Automated Detection of Complex Tactical Patterns in Football

Using Machine Learning Techniques to Identify Tactical Behavior

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. Pascal Bauer
Sankt Ingbert

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 the project is the thesis of Gabriel Anzer, 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 Gabriel Anzer, Joshua Wyatt Smith (PhD) and Prof. Dr. Ulf Brefeld, for informative collaborations and discussions building a central component of the work presented. I also 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 spend much time evaluating (intermediate) results. I would also like to cordially thank Dr. Stephan Nopp, 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 for the womens national team) as well as Sebastian Geißler (former match-analyst of Borussia Mönchengladbach).

Contents

List of Publications

The following publications/submissions are in the core of the cumulative dissertation:

(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) 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

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

(V) Bauer, P., Anzer, G., Smith, J. W. (2022). Individual role classification for players defending corners in football (soccer)—Categorisation of the defensive role for each player in a corner kick using positional data. *Journal of Quantitative Analysis in Sports (submitted)*.

(VI) 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

Additional, publications/submissions conducted within the scope of this research program:

(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) 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(6). https://doi.org/10.1177/1747954119879350

(iv) Herold, M., Kempe, M., Bauer, P., & Meyer, T. (2021). Attacking Key Performance Indicators in Soccer: Current Practice and Perceptions from the Elite to Youth Academy Level. *Journal of Sports Science and Medicine*, 20(1), 158–169. https://doi.org/10.52082/jssm.2021.158

(v) 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

(vi) Ric, A., Bradley, P., Shaw, L., Thies, H., Sumpter, D., López-felip, M. A.,Ade J. Dixon D., James A., Evans M., Gómez-díaz A., Harrison H., Laws A., Petersen M., Seirul P., Robertson S., Pollard, R., Bransen, L., Kempe M., & Bauer, P. (2021). Football Analytics 2021: The Role of Context in Transferring Analytics to the Pitch. *Barça Innovation Summit 2020, Barcelona*.

Abstract

Football tactics is a topic of public interest, where decisions are predominantly made based on gut instincts from domain-experts. Sport science literature often highlights the need for evidence-based research on football tactics, however the limited capabilities in modeling the dynamics of football has prevented researchers from gaining usable insights. Recent technological advances have made high quality football data more available and affordable. Particularly, *positional data* providing player and ball coordinates at every instance of a match can be combined with *event data* containing spatio-temporal information on any event taking place on the pitch (e.g. passes, shots, fouls). On the other hand, the application of machine learning methods to domain-specific problems yields a paradigm shift in many industries including sports.

The need for more informed decisions as well as automating time consuming processes—accelerated by the availability of data—has motivated many scientific investigations in football analytics. This thesis is part of a research program combining methodologies from sports and data science to address the following problems: the synchronization of positional and event data, objectively quantifying offensive actions, as well as *the detection of tactical patterns*. Although various basic insights from the overall research program are integrated, this thesis focuses primarily on the latter one.

Specifically, positional and event data are used to apply machine learning techniques to identify eight established tactical patterns in football: namely *high-/mid-/low-block* defending, *build-up/attacking play* in the offense, *counterpressing* and *counterattacks* during transitions, and patterns when defending corner-kicks, e.g. *player-/zonal-* or *post-marking*. For each pattern, we consolidate definitions with football experts and label large amounts of data manually using video recordings. The inter-labeler reliability is used to ensure that each pattern is well-defined. Unsupervised techniques are used for the purpose of exploration, and supervised machine learning methods based on expert-labeled data for the final detection. As an outlook, semi-supervised methods were used to reduce the labeling effort. This thesis proves that the detection of tactical patterns can optimize everyday processes in professional clubs, and leverage the domain of tactical analysis in sport science by gaining unseen insights. Additionally, we add value to the machine learning domain by evaluating recent methods in supervised and semi-supervised machine learning on challenging, real-world problems.

1 Introduction

In 1987, Bate declared (association) football as a *'game of opinions'* in which *'coaches and managers base strategy and tactics on their own opinions'*. Hence, he motivated that team strategy should be based on something more substantial rather than opinions and instincts. More than 30 years later, Kuper (2018) maintains a disproportionately high number of decisions are made based on gut instincts in professional football. Aiming to build more evidence around football tactics, Reep and Benjamin (1968) were the first to systematically annotate data from professional football matches. Since then, various studies on tactical performance analysis used manually acquired statistics to conduct studies on a substantial basis (Camerino, Chaverri, Anguera, & Jonsson, 2012; Borrie, Jonsson, & Magnusson, 2002; Gould & Gatrell, 1979). The major research question for such investigations in football was to study the efficiency of tactics and strategy (Sarmento et al., 2018, 2014) more objectively in order to support decision-making (Desporto, 2009). Following the idea of Reep and Benjamin (1968), various sport science researchers utilized coding systems to manually annotate match logs. By doing so, they captured a rough extraction of a football match to answer pre-defined research questions (Camerino et al., 2012; Alcock, 2010; Borrie et al., 2002).

The growing public interest in football and recent advances in technology has enabled exhaustive data collection across all professional football leagues (Beal, Norman, & Ramchurn, 2019; Seidl, 2019). Companies[1] commercialized the systematic collection of so called *event data* across several professional competitions (Lucey, Oliver, Carr, Roth, & Matthews, 2013). Based

[1] such as Sportec Solutions AG, Statsperform, Statsbomb or Wyscout

on a dedicated definition catalog, defining basic events (e.g., passes, shots, duels, ...) in detail, manual operators annotate each event with the support of software-tools (Pappalardo et al., 2019). Even though event data comprise the central actions of a football match, they only contain information regarding the few players directly involved in a ball action (Borrie et al., 2002). This problem was solved through the introduction of global positioning systems (Hennessy & Jeffreys, 2018) and the latest improvements in image processing and computer vision (Manafifard, Ebadi, & Abrishami Moghaddam, 2017; D'Orazio & Leo, 2010; Barris & Button, 2008). So called *positional data* (often also referred to as *movement* or *tracking data*), containing the position of every player and the ball across the whole match, became available and affordable in sports (Manafifard et al., 2017; Stein et al., 2017).

Not limited to football, the combination of positional and event data (often also referred to as play-by-play data) enabled a change in paradigm for sport science (Sarmento et al., 2018; Link, 2018; Rein & Memmert, 2016) as well as in the everyday business of clubs and federations (Herold, Kempe, Bauer, & Meyer, 2021; Andrienko et al., 2017; Herold et al., 2019). Besides the well-known Moneyball-story (MacLennan, 2005)—gaining notoriety whereby a statistician provided evidence for transfer decisions to Oakland Athletics's general manager Billy Beane—a whole community of statistical researchers emerged in baseball (so called Sabermetrics) (Albert, 2010; Baumer & Zimbalist, 2014) and later in basketball analytics (so called APBRmetrics) (Stephanos, Husari, Bennett, & Stephanos, 2021; Schumaker, Solieman, & Chen, 2010). More recently, researchers tried to transfer the learnings to other invasion sports such as American football (Atmosukarto, Ghanem, Ahuja, Ahuja, & Muthuswamy, 2013), Australian rules football

(Sampaio & Maças, 2012), team handball (Pfeiffer & Perl, 2015), rugby (Bunker, Fujii, Hanada, & Takeuchi, 2020) and (association) football (F. R. Goes, Meerhoff, et al., 2020; Lucey, Oliver, et al., 2013), establishing the research domain of *sports analytics* (Araújo, Couceiro, Seifert, Sarmento, & Davids, 2021; Beal et al., 2019; Link, 2018; Morgulev, Azar, & Lidor, 2018; Schumaker et al., 2010).

This dissertation is not only motivated from a football perspective. Recent success in machine learning, a research domain established since the 1950's (Samuel, 1959) warrants application domains—notably practical tasks a machine learning algorithm can learn from data—to explore and evaluate new methods. In this context, the dynamic nature of football, often referred to as the most complex of invasion sports (Tuyls et al., 2021) provides various problems challenging machine learning researchers. In a position paper, Tuyls et al. (2021) outlines how applications of artificial intelligence in football—due to its modeling complexity, dynamics and the ubiquitous human component—can serve as a valuable playing ground for machine learning research.

We focus on tactical (rather than physical) performance analysis in football, and on investigations to improve decision making when applying tactics (rather than result-predictions). Relevant studies include probabilistic models to quantify goal scoring probabilities using expected goals (xG) values (Anzer & Bauer, 2021; Robberechts & Davis, 2020; Lucey, Bialkowski, Monfort, Carr, & Matthews, 2014), pass completion probabilities using expected pass (xPass) values (Anzer & Bauer, 2022; Spearman, Basye, Dick, Hotovy, & Pop, 2017), or the goal scoring probability at any time-point in the match using expected possession values (Fernández, Bornn, & Cervone, 2021; Spearman, 2018). The above described metrics allow more contextual evaluation of the key-events in football (e.g., passes, shots), facilitating to overcome otherwise limited, binary evaluations (e.g., passes: completed or not; shots: successful or not). Expected pass approaches are often combined with reward quantification of passes (Steiner, Rauh, Rumo, Sonderegger, & Seiler, 2019; Power, Ruiz, Wei, & Lucey, 2017; Rein, Raabe, & Memmert, 2017; Chawla, Estephan, Gudmundsson, & Horton, 2017; F. Goes, Schwarz, Elferink-Gemser, Lemmink, & Brink, 2021; Gómez-Jordana, Milho, Ric, Silva, & Passos, 2019; F. R. Goes, Kempe, Meerhoff, & Lemmink, 2019) allowing to quantify pass decisions considering risk and reward of pass options compared to alternatives. Another continuously addressed concept is the control of space and the development of movement models (Martens, Dick, & Brefeld, 2021; Brefeld, Lasek, & Mair, 2019; Fernandez & Bornn, 2018; Fujimura & Sugihara, 2005; Taki, Hasegawa, & Fukumura, 1996). Teams' playing styles (Beal, Chalkiadakis, Norman, & Ramchurn, 2020; Kempe, Vogelbein, Memmert, & Nopp, 2014; Vogelbein, Nopp, & Hökelmann, 2014) or formations (Gudmundsson, Laube, & Wolle, 2017) were studied across the course of a match or a whole season.

Various studies from the sports science domain (F. R. Goes, Meerhoff, et al., 2020) used data to validate hypotheses via deductive reasoning (Pappalardo et al., 2019; Sarmento et al., 2018; Duarte et al., 2013; Bartlett, Button, Robins, Dutt-Mazumder, & Kennedy, 2012). On the other hand, more data-driven approaches aim to investigate interesting patterns directly from the high-dimensional, spatio-temporal data. Accordingly, two major approaches can be distinguished: the *supervised* detection of predefined patterns (Müller-Budack, Theiner, Rein, & Ewerth, 2019; Chawla et al., 2017), as well as the *unsupervised* exploration of patterns (Decroos, Van Haaren, & Davis, 2018; Knauf, 2014).

The detection of pre-defined tactical patterns conducted by teams in specific situations is a central issue in basketball and American football analytics, nevertheless, it is not adequately addressed in football. In the following we define a *tactical pattern* as a well-specified, repeatable and coordinated movement of a team (or a group of members) conducted in a specific situation of a match. One relevant aspect of tactical patterns in our context is that they are uniquely identifiable by experts (see Section 2.3 for details). Another established concept in invasion sports is *game-states*, splitting each match into offense and defense on the highest level (Gréhaigne, Bouthier, & David, 1997). In football, transitions to offense or defense as well as set-pieces are often considered as separate game-states (Wei, Sha, Lucey, Morgan, & Sridharan, 2013). An overview of the game-states in football is shown in Figure 1. A team falls into the transition to offense phase after winning the ball (and vice versa), whereas set-pieces are separated since they start with a stoppage in the match.

In 2002, Borrie et al. stated that behavior in team sports contains more patterns than the the human eye can observe. For a single match, Laird and Waters (2008) showed that coaches have a limited recall in reconstructing relevant scenes from the match. Hence, the objective analysis of tactical patterns is of high interest for practitioners and for sport science research (Gudmundsson et al., 2017; Stein et al., 2017). Annotating tactical patterns by inspecting video-footage is time-consuming (Perse, Kristan, Perš, & Kovacic, 2006; Gudmundsson et al., 2017) and often subjective (Perše, Kristan, Kovačič, Vučkovič, & Perš, 2009). As a consequence, sample sizes that can be investigated (e.g., for trend analysis over multiple seasons) are limited due to time-constraints. Consequently, the main objective of this thesis is to automate this process using positional and event data, i.e. to answer the following research question: *(How) can tactical patterns be detected automatically using machine learning algorithms based on positional and event data?*

To answer this question, the remainder of this thesis is structured as follows: Section 2 describes the basic preliminaries for all investigations, starting with a detailed description of positional and event data (Section 2.1). The basics in machine learning methodologies for sports applications are provided in Section 2.2. Finally, Section 2.3 describes related work on the detection of tactical patterns in order to derive the above presented definition of a tactical pattern and to motivate the experimental studies presented in Section 3. In studies I & II (Sections 3.1 and 3.2) interesting tactical patterns are explored using unsupervised techniques, while studies III–VI aim to detect tactical patterns along the five established game-states shown in Figure 1 (offensive, defensive, transition to offense, transition to defense and set-pieces). Study III (Section 3.3) focuses on transitions to defense, i.e. *counterpressing*. The detection of team-tactical patterns is extended to offensive (i.e., *build-up* and *attacking play*) and defensive patterns (i.e., *low-/mid-* and *high-block*) in study IV (Section 3.4). Tactical patterns during set-pieces (i.e., *player-* or *zonal-marking* during corner kicks) are detected in study V (Section 3.5). Study VI (Section 3.6) presents an outlook of offensive transitions, i.e. *counterattacks*. All results are discussed in Section 4.

Besides the core contributions presented in this thesis (studies I–VI in Section 3), further publications have been achieved within the scope of this research program: In Herold et al. (2021) and Herold et al. (2019) the relevance of machine learning applications for football were explored from a practical perspective. In Anzer and Bauer (2021) and Anzer and Bauer (2022)

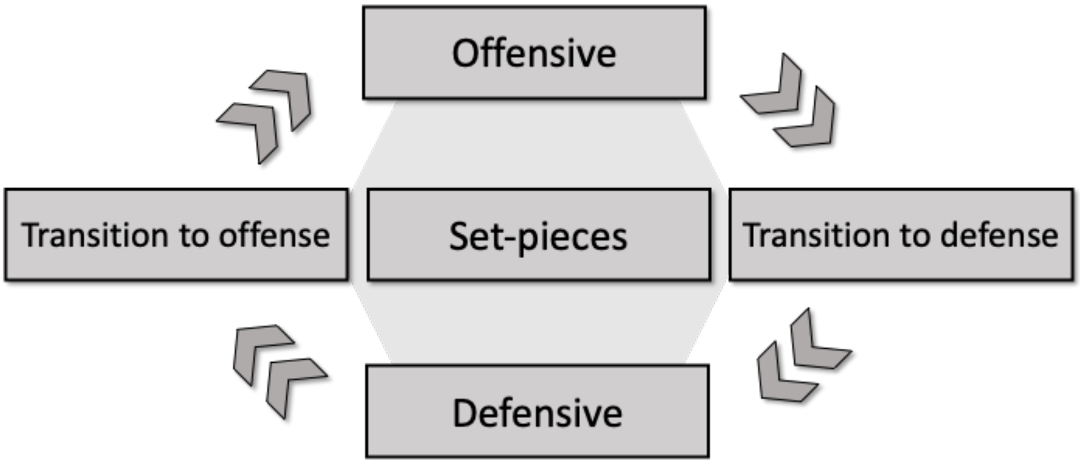

Figure 1: Overview of game-states in football.

expected goals (xG) and expected pass (xPass) models have been built which are used as a fundamental component in all empirical studies of this thesis. Additionally, two book contributions provide a more general overview on machine learning applications in football (Anzer, Bauer, & Höner, 2021), as well as on defensive organisation (Ric et al., 2021).

2 State of the Art

2.1 Positional and Event Data in Football

2.1.1 Event Data

Event data are logs of well-defined basic events describing a football match, such as passes, shots, fouls, substitutions, corners etc. Besides the time-stamp, the involved players (e.g., passer, pass receiver), as well as several sub-attributes for each event (e.g., high/low pass, completed/not completed), an estimated position on the where on the pitch the event happened is usually contained (Pappalardo et al., 2019; Stein et al., 2017). For a long time, event data were collected explicitly to answer pre-defined research questions (F. R. Goes, Meerhoff, et al., 2020). For example, several approaches focused on annotating passes (Reep & Benjamin, 1968; Camerino et al., 2012; Alcock, 2010; Borrie et al., 2002), others annotated corners manually (Pulling, 2015; Casal, Maneiro, Ardá, Losada, & Rial, 2015; Schmicker, 2013). Following more of a bottom-up approach, event data are nowadays collected systematically across matches and seasons (Pappalardo et al., 2019) describing a match as an ordered sequence of events (Stein et al., 2017). Companies like Stats Perform[2], Sportec Solutions[3], and Statsbomb[4] developed their own event data catalog respectively, defining events and several attributes in great detail.[5] Basic event data catalogs are also described in Bialkowski et al. (2016); Pappalardo et al. (2019) or Stein et al. (2017). Stein et al. (2017) pointed out that the accordance of definitions differ per event. For example, events like throw-ins, kick-offs or corner kicks are well-defined, whereas different interpretations of tackles/duels, crosses or even successful passes cause huge differences qua definition and/or subjective annotations.

