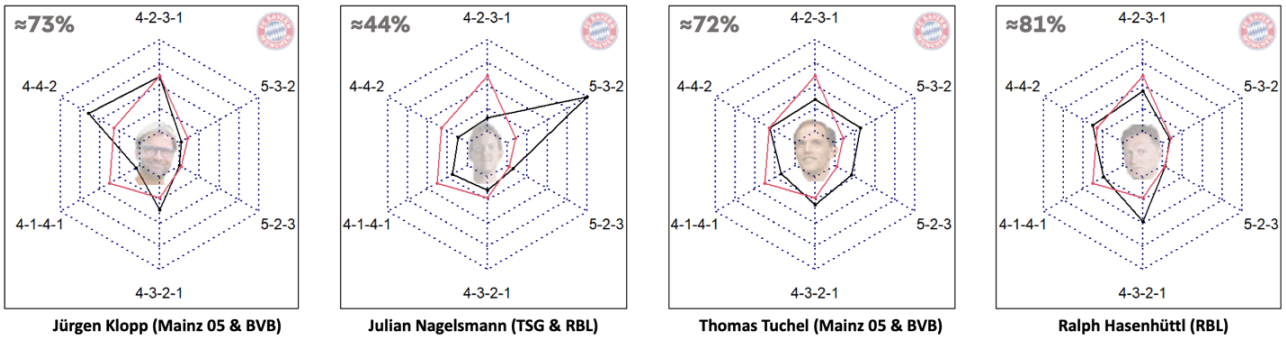

Fig. 8: Formation similarity. Who is the best fit for FC Bayern Munich? Top left the similarity of each coach compared to FC Bayern is displayed.

6 Discussion

The availability of accurate and league-wide tracking data has motivated several research investigations into team formations, the basis of team-tactics in football. The main objective of this paper was to detect phases of play as a preliminary for contextualized formation analysis. Previous work has attempted to detect only single specific phases of play, such as counterattacking (Fassmeyer et al. 2021; Hobbs et al. 2018) or counterpressing (Bauer et al. 2021). For the first time, we present a method for classifying games into five distinct phases of play. While the phases of plays used in our approach are well established among football experts, their exact definitions may vary depending on a club's playing philosophy. The definitions we used in the labeling process were consolidated among professional match analysts of German Bundesliga clubs. In future work, a proper qualitative study, that formalizes and extends the framework presented in Fig. 3 should be conducted in order to have a proper scientific baseline for further investigations on phases of play—a well established theory in professional football. In this context, our work shows, that (a) phases of play can be defined and identified by experts with an appropriate accordance, and (b) that these phases of play influence the collective behavior of teams (i.e. their formations) significantly.

We used this time-domain classification to measure team formations in distinct phases of play, achieving a spatial classification. Phases of play measurement and classification of formations represents a major step towards decrypting the complexities of strategy in football and provide a new insight into the tactical preferences of individual managers and coaches. While the methodology for the formation classification is mostly similar to the one introduced in Shaw et al. (2019), a crucial difference is not only that five different phases of play are considered separately, but also how closely subject experts were involved throughout the whole project. Selecting the final number of clusters purely on a statistical measure, would not lead to the same results as when taking expert-knowledge into consideration as well. This interplay between data-science and domain experts also turned out to be beneficial for the contextualisation of the clusters, as well as for the identification of meaningful use-cases (see also Andrienko et al. (2019), Herold et al. (2019), Goes et al. (2020a), and Rein et al. (2016)).

The benefit of our approach to practitioners is threefold: by automatically detecting phases of play of the next opponent over an arbitrary number of their previous games we save the match analysis departments significant amounts of time. An objective long-term analysis enables us to assess which formations are the most effective counter to a particular reference formation, drastically supporting a coaches decision-making process of how to

approach the next opponent. Last but not least, we show a unique use-case for club decision-makers on how to quantify candidate coaches' tactical style and identify those that represent continuity to the current playing style of the club.

Besides these applications, the full potential of this approach is yet to be unlocked. Future studies could analyse the interplay of different formations more thoroughly and control for confounding factors. On one hand, quantitative tendencies should always be evaluated by qualitative analysis, i.e. by analysing video footage of formation-pairings of interest to generate expert-based ad- and disadvantages when playing a specific formation (against another). On the other hand, the most critical confounding factor (the strength of a team playing a formation) should be modelled with a rating system of teams (see e.g., Baysal et al. (2016)) and used to validate the hypothesis presented in Section 5.1 Additionally, when evaluating a coach's tactical fingerprint, all phases of play as well as other factors could be taken into consideration.

Acknowledgements

This work would not have been possible without the perspective of professional match-analysts from world class teams who helped us to define relevant features and spend much time evaluating (intermediate) results. We would cordially like to thank Dr. Stephan Nopp and Christofer Clemens (head match-analysts of the German mens National team), Jannis Scheibe (head match-analyst of the German U21 mens national team), Leonard Höhn (head match-analyst of the German women national team) as well as Sebastian Geißler (former match-analyst of Borussia Mönchengladbach). Additionally, the authors would like to thank Dr. Hendrik Weber and Deutsche Fußball Liga (DFL) / Sportec Solutions GmbH for providing the positional and event data.