A major flaw with primarily manually annotated event data is inaccuracies due to human errors (Stein et al., 2017), especially for the position on the pitch (typically drawn on a digital coordinate system) and the timestamp (usually delayed due to the human reaction time). Making use of automatically collected positional data to correct timestamp and location of event data is, consequently, a relevant issue. The issue of synchronizing positional and event data is only rarely addressed as a limitation (Spearman et al., 2017), current approaches typically use either

[2]Statsperform LLC, Chicago, https://www.statsperform.com/.; former AMISCO, Prozone and Opta.

[3]Sportec Solutions AG, Munich, Germany https://www.sportec-solutions.de/index.html.

[4]Statsbomb Services Limited https://statsbomb.com/.

[5]Note that event data are also collected on a systematic basis in other sports and often referred to as play-by-play data in other sports, e.g. basketball (Vračar, Štrumbelj, & Kononenko, 2016).

positional or event data, or rely on exhaustive manual annotations aligning event data with other data sources. Within the scope of this research program, the issue of synchronizing event data and positional data is addressed in Anzer and Bauer (2021) for shots and in Anzer and Bauer (2022) for passes or more generally in Anzer (2021). All event data used in this dissertation are spatio-temporally synchronized with the positional data, allowing us to effectively use complementary information from both data-sources.

2.1.2 Positional Data

Event data focuses on events with the ball, whereas Link and Hoernig (2017) pointed out that a player possesses the ball on average only for less than three minutes per match (depending on the position).[6] Accordingly, literature has expressed a haunting need to gather more information on off-the-ball activities (F. R. Goes, Meerhoff, et al., 2020; Vilar, Araújo, Davids, & Travassos, 2012; Borrie et al., 2002). *Positional data*, often also referred to as *tracking* or *movement data* capture the exact positions of players at any time within a match (typically with a frequency of 10 or 25 Hz). Positions are transformed into a two-dimensional Cartesian coordinate system determined by the pitch surroundings (Stein et al., 2017; Andrienko et al., 2017). For the ball, positional data typically contain a third dimension—the height relative to the pitch surface.

For the acquisition of positional data, one can differentiate between sensor-based solutions (i.e., global or local positioning systems) and optical tracking systems. Optical tracking systems ap-

[6] Average possession times according to Link and Hoernig (2017): 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)

ply computer vision algorithms to video footage collected from dedicated tracking cameras covering multiple perspectives of a football match (Taberner et al., 2020; Stein et al., 2017). In the past, GPS tracking data were predominantly used in sport science literature, especially for physical performance analysis (see Low et al. (2020) for an overview). For tactical analysis, various shortcomings limited their usage: GPS data provide longitude and latitude coordinates, which describes an object's position relative to the earth surface (McNeff, 2002). The transformation to the pitch-centered coordinate system as well as the stadium infrastructure (disturbing the connection between GPS receiver) cause significant inaccuracies on player positions (Sathyamorthy, Shafii, Amin, Jusoh, & Ali, 2015; Pons et al., 2019). Additionally, various practical limitations makes it hard to consistently use sensor-based data (e.g., missing data of opponent and ball, players conceiving the devices as disturbance, single devices can break during a match) (F. R. Goes, Meerhoff, et al., 2020; Low et al., 2020; Buchheit et al., 2014; Hennessy & Jeffreys, 2018). With recent developments in computer vision and video processing, optical tracking systems turned out to be an alternative with sufficient and increasing accuracy (Linke, Link, & Lames, 2020; Taberner et al., 2020; Linke, Link, & Lames, 2018). A wide range of research has been conducted on optical tracking systems (see Manafifard et al. (2017) for an overview). Even though, the evaluation of positional data accuracy is an ill-posed problem due to missing ground-truth information, experimental evaluation studies are conducted in great detail (Linke et al., 2020; Taberner et al., 2020; Linke et al., 2018; Cardinale, 2006), stating that cutting-edge systems track player positions with an error of less then 10 cm (Linke et al., 2020).

2.1.3 Dataset of Bundesliga and German National Team

For the remainder of this thesis, we make use of event data collected by Sportec Solutions following the official event data catalog for German Bundesliga.[7] Positional data are generated via different generations of the Chyronhego TRACAB system,[8] which has been validated in Linke et al. (2020). The dataset, consisting of several seasons of Bundesliga data as well as national team matches is described in more detail in the respective publications, as well as in Anzer (2021). A detailed description of the annotation of shots can be found in Anzer and Bauer (2021), more details on passing events are pointed out in Anzer and Bauer (2022).

The data used in this study are property of the Deutsche Fußball-Bundesliga,[9] as well as the Deutsche Fußball-Bund[10] and cannot be shared publicly. However, open-source positional data (Pettersen et al., 2014)[11] and event data (Pappalardo et al., 2019)[12] can be used for reproduction. These open source data-sets provide the scientific community the option to reconstruct and evaluate the approaches presented in this thesis.

2.2 Machine Learning Basics for Sports Applications

Goodfellow, Bengio, and Courville (2016) defines machine learning as the ability of algorithms to learn from data. Although machine learning methods have been researched for decades (Samuel,

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

[8] https://tracab.com/products/tracab-technologies/ or https://chyronhego.com/wp-content/uploads/2019/01/TRACAB-PI-sheet.pdf

[9] https://www.dfl.de/de/

[10] https://www.dfb.de/index/

[11] 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).

[12] 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).

1959), its recent success was only enabled through the availability of vast amounts of data and the affordability of computing power needed to process such volumes of data (Jordan & Mitchell, 2015). A major methodological distinction in machine learning is made between supervised and unsupervised learning (Alloghani, Al-Jumeily, Mustafina, Hussain, & Aljaaf, 2020). While *unsupervised machine learning* explores patterns in data with little or no human guidance (Gentleman & Carey, 2008), *supervised machine learning* uses human annotations to imitate experts performing tasks (Sing, Thakur, & Sharma, 2016). The biggest category of unsupervised machine learning approaches solve clustering problems, while supervised machine learning algorithms are typically applied to classification or regression tasks (Goodfellow et al., 2016). Unsupervised algorithms are thus purely data-driven, unbiased by expert assumptions, well-suited to explore unknown data, and have the ability to disclose unknown patterns in data (Goodfellow et al., 2016; Alloghani et al., 2020). On the contrary, supervised algorithms require considerable human effort (i.e., for data labeling) and a clearly defined task. However, once a supervised machine learning algorithm performs a given task with sufficient accuracy, they can be used in production to assume repetitive tasks that human experts would perform otherwise. To reduce the human effort required for labeling in supervised machine learning, *semi-supervised approaches* emerged as a third category (Goldberg, 2009). The basic idea is that human expertise is still used for labeling, but supported by more automated approaches to reduce the amount of labels required. For example, weak supervision (applied in study V, Section 3.5) uses rule-based approaches to create weakly labeled data, which are used in combination with expert-labeled data to train supervised machine learning algorithms (Ratner et al., 2017). Another method to increase the amount of labeled data without human effort is data augmentation (Dyk & Meng, 2001), used also in studies V & VI (Sections 3.5 & 3.6). Here, labeled data are slightly modified to create artificial data.

On the highest level, supervised machine learning algorithms aim to learn a function f taking input data $x \in X$ to predict pre-defined labels $y \in Y$: $f(x) \approx y$. Hereby, x can be seen as a reproduction of the real-world phenomena (e.g., a sequence of a football match), typically transformed into a lower dimensional space in order to reduce the complexity (e.g., through expert-driven feature extraction from raw data). y describes the label (target variable) in a pre-defined label space Y. For regression tasks y is a continuous variable, for classification task y is discrete. Compared to traditional statistics, inductive reasoning (i.e., generalisation) is performed via using three different sub-sets of data: training data X_{train}, Y_{train}, test data X_{test}, Y_{test}, and new data X_{new}. The training and testing labels (Y_{train} and Y_{test}) are available through expert-annotation, whereas the target variables for X_{new} are unknown. Mathematically, f is trained as an optimization task learning to chose the best free parameters $p \in P$ in f_p to minimize the prediction error on the training data: $\min_p \sum_{x_i, y_i \in X_{train}, Y_{train}} |f_p(x_i) - y_i|$. With sufficient flexibility in f (i.e., a high dimensional parameter space P), this optimization task can perform well by overfitting f on the training data. Thus, generalisation in supervised machine learning is evaluated on test data X_{test}, Y_{test}, that are not used in the training process. The error on the test data is described by $\sum_{x_i, y_i \in X_{test}, Y_{test}} |f_p(x_i) - y_i|$ (p optimized via model training) and typically used to evaluate the accuracy of the model.

For f, a plethora of algorithm families is established (Mohri, Rostamizadeh, & Talwalkar, 2014; Goodfellow et al., 2016). Relying typically on a feature extraction, *extreme gradient boosting (XGBoost)* classifiers, from the *family of tree-based algorithms*, achieve good results in various domains (T. Chen & Guestrin, 2016). XGBoost is used in Anzer and Bauer (2022, 2021), as well as in study III of this thesis (Bauer & Anzer, 2021). More details on the methodology can also be found in Anzer (2021). A closely related methodology, random forests, has also been applied in sports analytics (Karsten et al., 2017).

Besides tree-based classifiers, *support vector machines* are predominantly used to perform (binary) classification tasks (Pisner & Schnyer, 2019), and have also been applied to sports analytics (B. W. Hosp, Schultz, Honer, & Kasneci, 2021; Atmosukarto et al., 2013). Support vector machines are used to detect counterattacks in study VI of this dissertation.

Artificial neural networks (Shinde & Shah, 2018)—another type of machine learning algorithm that can be used for various types of tasks—achieved groundbreaking results in image (and video) classification tasks (Krizhevsky, Sutskever, & Hinton, 2012), natural language processing (Otter, Medina, & Kalita, 2021), autonomous driving (Shinde & Shah, 2018), playing video (Mnih et al., 2013) or board games (Ghory, 2004), and many more (e.g., Najafabadi et al. (2015)). Although deep architectures of neural networks (typically referred to as deep learning) require vast amounts of (labeled) data and sizeable computing power, their largest benefit is that they can often be fed with raw data, i.e. without the need for exhaustive feature engineering. In this dissertation, *convolutional neural networks*, a special family of neural networks optimized to handle images (see Zhang et al. (2019) for an overview), are used in studies IV & V (Sections 3.4 & 3.5). In study V we also apply the concept of *long short-term memory networks* introduced by Hochreiter and Schmidhuber (1996) to efficiently handle temporal data. A combination of convolutional layers with long-short term memory units has been used in football to perform classification tasks on eye-tracking data (B. Hosp, Schultz, Kasneci, & Höner, 2021). Grunz, Memmert, and Perl (2009, 2012) applied self organising maps, a modification of neural networks that are able to approach unsupervised problems, to different sports.

Further machine learning algorithms have been successfully applied to detect tactical patterns in invasion sports. Logistic regressions (Chawla et al., 2017; Mcintyre, Brooks, Guttag, Wiens, & Arbor, 2016), Bayesian classifiers, multi-model densities (Li, Chellappa, & Zhou, 2009), pattern templates (Intille & Bobick, 1999; Siddiquie, Yacoob, & Davis, 2009; Li & Chellappa, 2010; Perše et al., 2009, 2009), bag-of-word algorithms (Montoliu, Martín-Félez, Torres-Sospedra, & Martínez-Usó, 2015) and k-nearest neighbours (Bialkowski et al., 2015) were used to solve supervised problems.

T-pattern analysis (Borrie et al., 2002; Hernández-mendo, 2006), multi-scale matching (Hirano & Tsumoto, 2004), multiple modifications of hierarchical clustering (Bialkowski et al., 2014), convolutional kernels (Knauf, 2014) and the Delauany method (Narizuka & Yamazaki, 2019) addressed unsupervised tasks.

Hierarchical clustering, the most established unsupervised technique, is used in studies II & IV.

A general introduction on supervised and unsupervised machine learning applications in football is presented in Anzer, Bauer, and Höner (2021). Fujii (2021) and Araújo et al. (2021) outline an overview of machine learning methodologies applied to sports. In Herold et al. (2019) we give an overview of machine learning applications applied to quantify offensive play in football. Technical details on the above presented algorithms can be found in Goodfellow et al. (2016) and (Han, Kamber, & Pei, 2012).

In order to train a machine learning algorithm (either supervised or unsupervised) the underlying data has to be modeled into an appropriate structure, which poses technically demanding challenges. Positional data are a high-dimensional, spatio-temporal abstraction of a football match (Stein et al., 2017). Typically, the dimensionality of the input data X has to be reduced (Lucey, Bialkowski, et al., 2013) to simplify the problem. A traditional approach is to use subject-matter expertise to extract features and perform the prediction task from a lower-dimensional feature space (Duarte et al., 2013). Second, the movement of players for the same tactical pattern can vary dramatically for different scenarios (Perse et al., 2006; Li & Chellappa, 2010; Stracuzzi et al., 2011), which makes the definition of similarity metrics challenging. A third major problem when dealing with positional data is to project player positions in a permutation-invariant space (Wei et al., 2013). The ordering problem in invasion sports is addressed by describing positional data as images (Dick & Brefeld, 2019; S. Zheng, Yue, & Lucey, 2016; K.-C. Wang & Zemel, 2016), artificial orderings using heuristics (Le, Yue, Carr, & Lucey, 2017; Fujii, 2021) or, again, extracting features from the raw data. More recently, different approaches show that graph neural networks can serve as an efficient procedure to solve the permutation-invariance problem (Dick, Tavakol, & Brefeld, 2021; Sun, Karlsson, Wu, Tenenbaum, & Murphy, 2019; Yeh, Schwing, Huang, & Murphy, 2019).

2.3 Data-Driven Detection of Tactical Patterns in Sport

2.3.1 Tactical Patterns in Invasion Sports

Gréhaigne, Godbout, and Bouthier (1999) separated two often confounded terms: *strategy*—the plan of a team made before a match—and *tactics*—decisions that are conducted during a match as a response to the dynamic environment. In this context, they presented a sub-category of tactics, called schemas of play, describing organized, collective and repeated patterns. However, Rein and Memmert (2016) claimed that a clear distinction between tactics and strategy is challenging, since any real-time interaction will be prone by the a priori strategy. Even though the concept was originally designed for football, tactics and strategy is often used synonymously for invasion sports in general. A team's collective movement in invasion sports is clearly coordinated and contains patterns and aspects of synchronicity (see Sarmento et al. (2018) or Courel-Ibáñez, McRobert, Toro, and Vélez (2017) for reviews). Rather than following one fixed team or game tactic during a whole match, the current *game-state* (i.e., offensive, defensive, transition, set-piece) and further sub-categories (e.g., counterattack or counterpressing for transitions in football) significantly influence team behavior in invasion sports (Alexander, Spencer, Sweeting, Mara, & Robertson, 2019; Wei et al., 2013). Thus, team behavioral patterns have to be investigated in finer granularity, especially along game-states and the respective sub-categories.

For tactical patterns, literature lacks an established definition, rather a plethora of namings have been used for one and the same concept: schemas of play (Gréhaigne et al., 1999), cooperative plays (Hojo, Fujii, Inaba, Motoyasu, & Kawahara, 2018), game-phases (Perse et al., 2006; Lucey et al., 2014), events (Pfeiffer

& Perl, 2015), group motion patterns (Li & Chellappa, 2010), movement patterns (Stein et al., 2017; Gudmundsson & Horton, 2017), or tactical patterns (Q. Wang, Zhu, Hu, Shen, & Yao, 2015; Kempe, Grunz, & Memmert, 2015; Knauf, 2014; Grunz et al., 2012, 2009). While most authors used the respective term, without referring to an explicit definition, Q. Wang et al. (2015) defined tactical patterns as a series of frequently used ball-passing combinations, stating that this definition fits their purpose of analyzing passing patterns best. In football, Lucey et al. (2014) used the term game-phase for the same phenomena (less focused on passing behavior)—patterns conducted by the whole team in specific situations of a match—and analyzed seven exemplary categories (corners, free-kicks, penalties, set-pieces, open play, counterattacks). To be a good candidate for such a pattern that can be detected using a supervised machine learning approach, sundry authors highlight that the patterns must be well-defined and widely used (Hojo et al., 2018; Bialkowski et al., 2015; Montoliu et al., 2015; K.-C. Wang & Zemel, 2016).

In the following, we define a *tactical pattern* as a repeatable, coordinated movement of a team (or a group of members) conducted in specific situations, which can be uniquely identified by experts (Kempe et al., 2015; Q. Wang et al., 2015; Grunz et al., 2012).

2.3.2 The Detection of Tactical Patterns in Invasion Sports

The relevance of automatically detecting such tactical patterns in team-sports is highlighted in Desporto (2009) or Gudmundsson et al. (2017). All known studies following the goal of automatically detecting tactical patterns in invasion sports using positional data, event data and/or video footage are listed in Table

1.[13] In the following we pose the related work on tactical pattern detection in invasion sports along Table 1.

In 1999, Intille and Bobick introduced a pioneering approach to detect a pre-defined offensive play in American football using (manually drawn) movement trajectories of 29 attacking plays. The task of recognizing attacking plays remains a predominant problem in American football analytics (Stracuzzi et al., 2011; Li & Chellappa, 2010; Siddiquie et al., 2009; Li et al., 2009). High-level game-phases (offense, defense, kickoff, punt, field-goal plays) were detected directly from video-footage by S. Chen et al. (2014). Atmosukarto et al. (2013). Hochstedler and Gagnon (2017) detected the static formations (i.e., player roles) at the start of each play.

Many studies have also been conducted in basketball (see Stephanos et al. (2021) or Courel-Ibáñez et al. (2017) for reviews), focusing on the primary objective of automatically detecting lower level tactics performed by groups of players, i.e. screening[14] (Hojo et al., 2018; Mcqueen, Wiens, & Guttag, 2014; Perše et al., 2009; Perse et al., 2006), defensive counter-strategies on screening (Tian, De Silva, Caine, & Swanson, 2020; Mcintyre et al., 2016) or cutting[15] (Perse et al., 2006). Those group-tactical behaviors were often gathered as a sequence of actions to analyse whole attacking plays (Kempe et al., 2015; J. Wang & Zhang, 2015; H. T. Chen,

[13] Studies analyzing rather technical than tactical patterns (Schmidt, 2012) or studies analyzing physical patterns for the purpose of injury prediction (Kelly, Coughlan, Green, & Caulfield, 2012; Cai et al., 2018) were excluded, since those do not suit our definition of tactical patterns. We also exclude studies focusing purely on basic event detection from video footage (Pouyanfar & Chen, 2016; Kolekar, Palaniappan, Sengupta, & Seetharaman, 2009; Ekin, Tekalp, & Mehrotra, 2003) or from positional data (Stein et al., 2019; Richly, Bothe, Rohloff, & Schwarz, 2016; Motoi et al., 2012; Gudmundsson & Wolle, 2010; M. Zheng & Kudenko, 2010).

[14] A group-tactical pattern, where an offensive player (not in possession of the ball) legally blocks a defender in order to pretend him/her from defending himself/herself or his/her team-mate in possession of the ball.

[15] Cutting describes a sudden off-ball movement of a player in order to get rid of his defender.

Chou, Fu, Lee, & Lin, 2012; Perše et al., 2009; Perse et al., 2006) or to identify players or teams just by recognizing their playing styles (Mehrasa, Zhong, Tung, Bornn, & Mori, 2018).

From various approaches investigating screen plays in basketball, Hojo et al. (2018) presented an extended study improving the state-of-the-art results. First, they properly define different types of screen plays. Based on this definition, data were expert-labeled in a controlled experiment with staged screens, and later in real game situations. In both experiments positional data were recorded. For the binary classification (screen play or not) their thorough experimental set-up exceeds prior results with an area under the curve of 0.941 for on-ball and 0.855 for off-ball screens. Although different labelers were involved, no proper inter-labeler reliability study was conducted.

The data-driven detection of tactical patterns has also been investigated in ice-hockey (Mehrasa et al., 2018), Australian rules football (Rennie et al., 2020; Alexander et al., 2019), team handball (Pfeiffer & Perl, 2015) as well as in rugby (Bunker et al., 2020; Karsten et al., 2017).[16]

In Basketball (Hojo et al., 2018; K.-C. Wang & Zemel, 2016; Mcintyre et al., 2016; Mcqueen et al., 2014) and in American football (Hochstedler & Gagnon, 2017; Atmosukarto et al., 2013; Li et al., 2009; Siddiquie et al., 2009), the application of supervised machine learning methods using manual expert-annotations is established (see also Table 1). The relevance of clearly recognizable patterns is highlighted in various studies, nevertheless, the inter-labeler reliability is only analyzed as an indicator for agreement in football (Chawla et al., 2017) and rugby (Bunker et al., 2020). Unsupervised approaches often aim for the creation of

[16] Again, this dissertation focuses on pattern detection for invasion sports. Nevertheless, several studies analyzed patterns in other sports, for example volleyball in Van Haaren, Shitrit, Davis, and Fua (2016).

new, unknown insights, whereas process automation for coaches and match analysts is mentioned as the primary aim of supervised approaches.

2.3.3 The Detection of Tactical Patterns in Football

The first known approach analyzing tactical patterns stems from football: As early as in 1968, Reep and Benjamin (1968) manually annotated the passing sequences across 54 matches in the English Premier league as well as 47 matches from other competitions. Building on that pioneering work, a majority of approaches presented in football focus on the exploratory detection of attacking patterns (Decroos et al., 2018; Hobbs, Power, Sha, Ruiz, & Lucey, 2018; Van Haaren, Dzyuba, Hannosset, & Davis, 2015; Montoliu et al., 2015; Fernando, Wei, Fookes, Sridharan, & Lucey, 2015; Niu, Gao, & Tian, 2012; Borrie et al., 2002), build-up patterns (Knauf, 2014; Grunz et al., 2012) and patterns in pass sequences (Chawla et al., 2017; Brooks, Kerr, & Guttag, 2016; J. Wang & Zhang, 2015; Hernández-mendo, 2006; Hirano & Tsumoto, 2004). Defensive corner kick strategies have been investigated in Shaw and Gopaladesikan (2021) and Power, Hobbs, Ruiz, Wei, and Lucey (2018). More frequently addressed (compared to basketball and American football) are studies on team formations (Narizuka & Yamazaki, 2019; Müller-Budack et al., 2019; Shaw & Glickman, 2019; Bialkowski et al., 2014, 2015, 2016; Wei et al., 2013). Note that many studies aim to detect one team formation aggregated over a whole match, which does not fit our definition of tactical patterns (see Bauer, Anzer, and Shaw (2022) for more details). However, due to the complexity of the game, the analysis of formation patterns in football is a unique characteristic compared to American football (where only static formations at the beginning of each play are of interest) or to

Basketball (where fewer variations on formations exist).

In football, only a few studies built their work on expert-labeled data of pre-defined patterns (frequently used in basketball and American football) Montoliu et al. (2015) hand-labeled five types of attacking plays, Shaw and Gopaladesikan (2021) as well as Power et al. (2018) detected defensive tactics of defending corner kicks. In Chawla et al. (2017), two human experts annotated the reward of in total $2,932$ passes on a six-point Lickert scale (very good, good, marginally good, marginally bad, bad, very bad). The alignment among the labelers was monitored using Cohen's kappa—after labeling two matches, disagreements were consolidated. Finally, different classifiers were trained to detect three types of pass ratings (good, ok, bad)—multinominal logistic regression performed best on various metrics with an F_1-score of 0.748.

This thesis uses the idea of supervised machine learning to detect identifiable tactical patterns as well as established phases of play. By addressing team-tactical patterns in all game-states (offensive, defensive, transition and set-pieces) and by integrating football experts closely in our experiments, we guarantee a practical applicability of the results. As in Chawla et al. (2017), we make use of both positional and event data and emphasize the importance of a thorough labeling process.

Table 1: Overview of tactical pattern detection in invasion sports.

Study	Sport	Data	Methods	Pattern
Intille and Bobick (1999)	American Football	Positional data (manually drawn trajectories)	Supervised ML (bayesian classifiers, pattern template)	Attacking play (one particular group tactical pattern called *p51curl*-play)
Li et al. (2009)	American Football	Video footage	Supervised (expert-labeled, multi-model densities)	Attacking plays (*combo dropback, HITCH dropback, middle/wideleft/wideright run*)
Siddiquie et al. (2009)	American Football	Video footage	Supervised (expert-labeled, feature extraction)	Attacking plays (*left-/middle-/right-runs, option-/short-/rollout-/deep-pass*)
Li and Chellappa (2010)	American Football	Positional data	Supervised	Attacking plays
Stracuzzi et al. (2011)	American Football	Video footage	Supervised (pattern template)	(1) Individual running patterns; (2) Attacking plays
Atmosukarto et al. (2013)	American Football	Video footage	Supervsised (expert-labeled, support vector machines)	Team formations
S. Chen et al. (2014)	American Football	Video footage	Supervised (rule-based)	Game-states (*offense, defense, kickoff, non-punt, field goal plays*)
Hochstedler and Gagnon (2017)	American Football	Positional data	(1) Supervised; (2) Supervised (expert-labeled, feature extraction)	(1) Team formations; (2) Offensive routes

Table 1: Overview of tactical pattern detection in invasion sports.

Study	Sport	Data	Methods	Pattern
Perse et al. (2006)	Basketball	Positional data	Supervised (pattern template)	(1) Game-states (*offensive, defensive, time-out*); (2) Attacking pattern (*screen, move, player formation*)
Perše et al. (2009)	Basketball	Video footage, positional data	Supervised (pattern template)	(1) Game-state (*offensive/defensive/time-out*); (2) Attacking pattern (*screen plays, moves, starting formation*); (3) Attacking plays
Grunz et al. (2009)	Basketball, Football	Positional data	Unsupervised (self-organising maps)	Attacking plays
Mcqueen et al. (2014)	Basketball	Positional and event data	Supervised (expert-labeled, feature extraction)	Screen plays
Kempe et al. (2015)	Basketball	Positional data	(1) Supervised; (2) Unsupervised	(1) Detection of plays (*fastbreak, horns, high-pick*); (2) Tactical behaviours (*3x defensive, 6x offensive, 1x transition, 2x set-piece*)
K.-C. Wang and Zemel (2016)	Basketball	Positional data	Supervised (expert-labeled, neural network)	Attacking plays (*11 offensive plays with all players involved*)
Mcintyre et al. (2016)	Basketball	Positional data	Supervised (expert-labeled, logistic regression)	Defensive counter on screen plays

Table 1: Overview of tactical pattern detection in invasion sports.

Study	Sport	Data	Methods	Pattern
Mehrasa et al. (2018)	Basketball, Ice-Hockey	Positional data	Supervised (neural networks)	(1) Hockey-events (*pass, dump in/out, shot, carry, puck protection*); (2) Basketball team classification
Hojo et al. (2018)	Basketball	Positional data	Supervised (expert-labeled, feature extraction)	Subtypes of screen plays (*off-ball: down, flare, pin, black, flex, cross; on-ball: pick and roll, H and off*)
Tian et al. (2020)	Basketball	Positional data	Supervised	Defensive counter on screen plays (*switch and trap*)
Reep and Benjamin (1968)	Football	Annotated pass sequences	Descriptive statistics	Passing patterns
Borrie et al. (2002)	Football	Event data (purpose specific)	Unsupervised (T-pattern analysis)	Attacking plays
Hirano and Tsumoto (2004)	Football	Event data	Unsupervised (multi-scale matching, clustering)	Passing patterns
Hernández-mendo (2006)	Football	Event data	Unsupervised (T-Pattern analysis)	Passing patterns
Brooks et al. (2016)	Football	Event data	(1), (2) Supervised (feature extraction); (3) Descriptive statistics	(1) Team identification using passing patterns; (2) Prediction whether passing sequence ends in a shot; (3) Passing patterns

Table 1: Overview of tactical pattern detection in invasion sports.

Study	Sport	Data	Methods	Pattern
Grunz et al. (2012)	Football	Positional data	Unsupervised (self-organising maps; extended labeling conducted for evaluation)	Build-up patterns (*long game initiation* versus *short game opening*)
Niu et al. (2012)	Football	Video footage	Supervised (rule-based)	Attacking patterns (6 *pre-defined patterns; ground versus air attack from different starting points*)
Wei et al. (2013)	Football	Positional data	(1) Supervised; (2) Unsupervised (hierarchical clustering)	(1) Game-phase (*in-play, stoppages, highlights, different set-pieces*); (2) Team formations per game-phase
Vilar, Araújo, Davids, and Bar-Yam (2013)	Football	Positional data	Descriptive statistics	Teams covered areas
Bialkowski et al. (2014)	Football	Positional data	Unsupervised (agglomerative clustering)	Team formations
Knauf (2014)	Football	Positional data	Unsupervised (convolution kernels)	Build-up types
Montoliu et al. (2015)	Football	Video footage	Supervised (expert-labeled, bag-of-words)	Attacking plays (*ball possession; quick attacks, i.e., switching the attack and fast break; set-pieces, i.e., direct/indirect freekick, penalty, corner kick*)
Q. Wang et al. (2015)	Football	Event data	Unsupervised	Passing patterns
Fernando et al. (2015)	Football	Positional data	Unsupervised	Attacking plays
Van Haaren et al. (2015)	Football	Event data	Unsupervised	Attacking plays

Table 1: Overview of tactical pattern detection in invasion sports.

Study	Sport	Data	Methods	Pattern
Bialkowski et al. (2015)	Football	Positional data	(1) Supervised (k-nearest neighbours regression); (2) Supervised	(1) Team formations; (2) Team identities
Feuerhake (2016)	Football	Positional data	Unsupervised	Movement patterns
Bialkowski et al. (2016)	Football	Positional data	Unsupervised	Team formations
Chawla et al. (2017)	Football	Positional and event data	Supervised (expert-labeled, feature extraction, inter-labeler accordance)	Passing patterns (i.e., reward *good*, *ok*, *bad*)
Hobbs et al. (2018)	Football	Positional data	Hybrid	Counterattacking patterns
Power et al. (2018)	Football	Positional data	Supervised (expert-labeled, neural networks)	Defensive corner roles (team-level)
Decroos et al. (2018)	Football	Event data	Unsupervised	Attacking plays
Andrienko et al. (2019) (study I)	Football	Positional data	Visual analytics	(1) Team formations; (2) Counterpressing
Müller-Budack et al. (2019)	Football	Positional data	Supervised	Team formations
Narizuka and Yamazaki (2019)	Football	Positional data	Unsupervised (Delaunay method, hierarchical clustering)	Team formations
Shaw and Glickman (2019)	Football	Positional data	Unsupervised	Team formations (per game-state)
Shaw and Gopaladesikan (2021)[17]	Football	Positional data	Supervised (expert-labeled, feature extraction, XGBoost)	Defensive corner roles (player-level)

[17]Note that a slightly different version of that that paper can be found in Haaren et al. (2013).

Table 1: Overview of tactical pattern detection in invasion sports.

Study	Sport	Data	Methods	Pattern
Anzer, Bauer, and Brefeld (2021) (study II)	Football	Positional and event data	Unsupervised (hierarchical clustering)	Goal scoring patterns
Bauer and Anzer (2021) (study III)	Football	Positional and event data	Supervised (expert-labeled, feature extraction, XGboost)	Counterpressing
Bauer, Anzer, and Shaw (2022) (study IV)	Football	Positional and event data	(1) Supervised (expert-labeled, convolutional neural networks); (2) Unsupervised (hierarchical clustering)	(1) Phases of play; (2) Team formations
Bauer, Anzer, and Smith (2022) (study V)	Football	Positional and event data	Supervised (expert-labeled, convolutional neural networks, long-short term memory networks)	Defensive corner roles (player-level and player-marking assignment)
Fassmeyer, Anzer, Bauer, and Brefeld (2021) (study VI)	Football	Positional and event data	Semi-supervised (expert-labeled, variational autoencoder, support vector machines)	(1) Events (*corner kicks*, *crosses*); (2) Counterattacks
Vilar et al. (2012)	Futsal	Video footage	Descriptive statistics	Attacker-defender dyads
Lucey, Bialkowski, et al. (2013)	Field-Hockey	Positional data	Supervised (formation templates)	Team formations
Karsten et al. (2017)	Rugby union	Positional data (GPS)	Supervised (expert-labeled, random forest)	Scrum events
Bunker et al. (2020)	Rugby	Event data	Supervised (rule-based, inter-observer consistency)	Attacking plays

Table 1: Overview of tactical pattern detection in invasion sports.

Study	Sport	Data	Methods	Pattern
Pfeiffer and Perl (2015)	Team handball	Event data (purpose-specific)	Supervised (expert-labeled, neural networks)	Attacking plays

2.3.4 Phases of Play in Football

The concept of tactical patterns is used in different invasion sports. On the basis of Gréhaigne et al. (1997), we embed frequently occurring tactical patterns in football in a game-model. Figure 2 shows an overview of *phases of play*, further defined as tactical patterns that are established as sub-categories of game-states (i.e. offensive, defensive, transition to offense, transition to defense and set-pieces) in professional football.[18] Although minor differences occur due to different playing philosophies, the taxonomy in Figure 2 is consolidated in several discussions with professional coaches and match analysts from German Bundesliga teams and the German national teams (see Acknowledgements). Figure 2 differentiates between common game-states (offensive, defensive and transition) on the highest level. Set-pieces, uncontrolled possessions and situations where the ball is out of play are listed as a separate game-state respectively due to their specific characteristics (Wei et al., 2013). For transitions we further distinguish between those to offense, and those to defense. Inter alia, we embed well known strategies like *counterattacks* (Hughes & Lovell, 2019; Tenga, Holme, Ronglan, & Bahr, 2010) or *counterpressing* (Low et al., 2020; Hobbs et al., 2018) as optional tactics during transitions. In offense, *build-up*—advancing possession from the own goal behind the first defending line of the opponent—is separated from *attacking play*, in which a team has possession in the last third of the pitch and purely follows the objective of scoring a goal (Rein & Memmert, 2016; Plummer, 2013). The defensive game-state is separated based on the height and activity of a team defending their own goal. *Low-block* is a very passive strategy, where a team focuses on the protection

[18] Similar frameworks for phases of play has been presented for American football (Siddiquie et al., 2009) and for football (Wei et al., 2013).

of the own goal. On the other hand, teams can block the opponents possession (i.e., the build-up) actively and as close to the opponents goal as possible, typically declared as *high-block* (Power et al., 2017). *Midfield-block* or *mid-block* describes a moderate defending tactic. The eleven phases of play presented as sub-categories of six game-states in Figure 2, again, can contain further (highly individualized) tactical patterns, which are either performed by groups of players (i.e., one-twos, overlapping runs or other passing patterns) or present specific manifestations, for example different types of counterattacks presented in Hobbs et al. (2018).