References

Alexander, Jeremy P. et al. (2019). "The influence of match phase and field position on collective team behaviour in Australian Rules football". In: *Journal of Sports Sciences* 37.15, pp. 1699–1707. ISSN: 1466447X. DOI: 10.1080/02640414.2019.1586077. URL: https://doi.org/10.1080/02640414.2019.1586077 (cit. on p. 3).

Andrienko, Gennady et al. (2017). "Visual analysis of pressure in football". In: *Data Mining and Knowledge Discovery* 31.6, pp. 1793–1839. ISSN: 1573756X. DOI: 10.1007/s10618-017-0513-2 (cit. on p. 18).

Andrienko, Gennady et al. (2019). "Constructing Spaces and Times for Tactical Analysis in Football". In: *IEEE Transactions on Visualization and Computer Graphics* 27.4, pp. 2280–2297. DOI: 10.1109/TVCG.2019.2952129. URL: https://ieeexplore.ieee.org/document/8894420 (cit. on pp. 2, 3, 12).

António, Doutor et al. (2014). "The emergence of team synchronization during the soccer match: understanding the influence of the level of opposition, game phase and field zone". In: (cit. on p. 4).

Anzer, Gabriel and Pascal Bauer (2021a). "A Goal Scoring Probability Model based on Synchronized Positional and Event Data". In: *Frontiers in Sports and Active Learning (Special Issue: Using Artificial Intelligence to Enhance Sport Performance)* 3.0, pp. 1–18. DOI: 10.3389/fspor.2021.624475 (cit. on pp. 2, 9).

— (2022). "Expected Passes—Determining the Difficulty of a Pass in Football (Soccer) Using Spatio-Temporal Data". In: *Data Mining and Knowledge Discovery, Springer US*. ISSN: 1573-756X. DOI: 10.1007/s10618-021-00810-3 (cit. on p. 2).

Anzer, Gabriel, Pascal Bauer, and Ulf Brefeld (2021b). "The origins of goals in the German Bundesliga". In: *Journal of Sport Science*. DOI: 10.1080/02640414.2021.1943981. URL: https://www.tandfonline.com/doi/full/10.1080/02640414.2021.1943981 (cit. on p. 4).

Araújo, Duarte et al. (2021). *Artificial Intelligence in Sport Performance Analysis*. April. ISBN: 9781000380125. DOI: 10.4324/9781003163589 (cit. on p. 2).

Atmosukarto, Indriyati et al. (2013). "Automatic recognition of offensive team formation in american football plays". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 991–998. ISSN: 21607508. DOI: 10.1109/CVPRW.2013.144 (cit. on p. 2).

Balague, Natàlia et al. (2013). "Overview of complex systems in sport". In: *Journal of Systems Science and Complexity* 26.1, pp. 4–13. ISSN: 15597067. DOI: 10.1007/s11424-013-2285-0 (cit. on p. 1).

Bauer, Pascal and Gabriel Anzer (2021). "Data-driven detection of counterpressing in professional football—A supervised machine learning task based on synchronized positional and event data with expert-based feature extraction". In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-021-00763-7. URL: https://doi.org/10.1007/s10618-021-00763-7 (cit. on pp. 3, 5, 6, 12).

Baysal, Sermetcan and Pinar Duygulu (2016). "Sentioscope: A Soccer Player Tracking System Using Model Field Particles". In: *IEEE Transactions on Circuits and Systems for Video Technology* 26.7, pp. 1350–1362. ISSN: 10518215. DOI: 10.1109/TCSVT.2015.2455713 (cit. on pp. 2, 13).

Beetz, Michael et al. (2006). "Camera-based observation of football games for analyzing multi-agent activities". In: *Proceedings of the International Conference on Autonomous Agents* 2006, pp. 42–49. DOI: 10.1145/1160633.1160638 (cit. on p. 1).

Bialkowski, Alina et al. (2014a). "Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data". In: *IEEE International Conference on Data Mining, ICDM (Proceeding)* January, pp. 725–730. ISSN: 15504786. DOI: 10.1109/ICDM.2014.133 (cit. on p. 2).

Bialkowski, Alina et al. (2014b). ""Win at Home and Draw Away": Automatic Formation Analysis Highlighting the Differences in Home and Away Team Behaviors". In: *MIT Sloan Sports Analytics Conference* June 2016. URL: http://www.sloansportsconference.com/wp-content/uploads/2014/02/2014_SSAC_Win-at-Home-Draw-Away.pdf (cit. on p. 6).

Bialkowski, Alina et al. (2015). "Identifying team style in soccer using formations learned from spatiotemporal tracking data". In: *IEEE International Conference on Data Mining Workshops, ICDMW* January, pp. 9–14. ISSN: 23759259. DOI: 10.1109/ICDMW.2014.167 (cit. on pp. 2, 6).

Bialkowski, Alina et al. (2016). "Discovering team structures in soccer from spatiotemporal data". In: *IEEE Transactions on Knowledge and Data Engineering* 28.10, pp. 2596–2605. ISSN: 10414347. DOI: 10.1109/TKDE.2016.2581158 (cit. on pp. 2, 3, 6).