Following the primary objective—the detection of tactical patterns—the contribution of this thesis is outlined with reference to Figure 2. First, two exploratory studies I & II are presented analyzing purely data-driven patterns (Sections 3.1 and 3.2), helping us to derive the taxonomy presented in Figure 2. Study I (Section 3.1) presents an unsupervised approach using visual analytics to explore differences in team behavior per game-state. Another exploratory overview on analyzing goal scoring patterns (including counterpressing, counterattacks, high-block and mid-block defending) is presented in study II (Section 3.2). Transition to defense, particularly the detection of counterpressing is addressed in study III (Section 3.3). Counterattacks, a sub-category of transition to offense are addressed in study VI (Section 3.6). Study IV analyses team formations by first detecting the five phases of offensive and defensive play (Section 3.4). Finally, in study V (Section 3.5), defending patterns in set-pieces, i.e., corner kicks are detected.

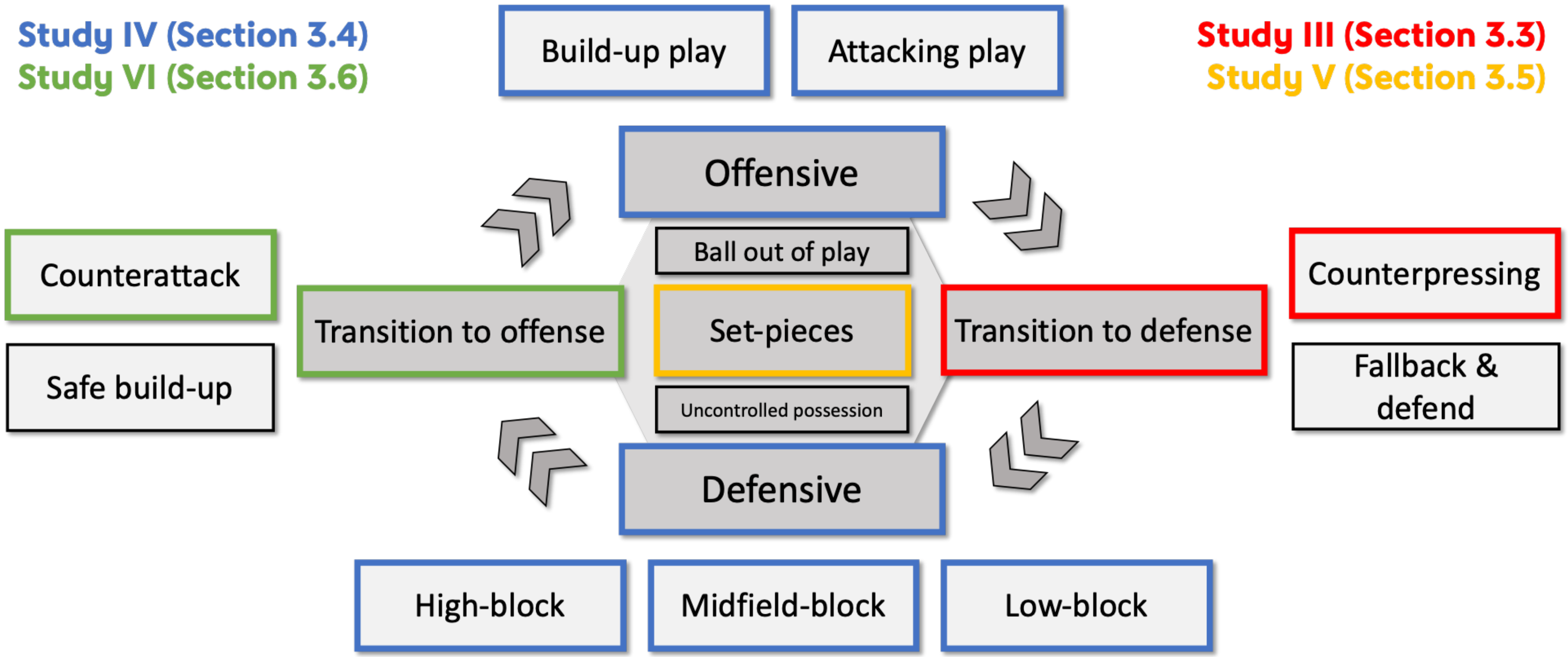

Figure 2: Overview of phases of play in football. The contribution of this dissertation is highlighted using different colours, in relation to the game-states (offensive, defensive, transitions to offense/to defense and set-pieces).

3 Empirical Studies

In the following sections we summarize the empirical studies. Instead of presenting detailed explanations, the respective summary focuses on the contribution of each study towards the above defined research question—details can be found in the respective full text attached in the Appendices.

In Sections 3.1 and 3.2 the importance of investigative, unsupervised approaches to determine further research questions is outlined. Building on these results, Sections 3.3, 3.4 and 3.5 present supervised machine learning approaches using close cooperation with football experts to automatically detect tactical patterns. Finally, Section 3.6 provides an outlook towards the potential of semi-supervised learning that can be used effectively for the purpose of tactical pattern detection.

The contribution to the empirical studies I, II & VI in Sections 3.1, 3.2 and 3.6 is a co-authorship, whereas the articles presented in studies III, IV & V (Sections 3.3, 3.4 and 3.5) are conducted as first-author.

3.1 Study I: Constructing Spaces and Times for Tactical Analysis in Football (Andrienko et. al. 2019)

Complex positional data can only be processed by experts using aggregations and high-level visualizations depicting interesting patterns. Visual analytics has been a relevant issue in football analytics (Andrienko et al., 2017; Sacha et al., 2017; Perin, Vuillemot, & Fekete, 2013; Wu et al., 2019). Study I (Andrienko et al., 2019) is the result of a common research project of visual analytics experts, data-scientists and football experts with the goal to explore tactical patterns using different methodologies. This work can thus be seen as a general introduction to the research project, using simple aggregation and query techniques with a strong emphasis on visualizations in order to explore tactical patterns and team tactics in the data.

Using a substantial portion of domain expertise, data have been (1) queried and filtered to select time intervals of interest, (2) aggregated to get an overview over different time intervals, and (3) visualized in a way that patterns can be explored from illustrations. First, we analyzed team formations. Traditional approaches study the formation of a team, specifically each player's role, related to the center of the pitch. By introducing and visualizing the team-space—player positions related to their teams' average position—we help practitioners to capture the interaction of teammates in greater detail. Second, by presenting various queries we show the dynamics of team formations and player rules. For example, we visualize differences in formations per game-state (offensive, defensive, transition), per score (e.g., teams falling back after goals), and how players interpret the same role differently. This exploratory analysis guided the research investigated in study IV (Bauer, Anzer, & Shaw, 2022) in Section 3.3. Lastly, transition scenarios have been analyzed and visualized. By proving the practitioner assumption, that the transition to defense strategy varies significantly per team, we motivate our work on the detection of counterpressing (Bauer & Anzer, 2021) presented in study III (Section 3.3).

The major limitation of this approach is that little evidence or usable insights can be derived by visualizations and/or the small amount of data used in the experiment (essentially one match of positional and event data). However, the exploratory approach and the collaboration of people with different expertise and perspectives turned out to be very beneficial for our research project. Even though the applied methods fall slightly out of scope compared with the rest of the thesis, various results of this project served as a motivation for the rest of the projects.

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

In professional football, match analysis departments categorize goals scored and received (e.g., open play versus set-pieces on the highest level) for their own matches (several times in a season) and for their upcoming opponent (on a weekly basis). Approaches differ drastically among experts due to a low accordance on well-defined and unique categories of goals scored. In addition to the time saved using automated categorization, the problem of goal origins has been studied for decades in the scientific domain (Reep & Benjamin, 1968; González-Ródenas et al., 2019; Njororai, 2013; Mitrotasios & Armatas, 2012).

In Anzer, Bauer, and Brefeld (2021) we follow an unsupervised approach to explore objective goal categories using a sample size that would exceed human capacities. Based on synchronized positional and event data consisting of $3,457$ goals from two seasons of German Bundesliga and 2$^{\text{nd}}$ Bundesliga (2018/20219 and 2019/2020), we devise a rich set of 37 features that can be extracted automatically, and propose an agglomerative hierarchical clustering approach to identify group structures. The features describe the attack leading to a goal and contain, for example, the duration of the attack, the location of the assist and the shot. Feature extraction, choosing the number of clusters and contextualizing the clusters based on video footage has been conducted in close collaboration with football experts. The results consist of 50 interpretable clusters revealing insights into scoring patterns.

The clustering found eight highly separated clusters (penalties, direct freekicks, kick and rush, one-two's, assisted by header, assisted by throw-in) and nine categories (e.g., corners) combining more granular patterns (e.g., five subcategories of corner-goals). Again, the insights motivated further work: One cluster is comprised of 124 goals later contextualized as goals after counterpressing by the experts (motivating Section 3.3). The clustering also revealed 93 goals after high-block pressing, as well as 73 goals after mid-block pressing (motivating Section 3.4).

By automating the analysis of goal scoring patterns using data, with 3,457 goals, we exceed the sample sizes used in traditional studies (Reep & Benjamin, 1968; González-Ródenas et al., 2019; Njororai, 2013; Mitrotasios & Armatas, 2012). Consequently, we are able to reveal patterns that could not have been detected as such using less goals. The major limitation, typical for unsupervised machine learning approaches, is that they can not be used seamlessly in practice. The meaningful results can rather be used to consolidate and define clear categories of goals to build a supervised machine learning approach in future projects.

3.3 Study III: Data-Driven Detection of Counterpressing in Professional Football (Bauer & Anzer 2021)

After losing the ball, a team conducts counterpressing if at least one player exerts (spatio and/or temporal) pressure on the ball carrier, or on the opponents close to the ball. The relatively young strategy of counterpressing is admittedly touched upon but never properly investigated in literature (Anzer, Bauer, & Brefeld, 2021; Low et al., 2020; Andrienko et al., 2019; Hobbs et al., 2018). Analyzing counterpressing is an important task for any professional match analyst in football, but is being done exclusively manually by observing video footage.

In study III (Bauer & Anzer, 2021) we present the first approach in football using supervised machine learning to detect a complex team-tactical pattern in open-play (see Table 1).[19] The primary purpose of this paper is to automatically identify this strategy using positional and event data. Together, with professional match analysis experts we discussed and consolidated a consistent definition of counterpressing, extracted 134 features and manually labeled 20,928 defensive transition situations from 97 professional football matches. The features describe the constitution of the teams using attributes like team-center, stretch-index (Santos, Theron, Losada, Sampaio, & Lago-Peñas, 2018), or pressure on the ball carrier (Andrienko et al., 2017) and were extracted at three discrete timestamps during a transition to defense (at the time of the ball possession change, plus one and two seconds after). This provides a drastic reduction of input data dimensionality. We present a comprehensive inter-labeler reliability with a pair-wise labeling accuracy of 82.01% and provide rule-based baseline models in order to give an indication of the task-complexity (area under the curve 60.2%). The trained XG-Boost model—with an area under the curve of 87.4% on the labeled test data—enabled us to judge how quickly teams can win the ball back with counterpressing strategies, how many shots they create or allow immediately afterwards, and to determine what the most important success drivers are. We applied this automatic detection on all matches from six full seasons of the German Bundesliga and quantified the defensive and offensive consequences when applying counterpressing for each team.

[19] Power et al. (2018) addressed defensive behaviors during corners as a team-tactical pattern using supervised machine learning techniques.

To capture all counterpressing situations manually, the full match has to be observed at least once using a dedicated tagging tool. Further efforts has to be spent to review the labels leading to an overall manual effort of roughly two hours per match. In two experimental studies, the effort was reduced by automatically suggesting 15–30 counterpressing scenes per match, which can be immediately observed. Consequently, automating the task saves analysts a tremendous amount of time, standardizes the otherwise subjective task, and allows to identify trends within larger data sets. We present an effective way of how the detection and the lessons learned from this investigation are integrated effectively into common match analysis processes. For example, Figure 3 shows how the outcome of the counterpressing detection is used in match-reports for the German national teams. Instead of manually screening the video footage of all transitions, the plot overviews the success rate of counterpressing (defined as a regaining possession within the five seconds). Considering only defensive transitions of the German team, Figure 3 shows that the detection leads an analyst directly to the 33 scenes of interest out of a total of 164 ball losses in the opposing half. One outcome of the study is that counterpressing is a risky tactic. Figure 3 shows the German U21 national team conceded two goals after unsuccessful counterpressing attempts (within 20 seconds after the counterpress). The green/red bullets show the ball losses before successful/unsuccessful counterpressing was conducted. In the bottom line, several performance indicators are benchmarked against Bundesliga average.

A major limitation of the approach is the exhaustive human endeavor required for such an experiment. In the presented study, a total of 97 matches (roughly 140 hours of video footage) had to be observed by at least one football expert. Not only

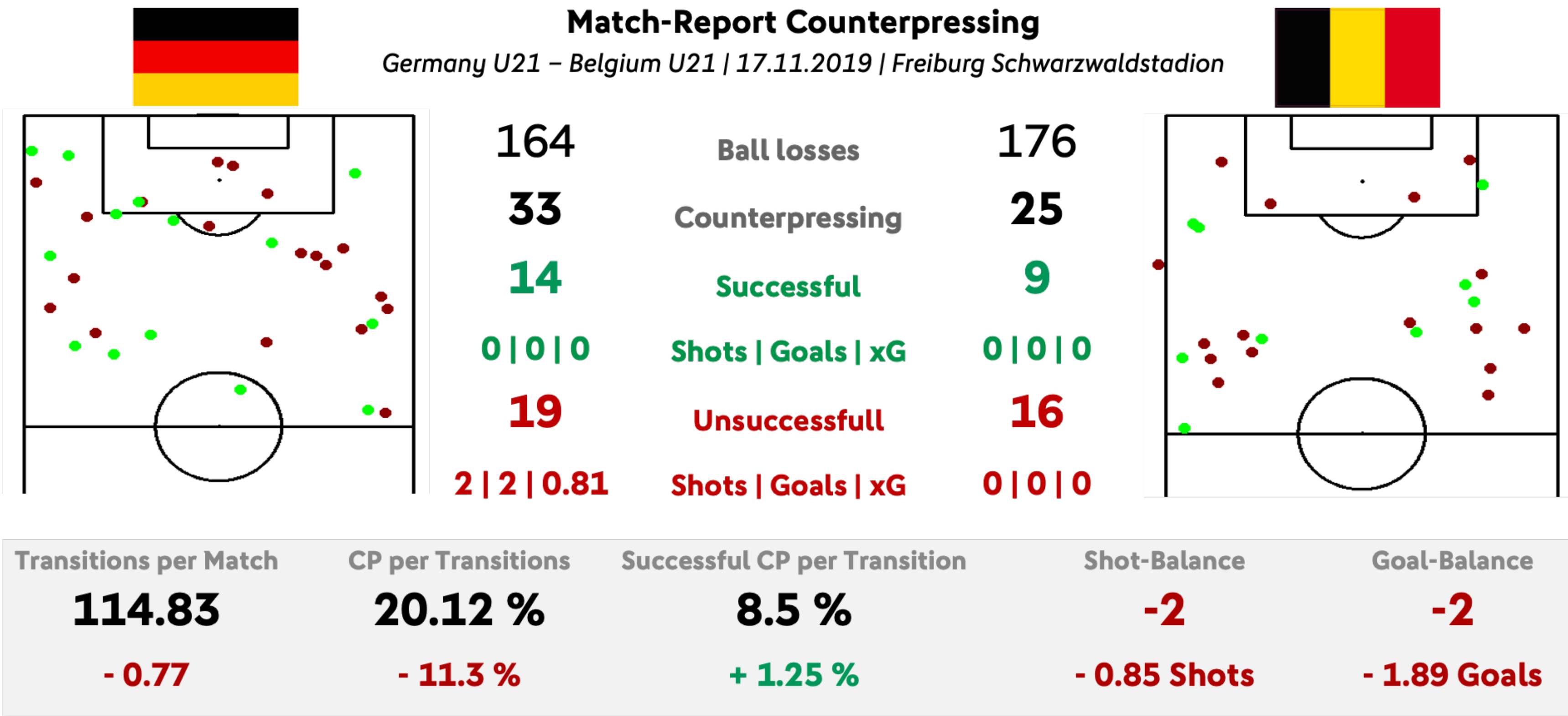

Figure 3: Counterpressing. Excerpt from match-report of the U21 national team match Germany against Belgium. This figure is copied from Bauer and Anzer (2021) and further described in the full text (see Appendix).

labeling, but also feature extraction is an elaborate process. Although many features could be re-used to identify other tactical patterns, using end-to-end algorithms handling the raw data, as presented in the next Section 3.4 (study IV), would seem to be more efficient.

3.4 Study IV: Putting Team Formations in Association Football into Context (Bauer, Anzer, & Shaw 2021)

Choosing the right team formation in football is a fundamental tactical and strategical decision for coaches. The availability of accurate positional data motivated ample research on team formations using supervised (Narizuka & Yamazaki, 2019; Shaw & Glickman, 2019; Bialkowski et al., 2016, 2015, 2014; Wei et al., 2013) and unsupervised approaches (Müller-Budack et al., 2019). However (as indicated in study I) formations change dynamically and therefore should not simply be aggregated over an entire match (Andrienko et al., 2019; Shaw & Glickman, 2019; Gudmundsson et al., 2017; Bialkowski et al., 2016; Lucey, Bialkowski, et al., 2013). Past literature focused primarily on aggregating player positions across all game-states using positional data (Wei et al., 2013; Bialkowski et al., 2014, 2015). Only Bialkowski et al. (2016) and Shaw and Glickman (2019) explored differences in team formations between different game-states, i.e. offensive, defensive and transition. However, prior work did not consider formations related to more granular phases of play like build-up versus attacking play within offense.

To address this gap in literature, in study IV (Bauer, Anzer, & Shaw, 2022), we first detect those phases of play from Figure 2 in which formations are of interest for practitioners (namely build-up, attacking play, high-/mid-/low-block) at each moment of the game using a convolutional neural network with an average F_1-score of 0.76. To train this model all phases of play of 97 matches have been labeled manually at each frame (in total 59 hours and 50 minutes of the above listed phases were labeled). Again, for this supervised machine learning approach, we present a verbose labeling experiment using definitions consolidated among experts, different labelers to analyze their inter-labeler accordance and baseline models indicating the complexity of the classification task. Detecting those phases, Figure 4 shows the average positioning of a team (playing from left to right) in the respective phases during a match, show that phases of play considerably influence team formations. The formations have been contextualized (i.e assigned to a 5–3–2 for the low-block phase of play) in close consultation with football experts. The ellipses show the 80% confidence region for each player.

We then measure and contextualize unique formations per phase of play by hierarchically clustering the respective sequences for seven seasons of German Bundesliga (2013/2014–2019/2020). Instead of over-simplifying a team's formation across the whole match into widely established three- or four-digit codes (e.g., 4–4–2 abbreviating 4 defender, 4 midfielder and 2 attacker), we provide an objective and granular representation of teams formations per tactical phase of play. Using the most frequently occurring phases of play, mid-block, we identify and contextualise six unique formations (4–2–3–1, 4–4–2, 4–1–4–1, 4–3–2–1, 5–3–2 and 5–2–3). The definitions of the formations, including the distinction between three and four lines, has been defined by the involved experts. A long-term analysis in the German Bundesliga allows us to quantify the efficiency of each formation against others, and also to present a helpful scouting tool to identify how well a coach's preferred playing style suits to a

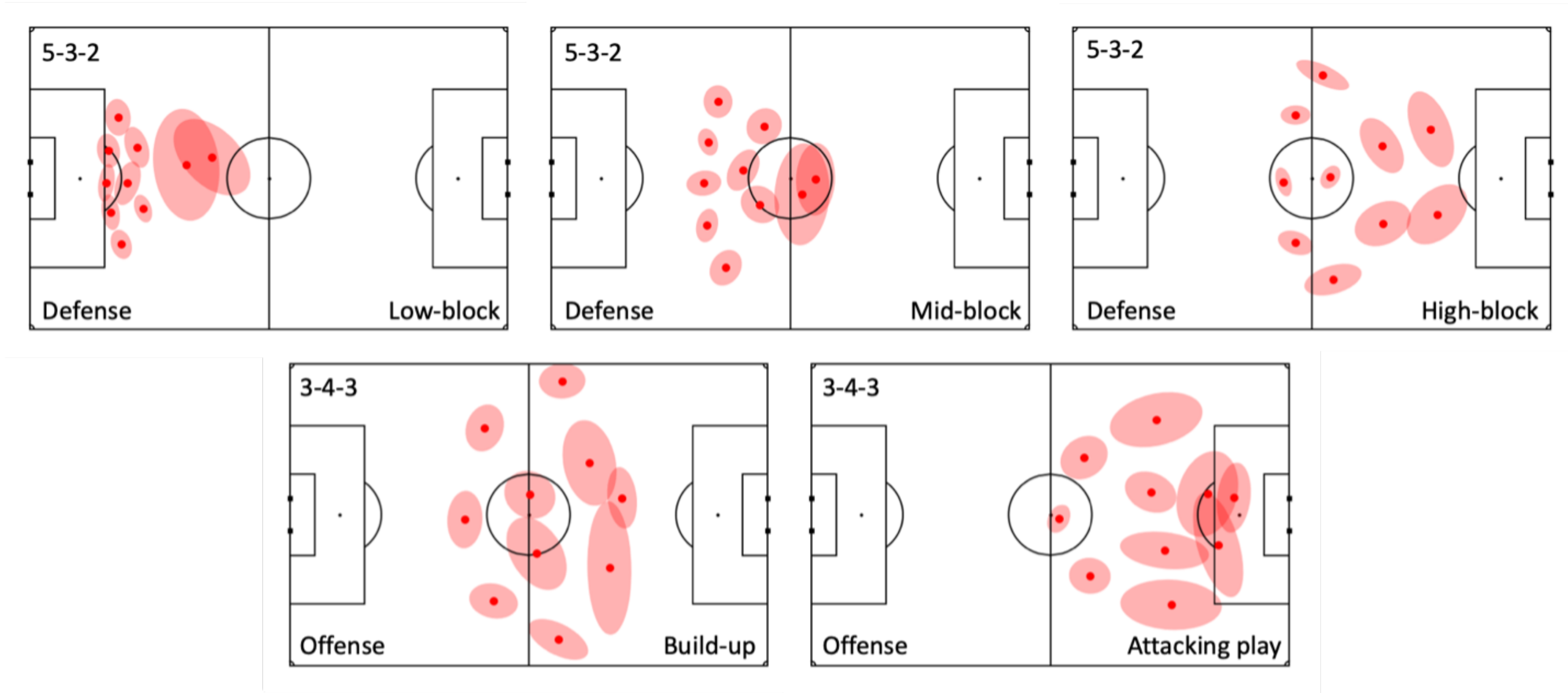

Figure 4: Team formations per phases of play. A similar Figure is shown and further described in Bauer, Anzer, and Shaw (2022).

potential club.

Following our research question—the detection of tactical patterns—this paper contains two major contributions:

(1) The supervised detection of five phases of play, namely low-/mid-/high block, build-up and attacking play.

(2) The unsupervised exploration of team formations per phase of play.

By transferring the raw positional data to images, and using convolutional neural networks, in (1) we present a major advantage compared to the method presented in study III (Section 3.3). This allows us to accurately detect five phases of play with just one trained algorithm and without manual feature extraction. For (2) we present an interpretation of team formations as tactical patterns, that should rather be defined on specific sequences than aggregated over a whole match or over basic game-states (offensive, defensive, transition). In future investigations, more focus has to be put on the analysis of formation efficiencies, which is only touched upon in the practical application of our approach.

3.5 Study V: Individual Role Classification for Players defending Corners in Football (Bauer, Anzer, & Smith 2021)

Choosing the right defensive corner-strategy is a crucial task for each coach in professional football.[20] Due to their repeatable and relatively static set-up, corners are an obligatory investigation for pattern analysis using positional data: Power et al. (2018) detects the defensive strategy on a team-level (player-marking, zonal-marking, hybrid) using expert-labeled data and

[20]Following many other teams, the German national squad recently hired a dedicated set-piece coach purely focusing on set-piece tactics https://www.kicker.de/als-hansi-anrief-dachte-ich-da-verarscht-mich-jemand-867987/artikel, accessed 28.08.2021.

neural networks. Shaw and Gopaladesikan (2021) extended the automated distinction between player- and zonal-marking to a prediction on a player-level. For this task, the ordering problem (mentioned in Section 2.2) becomes challenging. To overcome this issue, Shaw and Gopaladesikan (2021) extracted features on a player level and perform an XGBoost model predicting the role for each player (neglecting player interactions).

Our work in Bauer, Anzer, and Smith (2022) addresses this problem by combining a convolutional neural network (to handle the ordering-problem) with a long-short-term memory (to capture the temporal interaction of two players). By doing so, we identify which of seven well-established roles a defensive player conducted (*player-marking, zonal-marking, placed for counterattack, back-space, near-post, far-post*, and *short-defender*). Further, in case of player-marking we detect which attacking player is marked, which is a relevant extension to Shaw and Gopaladesikan (2021). We hand-labeled the role of each defensive player from 213 corners in 33 matches, where we then employ an augmentation strategy to increase the number of data points. The model achieves an overall weighted accuracy of 89.3%, and in the case of player-marking, we are able to accurately detect which offensive player the defender is marking 80.8% of the time. The performance of the model is evaluated against a rule-based baseline model, as well as by an inter-labeler accuracy.

For practical usage, we show three concrete use-cases on how this approach can support a more informed and fact-based decision making process for defensive corner strategies. Figure 5 shows an excerpt from a match-report using our algorithm. Each grey or colored line indicates a player's action for one corner of that match. For each defensive player (red ellipses) the roles per corner are shown (link to boxes on the left side). In case of player-marking, the links to the opponents (green ellipses) show which player was marked. To highlight the insights for coaches and analysts using the report, the figure also indicates who touched the ball first for each corner and whether an attacker was able to create a goal or shot (within 18 seconds after the corner).

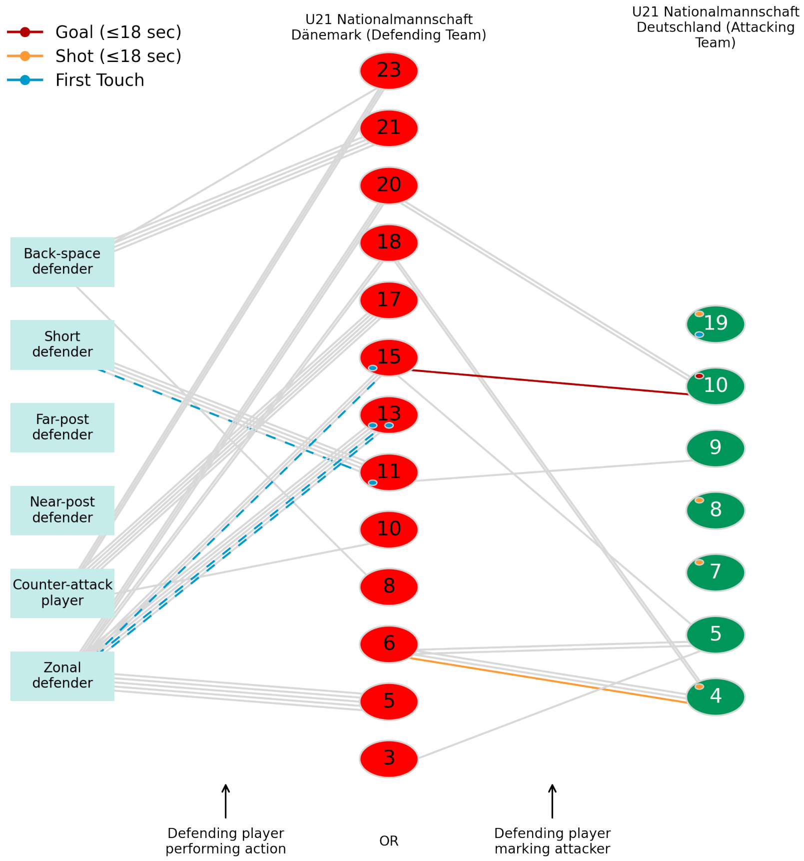

Figure 5: Defensive player roles for corners. Excerpt from match-report of the German U21 national team against Denmark. This figure is copied from Bauer, Anzer, and Smith (2022).

The largest limitation is the amount of manual effort required to annotate the training data. To reduce the labeling time, the data augmentation presented a massive improvement by multiplying the set of training data by a factor of ten without overfitting the model (due to few unique samples). Further, we support the labeling with rule-based suggestions, which are also used to perform a weak supervision approach, included as an outlook of the paper. Both for this approach, as well as for the phases of play detection in Section 3.4, the computing complexity when handling images using convolutional neural networks is another relevant limitation. Transferring the spatio-temporal positional data into a sparsely populated image actually increases the dimensionality of the input data. Although image classifications is a well researched area, where plenty of tweaks in the architecture of convolutional neural network can handle those high dimensional data, future work should consider using graph neural networks as indicated in Dick et al. (2021) or Stöckl, Seidl, Marley, and Power (2021) to model players interaction and learn individual classifications.

3.6 Study VI: Torward Automatically Labeling Situations in Football (Fassmeyer et. al. 2021)

Using semi-supervised methodologies to reduce the expert-labeling time required has become a relevant issue in machine learning research. Previous studies showed that there is also a need to apply such strategies when detecting tactical patterns in football. Study VI, Fassmeyer et al. (2021) can be seen as an outlook of the thesis in order to explore further semi-supervised approaches towards the detection of tactical patterns.

We split the problem into two parts and learn (1) a meaningful feature representation using variational autoencoders on unlabeled data at large scales, and (2) a large-margin classifier acting in this feature space but using only a few (manually) annotated examples of the situation of interest. Both a static (one fixed time frame) and a temporal (sequence of positional data) autoencoder is implemented using the transformation of positional data into an image also presented in studies IV & V. As a proof of concept for the novel methodology, corner kicks and cross events are detected by support vector machines applied to the encoded data in a lower-dimensional space. The detection is compared against event data available abundantly. Since the results are sufficient, even compared against other approaches aiming to derive event data from player positions (Stein et al., 2019; Richly et al., 2016; Motoi et al., 2012; Gudmundsson & Wolle, 2010; M. Zheng & Kudenko, 2010), we further applied the approach to counterattacks, a more complex tactical pattern that is of high interest for practitioners (Hobbs et al., 2018). For counterattacks only 60 positive examples (27 training data; 33 test data) of a single match had to be hand-labeled to achieve a sufficient accuracy (area under the curve: 0.912; F_1-score: 0.730) using the sequential variational autoencoder in combination with the support-vector machine classifier. For interpretability of the approaches, we investigate different false positives and false negatives of the detection in detail, showing that even the misclassifications are reasonable for experts.

The presented study suggests the potential of semi-supervised methods for the identification of tactical patterns. The approach focused on evaluating a new methodology, however, for practical usage further work has to be conducted, i.e. a generalisation to other patterns or the integration into an application.

4 Discussion

As shown in Table 1 and Figure 2, we present a substantial contribution towards the detection of tactical patterns in football. The first part of our research question—*whether complex tactical* *patterns in football can be detected automatically*—can be affirmed looking at Figure 2. All five standard offensive and defensive phases of play are detected in study IV (Section 3.4), counterpressing (Section 3.3) and counterattacks (Section 3.6) are detected in studies III & VI. Further, tactical patterns during corners, as an important example of a set-piece are investigated in study V (Section 3.5). Referring to the introductory statements of Richard Bate (Bate, 1987) and Simon Kuper (Kuper, 2018), we focus on two primary values for football (rather than claiming to provide groundbreaking insights that can fundamentally change tactics and strategy): On the one hand, we believe that the approaches presented can provide evidence for coaches' opinions and thus *support decision making on tactics and strategy*. Precisely, a simple frequency analysis on tactical patterns of interest (shown in Figures 3 and 5) can help to objectify otherwise biased opinions (Borrie et al., 2002), or to reconstruct all situations of interest entirely (Laird & Waters, 2008). Second, we present an effective integration of the *process automatization* into the everyday-business of professional clubs or federations, especially in studies III, IV & V. We show how repeated tasks, being conducted typically by dedicated match analysts observing hours of video footage, can be supported or fully realized by machine learning algorithms. By saving time in various use-cases (e.g., opponent analysis, team analysis, scouting of players or coaches) the applications presented can support the experts so that they can focus on qualitative analysis of scenes of interest rather than repeatedly annotating those scenes in the video footage.

For the second part of the research question—*the how of tactical pattern detection*—we present multiple ways using different methodologies: exploratory visual analytics (study I), unsupervised machine learning methods (study II, study IV), supervised methods (study III, study IV, study V), as well as semi-supervised approaches (study VI) and other labeling support methods (i.e., rule-based support, data augmentation and weak supervision in study V). For the detection of tactical patterns, the *interplay between unsupervised or other exploratory approaches* (e.g., visual analytics) and *supervised techniques* has proven to be a fruitful combination on the analysis of tactics and strategy in football. This dissertation indicates that the related learnings can be transferred to various domains, and aligns with Tuyls et al. (2021), claiming that football analytics can offer huge potential for the research domain of machine learning and artificial intelligence Transforming spatio-temporal multi-agent data into a permutation-invariant space is not only a problem in sports analytics (Battaglia et al., 2018). Recent improvements in graph neural networks can solve this problem, whereby positional data in football and basketball are favored testbeds due to the vividness of results (Dick et al., 2021; Games, 2019; Yeh et al., 2019; Sun et al., 2019; Kipf, Fetaya, Wang, Welling, & Zemel, 2018). A similar concept—stating that theoretical research domains can benefit from application areas and vice versa—has been proposed by Höner (2008) at the intersection of sport science and psychology. As mentioned in the introduction, various work has been conducted outside this thesis bridging the gap between sport and data science: In Herold et al. (2021) and Herold et al. (2019) the awareness of football practitioners and data-driven metrics is analyzed. In Anzer, Bauer, and Höner (2021), the basic concept of machine learning (explained with references to football examples) is presented in a match analysis textbook. Anzer and Bauer (2022) and Anzer and Bauer (2021) use granular metrics to objectively quantify passing and shooting performance. These metrics are not only established in sport science research (Herold et al., 2021) but also support performance evaluation in practice.