Bianchi, Federico, Tullio Facchinetti, and Paola Zuccolotto (2017). "Role revolution: Towards a new meaning of positions in basketball". In: *Electronic Journal of Applied Statistical Analysis* 10.3, pp. 712–734. ISSN: 20705948. DOI: 10.1285/i20705948v10n3p712 (cit. on p. 2).

Boon, Bart H. and Gerard Sierksma (2003). "Team formation: Matching quality supply and quality demand". In: *European Journal of Operational Research* 148.2, pp. 277–292. ISSN: 03772217. DOI: 10.1016/S0377-2217(02)00684-7 (cit. on p. 2).

Borrie, Andrew, Gudberg K. Jonsson, and Magnus S. Magnusson (2002). "Temporal pattern analysis and its applicability in sport: An explanation and exemplar data". In: *Journal of Sports Sciences* 20.10, pp. 845–852. ISSN: 02640414. DOI: 10.1080/026404102320675675 (cit. on p. 3).

Bradley, Paul S. et al. (2011). "The effect of playing formation on high-intensity running and technical profiles in English FA premier League soccer matches". In: *Journal of Sports Sciences* 29.8, pp. 821–830. ISSN: 02640414. DOI: 10.1080/02640414.2011.561868 (cit. on p. 2).

Brefeld, Ulf, Jan Lasek, and Sebastian Mair (2019). "Probabilistic movement models and zones of control". In: *Machine Learning* 108.1, pp. 127–147. ISSN: 15730565. DOI: 10.1007/s10994-018-5725-1. URL: https://doi.org/10.1007/s10994-018-5725-1 (cit. on p. 1).

Budak, Gerçek et al. (2019). "New mathematical models for team formation of sports clubs before the match". In: *Central European Journal of Operations Research* 27.1, pp. 93–109. ISSN: 16139178. DOI: 10.1007/s10100-017-0491-x (cit. on p. 2).

Carling, Christopher (2011). "Influence of opposition team formation on physical and skill-related performance in a professional soccer team". In: *European Journal of Sport Science* 11.3, pp. 155–164. ISSN: 17461391. DOI: 10.1080/17461391.2010.499972 (cit. on p. 2).

Casal, Claudio A. et al. (2015). "Analysis of corner kick success in elite football". In: *International Journal of Performance Analysis in Sport* 15.2, pp. 430–451. ISSN: 14748185. DOI: 10.1080/24748668.2015.11868805 (cit. on p. 7).

Chen, Sheng et al. (2014). "Play Type Recognition in Real-World Football Video". In: *IEEE Winter Conference on Applications of Computer Vision*, pp. 652–659. DOI: 10.1109/WACV.2014.6836040. (cit. on p. 3).

Cintia, Paolo et al. (Dec. 2015). "The harsh rule of the goals: Data-driven performance indicators for football teams". In: *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*. Institute of Electrical and Electronics Engineers Inc. ISBN: 9781467382731. DOI: 10.1109/DSAA.2015.7344823 (cit. on p. 2).

Danisik, Norbert, Peter Lacko, and Michal Farkas (Oct. 2018). "Football match prediction using players attributes". In: *DISA 2018 - IEEE World Symposium on Digital Intelligence for Systems and Machines, Proceedings*. Institute of Electrical and Electronics Engineers Inc., pp. 201–206. ISBN: 9781538651025. DOI: 10.1109/DISA.2018.8490613 (cit. on p. 2).

Decroos, Tom, Jan Van Haaren, and Jesse Davis (2018). "Automatic discovery of tactics in spatio-temporal soccer match data". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 223–232. DOI: 10.1145/3219819.3219832 (cit. on pp. 2, 3, 5).

Decroos, Tom et al. (2019). "Actions speak louder than goals: Valuing player actions in soccer". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1, pp. 1851–1861. DOI: 10.1145/3292500.3330758 (cit. on p. 2).

Dick, Uwe and Ulf Brefeld (2019). "Learning to Rate Player Positioning in Soccer". In: *Big Data* 7.1, pp. 71–82. ISSN: 2167647X. DOI: 10.1089/big.2018.0054 (cit. on p. 6).

Fassmeyer, Dennis et al. (2021). "Toward Automatically Labeling Situations in Soccer". In: *Frontiers in Sports and Active Living* 3.November. DOI: 10.3389/fspor.2021.725431 (cit. on pp. 3, 5, 12).

Fernandez, Javier and Luke Bornn (2018). "Wide Open Spaces : A statistical technique for measuring space creation in professional soccer". In: *MIT Sloan Sports Analytics Conference, Boston (USA)*, pp. 1–19 (cit. on p. 1).

Fernando, T et al. (2015). "Discovering Methods of Scoring in Soccer Using Tracking Data". In: *KDD Workshop on Large-Scale Sports Analytics*, pp. 1–4. URL: https://large-scale-sports-analytics.org/Large-Scale-Sports-Analytics/Submissions2015_files/paperID19-Tharindu.pdf (cit. on p. 3).

Fujii, Keisuke (2021). "Data-Driven Analysis for Understanding Team Sports Behaviors". In: *Journal of Robotics and Mechatronics* 33.3, pp. 505–514. ISSN: 0915-3942. DOI: 10.20965/jrm.2021.p0505 (cit. on p. 1).