The methodological approach, specifically the rigorousness while creating and evaluating labeled data in our studies, is comparable to Chawla et al. (2017). For example, Chawla et al. (2017) hand-labeled in total $2,932$ passes manually, whereas in study III we use $3,196$ expert-annotated counterpressing scenes for our detection. As in Chawla et al. (2017)[21] we show a high inter-labeler accordance in study III (pairwise labeler accuracy 82.01%), study IV (e.g., inter-labeler average F_1-score for mid-block 0.78) and study V (e.g., pairwise accuracy of player-marking detection 94.3%). However, the application on team-tactical patterns presented in this thesis is unique. For each tactical pattern detected in the supervised set-up (studies III, IV, V) we consolidate definitions in close collaboration with experts, create annotated data from different expert-labelers and steadily monitor their labeling accordance. A central learning of this thesis is that the inter-labeler reliability, as a measure of how well-defined a pattern is, should be monitored as an integral part of such studies. Related to this, the close *collaboration with practitioners* can be seen as another major strength of this dissertation. F. R. Goes, Brink, Elferink-Gemser, Kempe, and Lemmink (2020); Herold et al. (2019) and Rein and Memmert (2016) allege that collaborations among machine learning and sport science experts is key to success for the area of sports analytics. This is even more appreciable for manual expert-labeling, hand-crafted feature extraction, the contextualisation of the results and, last but not least, the concrete problem definition—all fundamental constituents of supervised learning. Hence, for projects using supervised ma-

[21] The inter-labeler reliability is presented as the Cohen's kappa. For the six class prediction (very good, good, marginally good, marginally bad, bad, very bad) the Cohen's kappa among the two labelers is 0.393, which is drastically improved with the three-class classification (good, bad, ok) yielding a Cohen's kappa of 0.697.

chine learning, we claim that the beneficiaries of the approach must be involved in project teams. Further, we maintain that there is a need for *data-literate match analysts and coaches* with a basic understanding of machine learning methodologies in order to transfer insights to practice. This will allow them to contribute to projects and to draw informed recommendations for actions based on data-driven results.

Whenever possible we compared our results against published benchmarks in the respective studies, especially in study III, IV & V (see Appendices). A comparison across sports and different patterns (with different complexities) warps reality, however, similar accuracies are achieved as presented in literature. With an area under the curve of 0.874, counterpressing—a complex team-tactical pattern in football—is detected in study III. The results are comparable to the detection of screen plays —a group-tactical pattern in basketball—in Hojo et al. (2018) (area under the curve: 0.855 off-ball, 0.941 on-ball). The F_1-score of 0.748, presented in Chawla et al. (2017) classifying passes according their reward is comparable to our counterpressing detection (study II; F_1-score: 0.67), the frame-wise detection of mid-blocks (F_1-score: 0.80) or build-up plays (F_1-score: 0.83) in study IV. Again, these comparisons have major flaws due to way different complexities of the patterns and confounding factors like data-quality, inter-labeler reliability, and many more, and should serve as a rough orientation only. To provide an idea of the task complexity, studies III, IV & V contain rule-based baseline models helping to give context to the accuracy of the machine learning model. According to this thesis, understandable baseline models should also be standardized for future work on the detection of tactical patterns.

4.1 Limitations and Future Work

Despite the great level of detail in which *data in football* are available, remaining inaccuracies, objectiveness, human error (i.e., for event data collection), and missing information (e.g., player pose) can still be seen as a pitfall of all data-driven approaches in football. The major shortcoming of optical tracking systems is, on the one hand, that they aggregate complex movements of a player to a two-dimensional position, neglecting players orientation and pose (Arbués-Sangüesa, Haro, Ballester, & Martín, 2019; Arbues-Sanguesa, Martin, Fernandez, Ballester, & Haro, 2020). On the other hand, they require an exhaustive set-up on-site with cameras from different, demanding positions (Manafifard et al., 2017). Thus, the acquisition of accurate positional data from basic video-footage (e.g., television signals), as well as appropriate signal processing to compare positional data from different tracking systems (Taberner et al., 2020) are relevant issues for future research. Another pertinent issue is the automated creation of event data from video footage (Pouyanfar & Chen, 2016; Kolekar et al., 2009; Ekin et al., 2003) or from positional data (Stein et al., 2019; Richly et al., 2016; Motoi et al., 2012; Gudmundsson & Wolle, 2010; M. Zheng & Kudenko, 2010). Although event data collection uses sophisticated software-support to make the hand-crafted annotation process as efficient as possible, automating and objectifying this manual task would leverage sports analytics to another level. However, given the recent success in data collection and data quality enabling sports analytics to emerge to a growing research area, we expect substantial further development on the accuracy and granularity of data in all sports.

Nevertheless, positional and event data will always be an abstraction of the real world proceedings, which does not capture all facets of a football match. Thus, more interdisciplinary approaches are required to cover confounding factors (e.g., psychological components) overlooked in the digital reproduction. Establishing the *interplay between qualitative and quantitative methodologies* poses a great potential for future investigations. In the studies presented, the formulation of definitions is often pragmatic and purpose-driven in order to perform the detection task with preferably high accuracy. We state that more sophisticated qualitative studies, grounded solidly on match analysis literature and rigorous expert interviews, should be established to derive proper definitions (e.g., for tactical patterns in general or specific tactical patterns like counterpressing) and also to develop and consolidate taxonomies as presented in Figure 2.

As a third limitation, the *exhaustive labeling effort* required to perform supervised machine learning does not meet the practical requirements of club-, team-, or coach-specific interpretations of tactical patterns. Evolving patterns and the high fluctuation of key-roles in professional football (i.e., coaches and managers) means integrating new philosophies and setting new priorities in a short amount of time. Although Grunz et al. (2012) affirmed that it is hard to formulate precise definitions for tactical patterns, and consequently, that rule-based definitions are inappropriate, study V (Bauer, Anzer, & Smith, 2022) shows that heuristics can add value to supervised machine learning approaches in different aspects. They can serve as baseline models (see also studies III & IV), simplify the human labeling process in complex scenarios (like individual player annotations at corners in study V), and even create weakly supervised training data. Studies V & VI also shows that data augmentation (Bauer, Anzer, & Smith, 2022) and variational autoencoders (Fassmeyer et al.,

2021) can drastically reduce labeling efforts without loosing accuracy. However, labeling support methodologies such as active learning (Druck, Settles, & McCallum, 2009), weak supervision (Ratner et al., 2017), semi-supervised learning (Cholaquidis, Fraiman, & Sued, 2020) or transfer learning (Panigrahi, Nanda, & Swarnkar, 2021) are of high relevance for the area of football (and sports) analytics and should guide future work of the detection of tactical patterns.

Another general problem in sports analytics is *comparability and reproducibility*. Positional and event data are typically the confidential property of leagues or clubs. As a consequence, only researches working for these organisations or closely related research groups can access the respective data. In theory, the increasing availability of open-source datasets (see Section 2.1) allows for reproducibility. However, in practice, different definitions for event data and tactical patterns, varying data-accuracies and other factors constrain comparisons. To ensure reproducibility, a wide range of researchers need to access the exact same dataset to transparently compare their results. Recently, in American football[22], basketball[23] and football[24] positional and event data were open-sourced within (public) competitions. These competitions not only attracted plenty of (machine learning) research groups (which would rarely access sports data otherwise), but also laid the foundation for solid scientific progress in the area of sports analytics. In football, an objective comparison of the top European leagues, as well as the

[22]The NFL Big Data Bowl is hosted annually since 2018/2019 by the National Football League (NFL) endowed with a price of up to 100.000$ (https://operations.nfl.com/gameday/analytics/big-data-bowl/).

[23]In 2019, the National Basketball Association (NBA) hosted a closed hackathon (https://hackathon.nba.com/).

[24]The DFB-Akademie, Eintracht Frankfurt and the Sportec Solutions AG hosted a *Hackathon* with selected participants in 2020 (https://www.dfb-akademie.de/hackathon2/-/id-11009109).

big international competitions on a team and federation level, exploits a huge potential for future work.

5 Conclusion

Concluding the overall research program, positional and event data can help practitioners to make informed decisions instead of purely relying on their gut instincts. Methods of machine learning can further help to capture and model the dynamics and complexity in football. Purely applying these methods to positional and event data will not suffice, they should rather be integrated into interdisciplinary approaches combined with qualitative research and domain-experts.

For the detection of tactical patterns, this dissertation shows these patterns across all phases of play are well-defined, can be identified by experts with a high accordance and can automatically be detected when applying machine learning algorithms on positional and event data. Additionally, this interdisciplinary approach does not only combine different methodologies (i.e. supervised machine learning with exploratory approaches), but attests to the effectiveness of integrating the results into the everyday business of professional football clubs. By doing so, this thesis shows how match analysts and coaches can save time for the otherwise manual task of annotating tactical patterns, and provides a baseline for evidence-based decisions on tactics and strategies.

References

Albert, J. (2010). Sabermetrics: The past, the present, and the future. *Mathematics and Sports*, 3–14. doi: 10.5948/UPO9781614442004.002

Alcock, A. M. (2010). Analysis of direct shots at goal from free kicks in elite women ' s football Analysis of Direct Shots at Goal from Free Kicks. *Dissertation School of Health and Human Sciences, Southern Cross University, Lismore, Australia*.

Alexander, J. P., Spencer, B., Sweeting, A. J., Mara, J. K., & Robertson, S. (2019). The influence of match phase and field position on collective team behaviour in Australian Rules football. *Journal of Sports Sciences*, *37*(15), 1699–1707. Retrieved from https://doi.org/10.1080/02640414.2019.1586077 doi: 10.1080/02640414.2019.1586077

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*. Springer, Cham. Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-22475-2_1 doi: 10.1007/978-3-030-22475-2-1

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*(4), 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*(6), 1793–1839. Retrieved from https://link.springer.com/article/10.1007/s10618-017-0513-2 doi: 10.1007/s10618-017-0513-2

Anzer, G. (2021). *Large Scale Analysis Offensive Performance in Football—Using Synchronized Positional and Event Data to Quantify Offensive Actions, Tactics, and Strategies* (Unpublished doctoral dissertation). Eberhard Karls University Tübingen.

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. 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*. Retrieved from https://link.springer.com/article/10.1007/s10618-021-00810-3 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. Retrieved from https://doi.org/10.4324/9781003160953 doi: 10.4324/9781003160953

Araújo, D., Couceiro, M., Seifert, L., Sarmento, H., & Davids, K. (2021). *Artificial Intelligence in Sport Performance Analysis* (1st Editio ed.). New York: Routledge. Retrieved from https://doi.org/10.4324/9781003163589 doi: 10.4324/9781003163589

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, Barcelona (Spain)*, 1–13.

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

Atmosukarto, I., Ghanem, B., Ahuja, S., Ahuja, N., & Muthuswamy, K. (2013). Automatic recognition of offensive team formation in american football plays. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 991–998. doi: 10.1109/CVPRW.2013.144

Barris, S., & Button, C. (2008). A review of vision-based motion analysis in sport. *Sports Medicine*, *38*(12), 1025–1043. doi: 10.2165/00007256-200838120-00006

Bartlett, R., Button, C., Robins, M., Dutt-Mazumder, A., & Kennedy, G. (2012). Analysing team coordination patterns from player movement trajectories in soccer: Methodological considerations. *International Journal of Performance Analysis in Sport*, *12*(2), 398–424. doi: 10.1080/24748668.2012.11868607

Bate, R. (1987). Football Chance: Tactics and Strategy. *Science and Football: Proceedings of the first World Congress of Science and Football*, 13–17.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... Pascanu, R. (2018). *Relational inductive biases, deep learning, and graph networks.*

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. Retrieved from https://link.springer.com/article/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)—Categorisation of the defensive role for each player in a corner kick using positional data. *Journal of Quantitative Analysis in Sports (submitted).*

Baumer, B., & Zimbalist, A. (2014). *The Sabermetric Revolution*. University of Pennsylvania Press. doi: 10.9783/9780812209129

Beal, R., Chalkiadakis, G., Norman, T. J., & Ramchurn, S. D. (2020). Optimising game tactics for football. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *2020-May*(May), 141–149.

Beal, R., Norman, T. J., & Ramchurn, S. D. (2019). Artificial intelligence for team sports: A survey. *Knowledge Engineering Review*, *34*, 1–37. doi: 10.1017/S0269888919000225

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*(10), 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)*(January), 725–730. doi: 10.1109/ICDM.2014.133

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., & Matthews, I. (2015). Identifying team style in soccer using formations learned from spatiotemporal tracking data. *IEEE International Conference on Data Mining Workshops, ICDMW*(January), 9–14. doi: 10.1109/ICDMW.2014.167

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*(10), 845–852. doi: 10.1080/026404102320675675

Brefeld, U., Lasek, J., & Mair, S. (2019). Probabilistic movement models and zones of control. *Machine Learning*, *108*(1), 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). Developing a data-driven player ranking in soccer using predictive model weights. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 49–55. doi: 10.1145/2939672.2939695

Buchheit, M., Allen, A., Poon, T. K., Modonutti, M., Gregson, W., & Di Salvo, V. (2014). Integrating different tracking systems in football: multiple camera semi-automatic system, local position measurement and GPS technologies. *Journal of Sports Sciences*, *32*(20), 1844–1857. Retrieved from http://dx.doi.org/10.1080/02640414.2014.942687 doi: 10.1080/02640414.2014.942687

Bunker, R., Fujii, K., Hanada, H., & Takeuchi, I. (2020). Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union. (October). Retrieved from http://arxiv.org/abs/2010.15377 doi: 10.31236/osf.io/g2bj8

Cai, Y., Wu, S., Zhao, W., Li, Z., Wu, Z., & Ji, S. (2018). Concussion classification via deep learning using whole-brain white matter fiber strains. *PLoS ONE*, *13*(5), 1–21. doi: 10.1371/journal.pone.0197992

Camerino, O. F., Chaverri, J., Anguera, M. T., & Jonsson, G. K. (2012). Dynamics of the game in soccer: Detection of T-patterns. *European Journal of Sport Science*, *12*(3), 216–224. doi: 10.1080/17461391.2011.566362

Cardinale, M. (2006). Validation of Prozone ® : A new video- based performance analysis system. *International Journal of Performance Analysis in Sport*, *6*(1), 108–116.

Casal, C. A., Maneiro, R., Ardá, T., Losada, J. L., & Rial, A. (2015). Analysis of corner kick success in elite football. *International Journal of Performance Analysis in Sport*, *15*(2), 430–451. doi: 10.1080/24748668.2015.11868805

Chawla, S., Estephan, J., Gudmundsson, J., & Horton, M. (2017). Classification of passes in football matches using spatiotemporal data. *ACM Transactions on Spatial Algorithms and Systems*, *3*(2). doi: 10.1145/3105576

Chen, H. T., Chou, C. L., Fu, T. S., Lee, S. Y., & Lin, B. S. P. (2012). Recognizing tactic patterns in broadcast basketball video using player trajectory. *Journal of Visual Communication and Image Representation*, *23*(6), 932–947. Retrieved from http://dx.doi.org/10.1016/j.jvcir.2012.06.003 doi: 10.1016/j.jvcir.2012.06.003

Chen, S., Feng, Z., Lu, Q., Mahasseni, B., Fiez, T., Fern, A., & Todorovic, S. (2014). Play Type Recognition in Real-World Football Video. *IEEE Winter Conference on Applications of Computer Vision*, 652–659. doi: 10.1109/WACV.2014.6836040.

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

Cholaquidis, A., Fraiman, R., & Sued, M. (2020). On semi-supervised learning. *MIT Press*, *29*(4), 914–937. doi: 10.1007/s11749-019-00690-2

Courel-Ibáñez, J., McRobert, A. P., Toro, E. O., & Vélez, D. C. (2017). Collective behaviour in basketball: A systematic review. *International Journal of Performance Analysis in Sport*, *17*(1-2), 44–64. doi: 10.1080/24748668.2017.1303982

Decroos, T., Van Haaren, J., & 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

Desporto, J. (2009). Trends of tactical performance analysis in team sports: bridging the gap between research, training and competition. *Revista Portuguesa de Ciências do Desporto*, *9*(1), 81–89. Retrieved from http://www.scielo.mec.pt/scielo.php?pid=S1645-05232009000100008&script=sci_arttext&tlng=es

Dick, U., & Brefeld, U. (2019). Learning to Rate Player Positioning in Soccer. *Big Data*, *7*(1), 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*, 174. Retrieved from https://www.frontiersin.org/article/10.3389/fspor.2021.682986 doi: 10.3389/fspor.2021.682986

D'Orazio, T., & Leo, M. (2010, 8). A review of vision-based systems for soccer video analysis. *Pattern Recognition*, *43*(8), 2911–2926. doi: 10.1016/j.patcog.2010.03.009

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

Duarte, R., Araújo, D., Folgado, H., Esteves, P., Marques, P., & Davids, K. (2013). Capturing complex, non-linear team behaviours during competitive football performance. *Journal of Systems Science and Complexity*, *26*(1), 62–72. doi: 10.1007/s11424-013-2290-3

Dyk, D. A., & Meng, X. L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1–50. Retrieved from https://www.tandfonline.com/doi/abs/10.1198/10618600152418584 doi: 10.1198/10618600152418584

Ekin, A., Tekalp, A. M., & Mehrotra, R. (2003). Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, *12*(7), 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). Retrieved from https://www.frontiersin.org/articles/10.3389/fspor.2021.725431/full 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. (2021). *A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions* (Vol. 110) (No. 6). Springer US. Retrieved from https://doi.org/10.1007/s10994-021-05989-6 doi: 10.1007/s10994-021-05989-6

Fernando, T., Wei, X., Fookes, C., Sridharan, S., & Lucey, P. (2015). Discovering Methods of Scoring in Soccer Using Tracking Data. *KDD Workshop on Large-Scale Sports Analytics*, 1–4. Retrieved from https://large-scale-sports-analytics.org/Large-Scale-Sports-Analytics/Submissions2015_files/paperID19-Tharindu.pdf

Feuerhake, U. (2016). Recognition of Repetitive Movement Patterns—The Case of Football Analysis. *ISPRS International Journal of Geo-Information*, *5*(11), 208. doi: 10.3390/ijgi5110208

Fujii, K. (2021). Data-Driven Analysis for Understanding Team Sports Behaviors. *Journal of Robotics and Mechatronics*, *33*(3), 505–514. doi: 10.20965/jrm.2021.p0505

Fujimura, A., & Sugihara, K. (2005, 6). Geometric analysis and quantitative evaluation of sport teamwork. *Systems and Computers in Japan*, *36*(6), 49–58. doi: 10.1002/scj.20254

Games, M.-a. S. (2019). A Graph Attention Based Approach for Trajectory Prediction in Multi-agent Sports Games. *Preprint (arXiv)*.

Gentleman, R., & Carey, V. J. (2008). Unsupervised Machine Learning. In *Bioconductor case studies* (pp. 137–157). Springer, New York, NY. Retrieved from https://link.springer.com/chapter/10.1007/978-0-387-77240-0_10 doi: 10.1007/978-0-387-77240-0

Ghory, I. (2004). Reinforcement learning in board games. *Department of Computer Science, University of Bristol*, 1–57. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Reinforcement+learning+in+board+games+.#0

Goes, F., Schwarz, E., Elferink-Gemser, M., Lemmink, K., & Brink, M. (2021). A risk-reward assessment of passing decisions: comparison between positional roles using tracking data from professional men's soccer. *Science and Medicine in Football*, 1–9. Retrieved from https://doi.org/10.1080/24733938.2021.1944660 doi: 10.1080/24733938.2021.1944660

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*(5), 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*(1), 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*(0), 1–16. Retrieved from https://doi.org/10.1080/17461391.2020.1747552 doi: 10.1080/17461391.2020.1747552

Goldberg, X. (2009, 6). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*, 1–116. doi: 10.2200/S00196ED1V01Y200906AIM006

Gómez-Jordana, L. I., Milho, J., Ric, , Silva, R., & Passos, P. (2019). Landscapes of passing opportunities in Football – where they are and for how long are available ? *Barça sports analytics summit, Barcelona (Spain)*, 1–14.

González-Ródenas, J., López-Bondia, I., Aranda-Malavés, R., Tudela Desantes, A., Sanz-Ramírez, E., & Aranda Malaves, R. (2019). Technical, tactical and spatial indicators related to goal scoring in European elite soccer. *Journal of Human Sport and Exercise*, *15*(1), 186–201. doi: 10.14198/jhse.2020.151.17

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Gould, P., & Gatrell, A. (1979). A structural analysis of a game: The Liverpool v Manchester united cup final of 1977. *Social Networks*, *2*(3), 253–273. doi: 10.1016/0378-8733(79)90017-0

Gréhaigne, J. F., Bouthier, D., & David, B. (1997). Dynamic-system analysis of opponent relationships in collective actions in soccer. *Journal of Sports Sciences*, *15*(2), 137–149. doi: 10.1080/026404197367416

Gréhaigne, J. F., Godbout, P., & Bouthier, D. (1999). The foundations of tactics and strategy in team sports. *Journal of Teaching in Physical Education*, *18*(2), 159–174. doi: 10.1123/jtpe.18.2.159

Grunz, A., Memmert, D., & Perl, J. (2009). Analysis and simulation of actions in games by means of special self-organizing maps. *International Journal of Computer Science in Sport*, *8*(1), 22–37.

Grunz, A., Memmert, D., & Perl, J. (2012). Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human Movement Science*, *31*(2), 334–343. Retrieved from http://dx.doi.org/10.1016/j.humov.2011.02.008 doi: 10.1016/j.humov.2011.02.008

Gudmundsson, J., & Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys*, *50*(2), 1–34. doi: 10.1145/3054132

Gudmundsson, J., Laube, P., & Wolle, T. (2017). *Movement Patterns in Spatio-Temporal Data* (Z. X. Shekhar S. Xiong H., Ed.). Boston: Springer. doi: 10.1007/978-3-319-17885-1

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

Haaren, J. V., Zimmermann, A., Renkens, J., Broeck, G. V. D., Beéck, T. O. D., Meert, W., & Davis, J. (2013). *Machine Learning and Data Mining for Sports Analytics* (No. September). doi: 10.1007/978-3-030-64912-8

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier Inc. doi: 10.1016/C2009-0-61819-5

Hennessy, L., & Jeffreys, I. (2018). The current use of GPS, its potential, and limitations in soccer. *Strength and Conditioning Journal*, *40*(3), 83–94. doi: 10.1519/SSC.0000000000000386

Hernández-mendo, A. (2006). Hidden patterns of play interaction in soccer using SOF-CODER. *Behaviour Research Methods*(3), 372–381.

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*(6). doi: 10.1177/1747954119879350

Herold, M., Kempe, M., Bauer, P., & Meyer, T. (2021). Attacking key performance indicators in soccer: Current practice and perceptions from the elite to youth academy level. *Journal of Sports Science and Medicine*, *20*(1), 158–169. Retrieved from https://doi.org/10.52082/jssm.2021.158 doi: 10.52082/jssm.2021.158

Hirano, S., & Tsumoto, S. (2004). Finding interesting pass patterns from soccer game records. *Knowledge Discovery in Databases: PKDD 2004; Springer Berlin Heidelberg*, *3202*, 209–218. doi: 10.1007/978-3-540-30116-5{_}21

Hobbs, J., Power, P., Sha, L., Ruiz, H., & Lucey, P. (2018). Quantifying the Value of Transitions in Soccer via Spatiotemporal Trajectory Clustering. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1–11.

Hochreiter, S., & Schmidhuber, J. (1996). Long Short-Term Memory. *Hardware-Software Co-Synthesis of Distributed Embedded Systems*, *9*(8), 13–39. doi: 10.1007/978-1-4757-5388-2

Hochstedler, J., & Gagnon, P. T. (2017). American Football Route Identification Using Supervised Machine Learning. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1–11.

Hojo, M., Fujii, K., Inaba, Y., Motoyasu, Y., & Kawahara, Y. (2018). Automatically recognizing strategic cooperative behaviors in various situations of a team sport. *PLoS ONE*, *13*(12), 1–15. doi: 10.1371/journal.pone.0209247

Höner, O. (2008). Basiert die Sportwissenschaft auf unterschiedlichen „Sorten" von Theorien? *Sportwissenschaft*, *38*(1), 3–23. doi: 10.1007/bf03356066

Hosp, B., Schultz, F., Kasneci, E., & Höner, O. (2021). Expertise Classification of Soccer Goalkeepers in Highly Dynamic Decision Tasks: A Deep Learning Approach for Temporal and Spatial Feature Recognition of Fixation Image Patch Sequences. *Frontiers in Sports and Active Living*, *3*(July). doi: 10.3389/fspor.2021.692526

Hosp, B. W., Schultz, F., Honer, O., & Kasneci, E. (2021). Soccer goalkeeper expertise identification based on eye movements. *PLoS ONE*, *16*(5 May), 1–22. doi: 10.1371/journal.pone.0251070

Hughes, M., & Lovell, T. (2019). Transition to attack in elite soccer. *Journal of Human Sport and Exercise*, *14*(1), 236–253. doi: 10.14198/jhse.2019.141.20

Intille, S. S., & Bobick, A. F. (1999). Framework for recognizing multi-agent action from visual evidence. *Proceedings of the National Conference on Artificial Intelligence*, 518–525.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science (Special Section on Artificial Intelligence)*, *349*(6245), 255–260.

Karsten, B., Baker, J., Naclerio, F., Klose, A., Antonino, B., & Nimmerichter, A. (2017). Validity of a Microsensor-Based Algorithm for Detecting Scrum Events in Rugby Union. *International Journal pf Sports Physiology and Performance*, *14*(2), 156-162. Retrieved from https://www.cochranelibrary.com/central/doi/10.1002/central/CN-01787161/full

Kelly, D., Coughlan, G. F., Green, B. S., & Caulfield, B. (2012). Automatic detection of collisions in elite level rugby union using a wearable sensing device. *Sports Engineering*, *15*(2), 81–92. doi: 10.1007/s12283-012-0088-5

Kempe, M., Grunz, A., & Memmert, D. (2015). Detecting tactical patterns in basketball: Comparison of merge self-organising maps and dynamic controlled neural networks. *European Journal of Sport Science*, *15*(4), 249–255. Retrieved from http://dx.doi.org/10.1080/17461391.2014.933882 doi: 10.1080/17461391.2014.933882

Kempe, M., Vogelbein, M., Memmert, D., & Nopp, S. (2014). Possession vs. Direct Play: Evaluating Tactical Behavior in Elite Soccer. *International Journal of Sports Science*, *4*(6A), 35–41. doi: 10.5923/s.sports.201401.05

Kipf, T., Fetaya, E., Wang, K. C., Welling, M., & Zemel, R. (2018). Neural relational inference for Interacting systems. *35th International Conference on Machine Learning, ICML 2018*, *6*, 4209–4225.

Knauf, K. (2014). Spatio-Temporal Convolution Kernels for Clustering Trajectories. *KDD Workshop on Large-Scale Sports Analytics, New York (USA)*.

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*(5), 298–312. doi: 10.4304/jmm.4.5.298-312

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems*, *25*, 1097–1105. doi: 10.1201/9781420010749

Kuper, S. (2018). *Soccernomics: Why England Loses; Why Germany, Spain, and France Win; and Why One Day Japan, Iraq, and the United States Will*

Become Kings of the World's Most Popular Sport. Nation Books. Retrieved from https://books.google.com/books/about/Soccernomics_2018_World_Cup_Edition.html?hl=de&id=OqS-swEACAAJ

Laird, P., & Waters, L. (2008). Eyewitness Recollection of Sport Coaches. *International Journal of Performance Analysis in Sport*, *8*(1), 76–84. doi: 10.1080/24748668.2008.11868424

Le, H. M., Yue, Y., Carr, P., & Lucey, P. (2017). Coordinated multi-agent imitation learning. In *34th international conference on machine learning, icml 2017* (Vol. 4, pp. 3140–3152).

Li, R., & Chellappa, R. (2010). Group motion segmentation using a spatio-temporal driving force model. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2038–2045. doi: 10.1109/CVPR.2010.5539880

Li, R., Chellappa, R., & Zhou, S. K. (2009). Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, *2009 IEEE*, 2450–2457. doi: 10.1109/CVPRW.2009.5206676

Link, D. (2018). Sports Analytics: How (commercial) sports data create new opportunities for sports science. *German Journal of Exercise and Sport Research*, *48*(1), 13–25. doi: 10.1007/s12662-017-0487-7

Link, D., & Hoernig, M. (2017). Individual ball possession in soccer. *PLoS ONE*, *12*(7), 1–15. doi: 10.1371/journal.pone.0179953

Linke, D., Link, D., & Lames, M. (2018). Validation of electronic performance and tracking systems EPTS under field conditions. *PLoS ONE*, *13*(7), 1–20. doi: 10.1371/journal.pone.0199519

Linke, D., Link, D., & Lames, M. (2020). Football-specific validity of TRACAB's optical video tracking systems. *PLoS ONE*, *15*(3), 1–17. doi: 10.1371/journal.pone.0230179

Low, B., Coutinho, D., Gonçalves, B., Rein, R., Memmert, D., & Sampaio, J. (2020). *A Systematic Review of Collective Tactical Behaviours in Football Using Positional Data* (Vol. 50) (No. 2). Springer International Publishing. Retrieved from https://doi.org/10.1007/s40279-019-01194-7 doi: 10.1007/s40279-019-01194-7

Lucey, P., Bialkowski, A., Carr, P., Morgan, S., Matthews, I., & Sheikh, Y. (2013). Representing and discovering adversarial team behaviors using player roles. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2706–2713. doi: 10.1109/

CVPR.2013.349

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

MacLennan, T. (2005, 5). Moneyball: The Art of Winning an Unfair Game. *The Journal of Popular Culture*, *38*(4), 780–781. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/j.0022-3840.2005.140_11.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.0022-3840.2005.140_11.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.0022-3840.2005.140_11.x doi: 10.1111/j.0022-3840.2005.140-11

Manafifard, M., Ebadi, H., & Abrishami Moghaddam, H. (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

Mcintyre, A., Brooks, J., Guttag, J., Wiens, J., & Arbor, A. (2016). Recognizing and Analyzing Ball Screen Defense in the NBA Learning to Classify Defensive Schemes. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1–10.

McNeff, J. G. (2002, 3). The global positioning system. *IEEE Transactions on Microwave Theory and Techniques*, *50*(3), 645–652. doi: 10.1109/22.989949

Mcqueen, A., Wiens, J., & Guttag, J. (2014). Automatically Recognizing on-Ball Screens. *MIT Sloan Sports Analytics Conference, Boston (USA)*, 1–10. Retrieved from https://pdfs.semanticscholar.org/6334/05a9abc6c9ac20045365a3f47192f8fa6544.pdf?_ga=2.111994950.378455875.1556236357-933593754.1556236357

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

Mitrotasios, M., & Armatas, V. (2012). Analysis of goal scoring patterns in the 2012 European Football Championship. *The Sport Journal*(50), 1–9.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. *Arxiv Preprint*, 1–9. Retrieved from http://arxiv.org/abs/1312.5602

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2014). Foundations in Machine learning. In *Springerbriefs in computer science* (Vol. 0, pp. 39–44).

Montoliu, R., Martín-Félez, R., Torres-Sospedra, J., & Martínez-Usó, A. (2015). Team activity recognition in Association Football using a Bag-of-Words-based method. *Human Movement Science*, *41*, 165–178. doi: 10.1016/j.humov.2015.03.007

Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, *5*(4), 213–222. Retrieved from https://doi.org/10.1007/s41060-017-0093-7 doi: 10.1007/s41060-017-0093-7

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*(1), 59–72. doi: 10.1007/s10044-011-0238-6

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 '*(September), 25–33. doi: 10.1145/3347318.3355527

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. doi: 10.1186/s40537-014-0007-7

Narizuka, T., & Yamazaki, Y. (2019). Clustering algorithm for formations in football games. *Scientific Reports*, *9*(1), 1–8. Retrieved from http://dx.doi.org/10.1038/s41598-019-48623-1 doi: 10.1038/s41598-019-48623-1

Niu, Z., Gao, X., & Tian, Q. (2012). Tactic analysis based on real-world ball trajectory in soccer video. *Pattern Recognition*, *45*(5), 1937–1947. Retrieved from http://dx.doi.org/10.1016/j.patcog.2011.10.023 doi: 10.1016/j.patcog.2011.10.023

Njororai, W. W. (2013). Analysis of goals scored in the 2010 world cup soccer tournament held in South Africa. *Journal of Physical Education and Sport*, *13*(1), 6–13. doi: 10.7752/jpes.2013.01002

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604–624. doi: 10.1109/TNNLS.2020.2979670

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), 1–15. Retrieved from http://dx.doi.org/10.1038/s41597-019-0247-7 doi: 10.1038/s41597-019-0247-7

Perin, C., Vuillemot, R., & Fekete, J. D. (2013). SoccerStories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2506–2515. doi: 10.1109/TVCG.2013.192

Perše, M., Kristan, M., Kovačič, S., Vučkovič, G., & Perš, J. (2009). A trajectory-based analysis of coordinated team activity in a basketball game. *Computer Vision and Image Understanding*, *113*(5), 612–621. doi: 10.1016/j.cviu.2008.03.001

Perse, M., Kristan, M., Perš, J., & Kovacic, S. (2006). A Template-Based Multi-Player Action Recognition of the Basketball Game. *CVBASE '06 - Proceedings of ECCV Workshop on Computer Vision*, 71–82.

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

Pfeiffer, M., & Perl, J. (2015). Analysis of tactical defensive behavior in team handball by means of artificial neural networks. *IFAC-PapersOnLine*, *28*(1), 784–785. doi: 10.1016/j.ifacol.2015.05.169

Pisner, D. A., & Schnyer, D. M. (2019). Support vector machine. *Machine Learning: Methods and Applications to Brain Disorders*, 101–121. doi: 10.1016/B978-0-12-815739-8.00006-7

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), 1–038.

Pons, E., García-Calvo, T., Resta, R., Blanco, H., del Campo, R. L., García, J. D.,

& Pulido, J. J. (2019). A comparison of a GPS device and a multi-camera video technology during official soccer matches: Agreement between systems. *PLoS ONE*, *14*(8), 1–12. doi: 10.1371/journal.pone.0220729

Pouyanfar, S., & Chen, S.-C. (2016). Semantic Event Detection Using Ensemble Deep Learning. *IEEE International Symposium on Multimedia (ISM), San Jose (USA)*, 203–208. doi: 10.1109/ism.2016.0048

Power, P., Hobbs, J., Ruiz, H., Wei, X., & Lucey, P. (2018). Mythbusting Set-Pieces in Soccer. *MIT Sloan Sports Analytics Conference, Boston*, *102*(2), 1–12.