Goes, F. R. et al. (2020a). "Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review". In: *European Journal of Sport Science* 0.0, pp. 1–16. ISSN: 15367290. DOI: 10.1080/17461391.2020.1747552. URL: https://doi.org/10.1080/17461391.2020.1747552 (cit. on pp. 2, 12).

Goes, Floris R. et al. (2020b). "The tactics of successful attacks in professional association football—large-scale spatiotemporal alanalysis of dynamic subgroups using position tracking data". In: *Journal of Sports Sciences* 39.5, pp. 523–532. DOI: 10.1080/02640414.2020.1834689 (cit. on p. 1).

Goutte, Cyril and Eric Gaussier (2005). "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation". In: *Lecture Notes in Computer Science*. Vol. 3408. Springer Verlag, pp. 345–359. URL: https://link.springer.com/chapter/10.1007/978-3-540-31865-1_25 (cit. on p. 6).

Gréhaigne, Jean Francis, Paul Godbout, and Daniel Bouthier (1999). "The foundations of tactics and strategy in team sports". In: *Journal of Teaching in Physical Education* 18.2, pp. 159–174. ISSN: 02735024. DOI: 10.1123/jtpe.18.2.159 (cit. on p. 2).

Grunz, Andreas, Daniel Memmert, and Jürgen Perl (2012). "Tactical pattern recognition in soccer games by means of special self-organizing maps". In: *Human Movement Science* 31.2, pp. 334–343. ISSN: 01679457. DOI: 10.1016/j.humov.2011.02.008. URL: http://dx.doi.org/10.1016/j.humov.2011.02.008 (cit. on p. 3).

Gudmundsson, Joachim and Michael Horton (2017a). "Spatio-temporal analysis of team sports". In: *ACM Computing Surveys* 50.2, pp. 1–34. ISSN: 15577341. DOI: 10.1145/3054132 (cit. on p. 1).

Gudmundsson, Joachim, Patrick Laube, and Thomas Wolle (2017b). "Movement Patterns in Spatio-Temporal Data". In: *Shekhar S., Xiong H., Zhou X. (eds) Encyclopedia of GIS. Springer, Cham.* DOI: 10.1007/978-3-319-17885-1{_}823 (cit. on pp. 1–3).

Haaren, Jan Van et al. (2013). *Machine Learning and Data Mining for Sports Analytics*. September, p. 2013. ISBN: 9783030649111. DOI: 10.1007/978-3-030-64912-8 (cit. on p. 2).

Herold, M. et al. (2019). "Machine learning in men's professional football: Current applications and future directions for improving attacking play". In: *International Journal of Sports Science and Coaching* 14.6. ISSN: 2048397X. DOI: 10.1177/1747954119879350 (cit. on pp. 2, 12).

Hobbs, Jennifer et al. (2018). "Quantifying the Value of Transitions in Soccer via Spatiotemporal Trajectory Clustering". In: *MIT Sloan Sports Analytics Conference, Boston (USA)*, pp. 1–11 (cit. on pp. 3, 5, 12).

Hochstedler, Jeremy and Paul T Gagnon (2017). "American Football Route Identification Using Supervised Machine Learning". In: *MIT Sloan Sports Analytics Conference, Boston (USA)*, pp. 1–11 (cit. on pp. 2, 3).

Intille, Stephen S. and Aaron F. Bobick (1999). "Framework for recognizing multi-agent action from visual evidence". In: *Proceedings of the National Conference on Artificial Intelligence*, pp. 518–525 (cit. on p. 2).

Janetzko, Halld'Or et al. (2015). "Feature-driven visual analytics of soccer data". In: *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings*. ISBN: 9781479962273. DOI: 10.1109/VAST.2014.7042477 (cit. on p. 9).

Kempe, Matthias, Andreas Grunz, and Daniel Memmert (2015). "Detecting tactical patterns in basketball: Comparison of merge self-organising maps and dynamic controlled neural networks". In: *European Journal of Sport Science* 15.4, pp. 249–255. ISSN: 15367290. DOI: 10.1080/17461391.2014.933882. URL: http://dx.doi.org/10.1080/17461391.2014.933882 (cit. on p. 3).

Kempe, Matthias et al. (2014). "Possession vs. Direct Play: Evaluating Tactical Behavior in Elite Soccer". In: *International Journal of Sports Science* 4.6A, pp. 35–41. ISSN: 2169-8791. DOI: 10.5923/s.sports.201401.05 (cit. on pp. 3, 4).

Kuhn, H.W. (1955). "The Hungarian method for the assignment problem". In: *Naval Research Logistics* 2, pp. 83–97. DOI: 10.1002/nav.3800020109 (cit. on p. 8).

Li, Ruonan and Rama Chellappa (2010). "Group motion segmentation using a spatio-temporal driving force model". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2038–2045. ISSN: 10636919. DOI: 10.1109/CVPR.2010.5539880 (cit. on p. 3).

Link, Daniel and Martin Hoernig (2017). "Individual ball possession in soccer". In: *PLoS ONE* 12.7, pp. 1–15. ISSN: 19326203. DOI: 10.1371/journal.pone.0179953 (cit. on p. 1).

Linke, Daniel, Daniel Link, and Martin Lames (2018). "Validation of electronic performance and tracking systems EPTS under field conditions". In: *PLoS ONE* 13.7, pp. 1–20. ISSN: 19326203. DOI: 10.1371/journal.pone.0199519 (cit. on p. 4).

Linke, Daniel, Daniel Link, and Martin Lames (2020). "Football-specific validity of TRACAB's optical video tracking systems". In: *PLoS ONE* 15.3, pp. 1–17. ISSN: 19326203. DOI: 10.1371/journal.pone.0230179 (cit. on p. 4).

Lucey, Patrick et al. (2013). "Representing and discovering adversarial team behaviors using player roles". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2706–2713. ISSN: 10636919. DOI: 10.1109/CVPR.2013.349 (cit. on pp. 2, 3).

Lucey, Patrick et al. (2014). ""Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data". In: *Proc. 8th Annual MIT Sloan Sports Analytics Conference*, pp. 1–9. URL: http://www.sloansportsconference.com/?p=15790 (cit. on p. 3).

Montoliu, Raúl et al. (2015). "Team activity recognition in Association Football using a Bag-of-Words-based method". In: *Human Movement Science* 41, pp. 165–178. ISSN: 18727646. DOI: 10.1016/j.humov.2015.03.007 (cit. on p. 3).

Müller-Budack, Eric et al. (2019). ""Does 4-4-2 exist?" – An analytics approach to understand and classify football team formations in single match situations". In: *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports (Nice, France)* MMSports '.September, pp. 25–33. DOI: 10.1145/3347318.3355527 (cit. on pp. 2, 3, 6).

Narizuka, Takuma and Yoshihiro Yamazaki (2019). "Clustering algorithm for formations in football games". In: *Scientific Reports* 9.1, pp. 1–8. ISSN: 20452322. DOI: 10.1038/s41598-019-48623-1. URL: http://dx.doi.org/10.1038/s41598-019-48623-1 (cit. on p. 2).

Olkin, I. and F. Pukelsheim (1982). "The distance between two random vectors with given dispersion matrices". In: *Linear Algebra and Its Applications* 48.C, pp. 257–263. ISSN: 00243795. DOI: 10.1016/0024-3795(82)90112-4 (cit. on p. 8).

Pantzalis, Victor Chazan and Christos Tjortjis (July 2020). "Sports Analytics for Football League Table and Player Performance Prediction". In: *11th International Conference on Information, Intelligence, Systems and Applications, IISA 2020*. Institute of Electrical and Electronics Engineers Inc. ISBN: 9780738123462. DOI: 10.1109/IISA50023.2020.9284352 (cit. on p. 2).

Pappalardo, Luca et al. (2019a). "A public data set of spatio-temporal match events in soccer competitions". In: *Scientific Data* 6.1, pp. 1–15. ISSN: 20524463. DOI: 10.1038/s41597-019-0247-7. URL: http://dx.doi.org/10.1038/s41597-019-0247-7 (cit. on p. 2).

Pappalardo, Luca et al. (2019b). "PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach". In: *ACM Transactions on Intelligent Systems and Technology* 10.5. ISSN: 21576912. DOI: 10.1145/3343172 (cit. on p. 2).

Perse, Matej et al. (2006). "A Template-Based Multi-Player Action Recognition of the Basketball Game". In: *CVBASE '06 - Proceedings of ECCV Workshop on Computer Vision*, pp. 71–82 (cit. on p. 3).

Pettersen, Svein Arne et al. (2014). "Soccer video and player position dataset". In: *Proceedings of the 5th ACM Multimedia Systems Conference, MMSys 2014 (Singapore, March 2014)*, pp. 18–23. DOI: 10.1145/2557642.2563677 (cit. on p. 4).

Pfeiffer, Mark and Jürgen Perl (2015). "Analysis of tactical defensive behavior in team handball by means of artificial neural networks". In: *IFAC-PapersOnLine* 28.1, pp. 784–785. ISSN: 24058963. DOI: 10.1016/j.ifacol.2015.05.169 (cit. on p. 3).

Power, Paul et al. (2017). ""Not all passes are created equal:" Objectively measuring the risk and reward of passes in soccer from tracking data". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Part F1296, pp. 1605–1613. DOI: 10.1145/3097983.3098051 (cit. on p. 4).

Redwood-Brown, A., W. Cranton, and C. Sunderland (2012). "Validation of a real-time video analysis system for soccer". In: *International Journal of Sports Medicine* 33.8, pp. 635–640. ISSN: 01724622. DOI: 10.1055/s-0032-1306326 (cit. on p. 4).

Reep, C. and B. Benjamin (1968). "Skill and Chance in Association Football Author". In: *Journal of the Royal Statistical Society* 131.4, pp. 581–585. ISSN: 14698005. URL: https://www.jstor.org/stable/2343726?seq=1 (cit. on p. 2).

Rein, Robert and Daniel Memmert (2016). "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science". In: *SpringerPlus* 5.1. ISSN: 21931801. DOI: 10.1186/s40064-016-3108-2 (cit. on pp. 2, 3, 12).

Ric, Angel et al. (2021). "Football Analytics 2021: The role of context in transferring analytics to the pitch". In: p. 158 (cit. on p. 2).

Sampaio, J. and V. MaçÃs (2012). "Measuring tactical behaviour in football". In: *International Journal of Sports Medicine* 33.5, pp. 395–401. ISSN: 01724622. DOI: 10.1055/s-0031-1301320 (cit. on p. 3).

Sarmento, Hugo et al. (2018). "What Performance Analysts Need to Know About Research Trends in Association Football (2012–2016): A Systematic Review". In: *Sports Medicine* 48.4, pp. 799–836. ISSN: 11792035. DOI: 10.1007/s40279-017-0836-6 (cit. on p. 1).

Shaw, Laurie and Mark Glickman (2019). "Dynamic analysis of team strategy in professional football". In: *Barça sports analytics summit*, pp. 1–13 (cit. on pp. 2, 3, 6–8, 12).

Siddiquie, Behjat, Yaser Yacoob, and Larry S Davis (2009). "Recognizing Plays in American Football Videos". In: *Technical Report, University of Maryland* 1.1, pp. 1–8. URL: http://www.researchgate.net/publication/228519111_Recognizing_Plays_in_American_Football_Videos (cit. on p. 3).

Stein, Manuel et al. (2017). "How to Make Sense of Team Sport Data: From Acquisition to Data Modeling and Research Aspects". In: *Data* 2.1, p. 2. ISSN: 2306-5729. DOI: 10.3390/data2010002 (cit. on pp. 2, 3).

Stracuzzi, David J et al. (2011). "An Application of Transfer to American Football : From Observation of Raw Video to Control in a Simulated Environment An Application of Transfer to American Football : From Observation of Raw Video to Control in a Simulated Environment". In: *AI Magazine* 32.2. DOI: 10.1609/aimag.v32i2.2336 (cit. on p. 3).

Taberner, Matt et al. (2020). "Interchangeability of position tracking technologies; can we merge the data?" In: *Science and Medicine in Football* 4.1, pp. 76–81. ISSN: 24734446. DOI: 10.1080/24733938.2019.1634279. URL: https://doi.org/10.1080/24733938.2019.1634279 (cit. on p. 4).

Teoldo, Israel, Júlio Manuel, and Pablo Juan Greco (2009). "Tactical Principles of Soccer Game: concepts and application". In: *Motriz. Journal of Physical Education. UNESP* 15.3, pp. 657–668. ISSN: 1980-6574. DOI: 10.5016/2488 (cit. on p. 2).

Thinh, Nguyen Hong et al. (Oct. 2019). "A video-based tracking system for football player analysis using Efficient Convolution Operators". In: *International Conference on Advanced Technologies for Communications*. Vol. 2019-Octob. IEEE Computer Society, pp. 149–154. ISBN: 9781728123929. DOI: 10.1109/ATC.2019.8924544 (cit. on p. 2).

Tierney, Peter J. et al. (2016). "Match play demands of 11 versus 11 professional football using Global Positioning System tracking: Variations across common playing formations". In: *Human Movement Science* 49.October, pp. 1–8. ISSN: 18727646. DOI: 10.1016/j.humov.2016.05.007. URL: http://dx.doi.org/10.1016/j.humov.2016.05.007 (cit. on p. 2).

Vilamitjana, Javier J. et al. (2021). "High-intensity activity according to playing position with different team formations in soccer". In: *Acta Gymnica* 51.March, pp. 2–7. ISSN: 23364920. DOI: 10.5507/AG.2021.003. URL: https://doi.org/10.5507/ag.2021.003 (cit. on p. 2).

Vilar, Luís et al. (2013). "Science of winning soccer: Emergent pattern-forming dynamics in association football". In: *Journal of Systems Science and Complexity* 26.1, pp. 73–84. ISSN: 15597067. DOI: 10.1007/s11424-013-2286-z (cit. on p. 2).

Vogelbein, Martin, Stephan Nopp, and Anita Hökelmann (2014). "Defensive transition in soccer - are prompt possession regains a measure of success? A quantitative analysis of German Fußball-Bundesliga 2010/2011". In: *Journal of Sports Sciences* 32.11, pp. 1076–1083. ISSN: 1466447X. DOI: 10.1080/02640414.2013.879671. URL: http://dx.doi.org/10.1080/02640414.2013.879671 (cit. on p. 3).

Wang, Jian and Jia Zhang (2015). "A win-win team formation problem based on the negotiation". In: *Engineering Applications of Artificial Intelligence* 44, pp. 137–152. ISSN: 09521976. DOI: 10.1016/j.engappai.2015.06.001. URL: http://dx.doi.org/10.1016/j.engappai.2015.06.001 (cit. on pp. 2, 3).

Wang, Kuan-Chieh and Richard Zemel (2016). "Classifying NBA Offensive Plays Using Neural Networks". In: *MIT Sloan Sports Analytics Conference*, pp. 1–9 (cit. on p. 6).

Wang, Yikang, Hao Wang, and Mingyue Qiu (July 2020). "Performance Analysis of Everton Football Club Based on Tracking Data". In: *Proceedings of 2020 IEEE International Conference on Power, Intelligent Computing and Systems, ICPICS 2020*. Institute of Electrical and Electronics Engineers Inc., pp. 49–53. ISBN: 9781728198736. DOI: 10.1109/ICPICS50287.2020.9202246 (cit. on p. 2).

Ward, Tr and H Joe (1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301, pp. 236–244 (cit. on p. 8).

Wei, Xinyu et al. (2013). "Large-scale analysis of formations in soccer". In: *2013 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2013*. DOI: 10.1109/DICTA.2013.6691503 (cit. on pp. 2, 6).

Wilson, Jonathan (2009). *Inverting the pyramid, a history of football tactics*. London: Orion, p. 374. ISBN: 978-1-4091-0204-5 (cit. on pp. 2, 9, 10).

Xu, Haoran (Mar. 2021). "Prediction on Bundesliga Games Based on Decision Tree Algorithm". In: *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*. IEEE, pp. 234–238. ISBN: 978-1-6654-1540-8. DOI: 10.1109/ICBAIE52039.2021.9389986. URL: https://ieeexplore.ieee.org/document/9389986/ (cit. on p. 2).

Yeh, Raymond A. et al. (2019). "Diverse generation for multi-agent sports games". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June, pp. 4605–4614. ISSN: 10636919. DOI: 10.1109/CVPR.2019.00474 (cit. on p. 2).

Zheng, Stephan, Yisong Yue, and Patrick Lucey (2016). "Generating long-term trajectories using deep hierarchical networks". In: *Advances in Neural Information Processing Systems* Nips, pp. 1551–1559. ISSN: 10495258 (cit. on p. 6).

Appendix

A Detecting Phases of Play with a CNN

A schematic visualization of the CNN-architecture is displayed in Fig. 9. The input images are of size 105x68

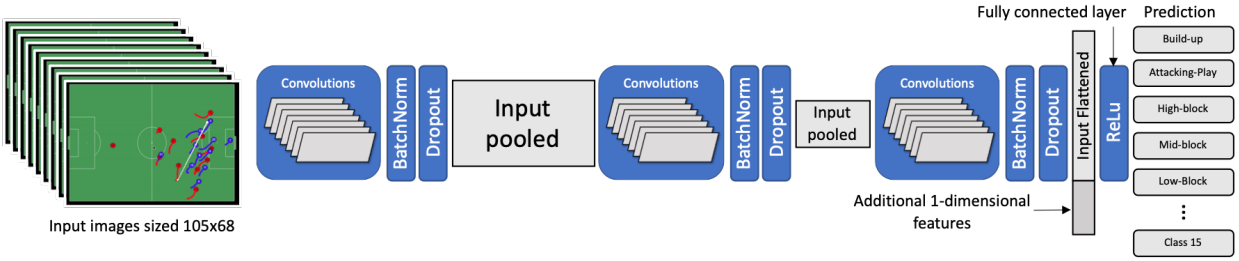

Fig. 9: Schematic architecture of the CNN predicting the phases of play.

pixels—corresponding to the typical dimensions of a football pitch in meters—and consist of up to nine layers (e.g. home-team positions, away-team positions, ball) containing information from a half-second period of the game. To feed time-related information to the CNN, player trajectories, weighted with a linearly decreasing function of time, were added to each image. To differentiate home team, away team and the ball, each information is imported as a separate layer. Additional layers contain smoothed speed values, which slightly improved the accuracy of our prediction. Finally, the CNN predicts one out of 15 possible phases of play[14] for each frame, although in this work only the phases shown in Fig. 1 in white boxes are taken into consideration. We split the labeled data into 75% training and 25% test data. On the training data we used a Bayesian hyper-parameter optimization and a 5-fold cross-validation. The final model has a batch size of 32 and was trained over 10 epochs. The imbalanced dispersion of the phases of play (see Table 2) was addressed by resampling and weighted inputs for each batch. The best performing CNN yielding the highest F_1-score on the test data consists of a base model with three convolutional layers, one fully connected layer and one concatenation with one-dimensional features. The additional features include for example a binary indicator whether the ball is in play or the game is interrupted during the corresponding frame. Another feature, which is included in the positional data, is the information which team is currently in possession of the ball. This base model is applied at 13 consecutive time points (roughly half a second) and the outputs are combined using a 1-D convolution. It uses a drop-out of 50% and a ReLu-activation function. To avoid noisy outcomes in the framewise prediction, the outcome is smoothed afterwards by joining short sequences to its neighbouring sequences until each phase of play lasts at least one second.

B Detecting Changes in Formation

As tactical changes in the team formation may occur at any point in the game, we need to identify the moment when this may have happened. We use the following steps to approximate the moment when a change may have occurred. Our approach is player specific; for example, if two wingers switch sides at half time, we want to identify this as a change of formation. For simplicity we use the out of possession formations as a reference, because they tend to be a bit more stable than while in possession. Therefore, we consider only the positional data of a team (excluding the goalkeeper), while the ball is in play and the opposing team is in ball possession.

We define the current formation position of a player as his average centered position, i.e. his mean average x and y coordinates relative to the team's center (see also (Andrienko et al. 2017)), between the start of this formation (e.g. the beginning of the match, or the latest identified formation change) and the current time, t. His current formation position is then compared to his position during the last three minutes of eligible frames up to time t. If the Euclidean distance between any player's current formation position and his three-minute rolling window position is greater than ten meters, we identify time t-minus-three minutes as the moment of a formation change and start to compute the current team formations starting at this time. Both thresholds were set by manually evaluating them on video footage with experts. Minor changes to these thresholds, do not strongly affect the presented results. Substituted players are compared to the position of the players they replaced. Using this algorithm over the past seven seasons of Bundesliga matches we identify on average 1.7 formation changes per match, which underpins the importance of this additional step to aggregating suitable sequences in our clustering step.

[14] These 15 phases of play contain further splits for the transition phases (e.g. counterattacking, counterpressing) and set-pieces.

C Clustering for Build-up

Fig. 10 displays the clustering outcome of the second relevant phase—the build-up phase. As discussed in Section 5.1, a major decision that has to be made by a team is whether to build up with two central defenders (formations #1, #2, #3, #4) or with three central defenders (formations #5 and #6).[15] In Fig. 10, formation #1 displays a

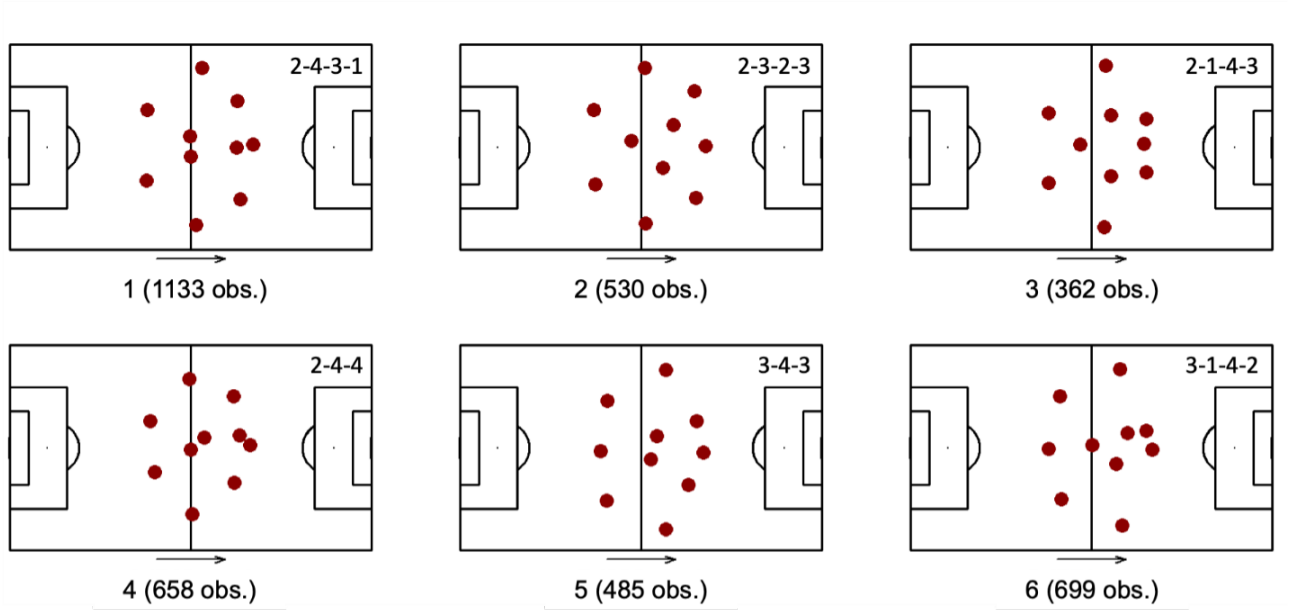

Fig. 10: Outcome of the clustering for build-up including the number of observations (obs.) of our sample.

2-4-3-1 with two central defenders playing on the same line and the full-backs pushed into midfield. In formation #4, one central midfielder clearly plays a more offensive role which allows the strikers not to participate in the build-up and rather plays a more offensive part, which was declared as a 2-4-4 by our experts. The formations shown in #2 (2-3-2-3) and #3 (2-1-4-3) also display similar patterns. The major difference is that the left and right striker tend to support the wing-back moving forward in #2, whereas in formation #3 all three strikers focus on playing in the center and leave the wings completely to the wing-backs. Formations #5 (3-4-3) and #6 (3-1-4-2) shows what our experts expected: building up with three central defenders provides a distinct flexibility during the build-up phase. A typical phenomenon when building up with three defenders is that the wing-backs have to conquer the wing-territories on their own, which should lead to a superiority in the center in both cases.

D Implementation Details

While the newly available positional data allows for novel insights, the sheer size poses a significant computational challenge for non-IT-focused organisations such as football clubs or federations. All implementations were made in Python. We implemented the CNN (Section A) using Keras and Tensorflow and trained it on a local GPU-Cluster. Additionally, we used sklearn to perform the training test data split. In order to enable rapid feedback loops with match analysts, the tracking data is locally stored in Parquet files , compressing them from 500mb to 20mb per match. This step not only saves storage in the analytics environment but also enables us to read in an entire match in less than a second. For the computations necessary in this paper, the code is parallelized whenever possible to speed up the analysis even further.

[15] Note that for the formation versus formation contemplation in Section 5.1, the hierarchical clustering is further aggregated to n=2, so that only three-defender versus two defender build-up is compared.