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

Pulling, C. (2015). Long Corner Kicks in the English Premier League. *Kinesiology*, *47*(2), 193–201.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, *11*(3), 269–282. doi: 10.14778/3157794.3157797

Reep, C., & Benjamin, B. (1968). Skill and Chance in Association Football Author. *Journal of the Royal Statistical Society*, *131*(4), 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*(1). doi: 10.1186/s40064-016-3108-2

Rein, R., Raabe, D., & Memmert, D. (2017). "Which pass is better?" Novel approaches to assess passing effectiveness in elite soccer. *Human Movement Science*, *55*, 172–181. Retrieved from http://dx.doi.org/10.1016/j.humov.2017.07.010 doi: 10.1016/j.humov.2017.07.010

Rennie, M. J., Kelly, S. J., Bush, S., Spurrs, R. W., Austin, D. J., & Watsford, M. L. (2020). Phases of match-play in professional Australian Football: Distribution of physical and technical performance. *Journal of Sports Sciences*, *38*(14), 1682–1689. Retrieved from https://doi.org/10.1080/02640414.2020.1754726 doi: 10.1080/02640414.2020.1754726

Ric, A., Bradley, P., Shaw, L., Thies, H., Sumpter, D., López-felip, M. A., ... Bauer, P. (2021). Football Analytics 2021: The role of context in transferring analytics to the pitch. *Barça sports analytics summit, Barcelona (Spain)*.

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*(March 2018), 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, Springer, Cham*, *1324*, 17–27. doi: 10.1007/978-3-030-64912-8

Sacha, D., Al-Masoudi, F., Stein, M., Schreck, T., Keim, D. A., Andrienko, G., & Janetzko, H. (2017). Dynamic Visual Abstraction of Soccer Movement. *Computer Graphics Forum*, *36*(3), 305–315. doi: 10.1111/cgf.13189

Sampaio, J., & Maças, V. (2012). Measuring tactical behaviour in football. *International Journal of Sports Medicine*, *33*(5), 395–401. doi: 10.1055/s-0031-1301320

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

Santos, A. B., Theron, R., Losada, A., Sampaio, J. E., & Lago-Peñas, C. (2018). Data-driven visual performance analysis in soccer: An exploratory prototype. *Frontiers in Psychology*, *9*, 1–12. Retrieved from https://www.frontiersin.org/article/10.3389/fpsyg.2018.02416 doi: 10.3389/fpsyg.2018.02416

Sarmento, H., Clemente, F. M., Araújo, D., Davids, K., McRobert, A., & Figueiredo, A. (2018). What Performance Analysts Need to Know About Research Trends in Association Football (2012–2016): A Systematic Review. *Sports Medicine*, *48*(4), 799–836. doi: 10.1007/s40279-017-0836-6

Sarmento, H., Marcelino, R., Anguera, M. T., CampaniÇo, J., Matos, N., & LeitÃo, J. C. (2014). Match analysis in football: a systematic review. *Journal of Sports Sciences*, *32*(20), 1831–1843. doi: 10.1080/02640414.2014.898852

Sathyamorthy, D., Shafii, S., Amin, Z. F. M., Jusoh, A., & Ali, S. Z. (2015). Evaluation of the accuracy of global positioning system (GPS) speed measurement via GPS simulation. *Defence S and T Technical Bulletin*, *8*(2), 121–128.

Schmicker, R. H. (2013). An application of satscan to evaluate the spatial distribution of corner kick goals in major league soccer. *International Journal of Computer Science in Sport*, *12*(2), 70–79.

Schmidt, A. (2012). Movement pattern recognition in basketball free-throw shooting. *Human Movement Science*, *31*(2), 360–382. doi: 10.1016/j.humov

.2011.01.003

Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). *Sports Data Mining* (Integrated ed.; R. Sharda & S. Voß, Eds.). Springer US. Retrieved from https://www.springer.com/de/book/9781441967299 doi: 10.1007/978-1-4419-6730-5

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, Barcelona (Spain)*, 1–13.

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

Shinde, P. P., & Shah, S. (2018, 7). A Review of Machine Learning and Deep Learning Applications. In *Proceedings - 2018 4th international conference on computing, communication control and automation, iccubea 2018.* Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ICCUBEA.2018.8697857

Siddiquie, B., Yacoob, Y., & Davis, L. S. (2009). Recognizing Plays in American Football Videos. *Technical Report, University of Maryland*, *1*(1), 1–8. Retrieved from http://www.researchgate.net/publication/228519111_Recognizing_Plays_in_American_Football_Videos

Sing, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310–1315. Retrieved from https://ieeexplore.ieee.org/abstract/document/7724478

Spearman, W. (2018). Beyond Expected Goals. *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*(1), 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

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*(2), 140–149. doi: 10.1007/s12662-019-00579-0

Stephanos, D. K., Husari, G., Bennett, B. T., & Stephanos, E. (2021). Machine learning predictive analytics for player movement prediction in NBA: Applications, opportunities, and challenges. *Proceedings of the 2021 ACMSE Conference - ACMSE 2021: The Annual ACM Southeast Conference*, 2–8. doi: 10.1145/3409334.3452064

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)*.

Stracuzzi, D. J., Fern, A., Ali, K., Hess, R., Pinto, J., Li, N., ... Shapiro, D. (2011). An application of transfer to American football: From observation of raw video to control in a simulated environment. *AI Magazine*, *32*(2), 107–125. doi: 10.1609/aimag.v32i2.2336

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.

Taberner, M., O'Keefe, J., Flower, D., Phillips, J., Close, G., Cohen, D. D., ... Carling, C. (2020). Interchangeability of position tracking technologies; can we merge the data? *Science and Medicine in Football*, *4*(1), 76–81. Retrieved from https://doi.org/10.1080/24733938.2019.1634279 doi: 10.1080/24733938.2019.1634279

Taki, T., Hasegawa, J. i., & Fukumura, T. (1996). Development of motion analysis system for quantitative evaluation of teamwork in soccer games. *IEEE International Conference on Image Processing*, 815–818. doi: 10.1109/icip.1996.560865

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*(3), 237–244. doi: 10.1080/02640410903502774

Tian, C., De Silva, V., Caine, M., & Swanson, S. (2020). Use of machine learning to automate the identification of basketball strategies using whole team player tracking data. *Applied Sciences (Switzerland)*, *10*(1), 1–16. doi: 10.3390/app10010024

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*(2020), 41–88. doi: 10.1613/JAIR.1.12505

Van Haaren, J., Dzyuba, V., Hannosset, S., & Davis, J. (2015). Automatically discovering offensive patterns in soccer match data. *International Symposium on Intelligent Data Analysis*, *9385*, 286–297. doi: 10.1007/978-3-319-24465-5

Van Haaren, J., Shitrit, H. B., Davis, J., & Fua, P. (2016). Analyzing volleyball match data from the 2014 world championships using machine learning techniques. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17*, 627–634. doi: 10.1145/2939672.2939725

Vilar, L., Araújo, D., Davids, K., & Bar-Yam, Y. (2013). Science of winning soccer: Emergent pattern-forming dynamics in association football. *Journal of Systems Science and Complexity*, *26*(1), 73–84. doi: 10.1007/s11424-013-2286-z

Vilar, L., Araújo, D., Davids, K., & Travassos, B. (2012). Constraints on competitive performance of attacker-defender dyads in team sports. *Journal of Sports Sciences*, *30*(5), 459–469. doi: 10.1080/02640414.2011.627942

Vogelbein, M., Nopp, S., & Hökelmann, A. (2014). Defensive transition in soccer - are prompt possession regains a measure of success? A quantitative analysis of German Fußball-Bundesliga 2010/2011. *Journal of Sports Sciences*, *32*(11), 1076–1083. Retrieved from http://dx.doi.org/10.1080/02640414.2013.879671 doi: 10.1080/02640414.2013.879671

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

Wang, J., & Zhang, J. (2015). A win-win team formation problem based on the negotiation. *Engineering Applications of Artificial Intelligence*, *44*, 137–152. Retrieved from http://dx.doi.org/10.1016/j.engappai.2015.06.001 doi: 10.1016/j.engappai.2015.06.001

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*(April 2016), 2197–2206. doi: 10.1145/2783258.2788577

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

Wu, Y., Xie, X., Wang, J., Deng, D., Liang, H., Zhang, H., . . . Chen, W. (2019, 1). ForVizor: Visualizing Spatio-Temporal Team Formations in Soccer. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 65–75. doi: 10.1109/TVCG.2018.2865041

Yeh, R. A., Schwing, A. G., Huang, J., & Murphy, K. (2019). Diverse generation for multi-agent sports games. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 4605–4614. doi: 10.1109/CVPR.2019.00474

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, M., & Kudenko, D. (2010). Automated event recognition for football commentary generation. *International Journal of Gaming and Computer-Mediated Simulations*, *2*(4), 67–84. doi: 10.4018/jgcms.2010100105

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

A Appendix—Study I: Constructing Spaces and Times for Tactical Analysis in Football

In the following, we present Andrienko et al. (2019), an accepted manuscript of an article published by IEEE Transactions on Visualization and Computer Graphics on April 1st 2021, available online: https://ieeexplore.ieee.org/document/8894420

Constructing Spaces and Times for Tactical Analysis in Football

Gennady Andrienko, Natalia Andrienko, Gabriel Anzer, Pascal Bauer, Guido Budziak, Georg Fuchs, Dirk Hecker, Hendrik Weber, and Stefan Wrobel

Abstract—A possible objective in analyzing trajectories of multiple simultaneously moving objects, such as football players during a game, is to extract and understand the general patterns of coordinated movement in different classes of situations as they develop. For achieving this objective, we propose an approach that includes a combination of query techniques for flexible selection of episodes of situation development, a method for dynamic aggregation of data from selected groups of episodes, and a data structure for representing the aggregates that enables their exploration and use in further analysis. The aggregation, which is meant to abstract general movement patterns, involves construction of new time-homomorphic reference systems owing to iterative application of aggregation operators to a sequence of data selections. As similar patterns may occur at different spatial locations, we also propose constructing new spatial reference systems for aligning and matching movements irrespective of their absolute locations. The approach was tested in application to tracking data from two Bundesliga games of the 2018/2019 season. It enabled detection of interesting and meaningful general patterns of team behaviors in three classes of situations defined by football experts. The experts found the approach and the underlying concepts worth implementing in tools for football analysts.

Index Terms—Visual analytics, movement data, coordinated movement, sport analytics, football, soccer.

1 Introduction

Football (soccer) is an exciting sport that attracts millions of players and billions of spectators worldwide. Wikipedia explains the basics as: *"Football is a team sport played with a spherical ball between two teams of eleven players. ... The game is played on a rectangular field called a pitch with a goal at each end. The object of the game is to score by moving the ball beyond the goal line into the opposing goal"* [1]. Although it was sufficient for G.Lineker to use just a single sentence to fully define football as *"Twenty-two men chase a ball for 90 minutes and at the end, the Germans always win"*, in reality football is very complex. 22+ players, the ball and 3 referees move and act in coordination within the teams and in competition between the teams. The game is defined by voluminous rules, characterized by complex interactions, and requires specific skills and sophisticated tactics.

Team managers (coaches) define team tactics and select a plan for each game that needs to be carefully implemented by the players. For winning a game, a team needs (1) skilled players in excellent physical conditions and (2) sophisticated tactics intelligently defined by coaches, effectively taught to players, trained, and carefully implemented in the game. While training is covered well by sport science, tactical analysis is still challenging. Data-driven tactical analysis requires understanding of information hidden in large volumes of game tracking data that include frequently sampled positions of players and the ball and numerous game events such as goal shots and goals, passes, tackles, possession changes, substitutions, fouls etc.

Professional football attracts tremendous interest and therefore is supported by industry and huge investments into infrastructure, players and coaches. Recent progress in football data collection, processing and analysis [2], [3] created new opportunities for providing data-driven insights into the game and, eventually, supporting a variety of stakeholders including coaches, medical staff of clubs, players, scouts, leagues, journalists and general public. Professional clubs nowadays intensively hire data scientists and some major clubs already have their own data analysis departments [4], [5]. Several companies develop software for supporting data collection, processing and statistical analysis and provide services to clubs delivering data and analysis results, including visualizations, which are mainly of illustrative nature. Analytical visualizations and visual analytics at large are still seeking their way to this domain.

This paper results from joint research and co-authorship of a group involving visual analytics researchers, data scientists, and football experts. The research goal of the group was to find approaches to extracting and understanding the general patterns of the **team behaviors** and **dynamics of changes** in relation to **events** and **context** in different classes of game **situations** as these situations develop. In the context of the paper, the term 'situation' refers to a combination of circumstances in which players behave, and the term 'episode' refers to the situation development in which the circumstances dynamically change. The overall goal involves the following sub-goals:

- *Gennady Andrienko and Natalia Andrienko are with Fraunhofer IAIS and City, University of London. E-mail: gennady.andrienko@iais.fraunhofer.de*
- *Gabriel Anzer is with Sportec Solutions GmbH, Germany*
- *Pascal Bauer is with DFB Akademie, Germany*
- *Georg Fuchs and Dirk Hecker are with Fraunhofer IAIS, Germany*
- *Guido Budziak is with TU Eindhoven, The Netherlands*
- *Hendrik Weber is with DFL Deutsche Fussball Liga GmbH, Germany*
- *Stefan Wrobel is with Fraunhofer IAIS and University of Bonn, Germany*

1) enable selection of groups of situations with particular characteristics and extraction of data pieces reflecting the development of these situations;
2) derive general patterns of team behaviors from the extracted data pieces;
3) enable comparison of general patterns corresponding to different groups of situations.

To achieve these sub-goals, our group has developed a framework including techniques for (1) query and data extraction (filtering), (2) integration of extracted data pieces into aggregate structures, which may involve (2*) space transformation, and (3) visualization of the aggregates for interpretation, exploration, and comparison. These components of the framework are briefly described below.

1. Query and filtering. This component enables selection of time intervals containing game episodes with target characteristics, which may refer to occurrences of specific game events (e.g., shots, passes in a given direction at a certain distance, etc.) and attributes, such as speed or acceleration, of the ball, teams, and selected players. Once a set of target intervals is selected according to event- or attribute-based query conditions, it is possible to further select intervals positioned in time in a specific way in relation to the target intervals, e.g., starting at a given time distance before or after the beginning or the end of each target interval and having a specified duration. This enables exploration of what had happened before and after the target episodes and at different stages of their development.

2. Aggregation. Trajectory fragments extracted from the selected intervals are integrated into aggregate structures, where each structure represents the behavior of one moving object (a player, the ball, the mass center of a team, etc.) and consists of a sequence of generalized positions corresponding to a sequence of interactive selections of time intervals. A generalized position is an aggregate of the positions of the object extracted from all selected intervals. It is represented by a central point, which may be the mean, median, or medoid of the extracted subset of positions, and one or more convex hulls covering chosen fractions (e.g., 50 and 75%) of this subset. Each sequence of generalized positions is represented by a pseudo-trajectory in an abstract temporal domain where the sequence of time stamps corresponds to the sequence of selections. The resulting sets of pseudo-trajectories of all players and the ball provide a generalized representation of collective behaviors in all situations with particular properties.

2*. Space transformation. As an additional means of abstraction and generalization, this component allows putting together similar movements that might take place in different parts of the pitch. The core idea is to replace the positions of the moving objects in the physical space (i.e., on the pitch) by corresponding positions in artificially constructed spaces, such as a team space, which represents the relative placements of the players within a team, or an abstract space with dimensions corresponding to some attributes. Generalized positions and pseudo-trajectories can be constructed from positions in artificial spaces in the same way as from the original positions in the pitch space. The pitch space is good for analyzing the tactics of the team movement while the team space is good for seeing the relative arrangement of the players and how it changes depending on circumstances.

3. Pattern visualization. To enable perception and interpretation of collective behavior patterns by an analyst, a set of aggregates (i.e., pseudo-trajectories) generated from selected episodes is represented visually. For comparative analysis of sets of aggregates generated for different teams, kinds of situations, and games, the pseudo-trajectories are put in a common spatial domain (i.e., the pitch space, team space, or attribute space) and aligned with respect to their abstract times.

While the framework makes use of previously existing techniques and approaches, it also incorporates novel ideas, specifically:

- new primitives for temporal queries allowing specification of relative time intervals (Section 4.1);
- a novel way of aggregating movement data that is suitable for bringing together temporally disjoint data pieces (Section 4.3);
- a data structure for representing aggregated movement data that allows the aggregates to be visualized and explored similarly to trajectories (Section 4.3.2);
- glyphs showing usual relative positions of players in their teams and providing hints at their roles (Section 4.2 and Fig. 4).

We demonstrate the effectiveness of the proposed framework in several case studies using real data from two Bundesliga games [6], [7] of the season 2018-19.

The remainder of the paper has the following structure. Section 2 introduces the main concepts concerning football tactics, describes the collection, contents, and structure of data from a football game, and presents the research problem we have addressed. After an overview of the related work (Section 3), we present our approach and components of the analytical framework in Section 4 and describe how we have applied it to three complex scenarios of team tactics analysis (Section 5). Section 6 discusses the overall approach and outlines directions for further work.

2 Background

2.1 Football tactics in a nutshell

Football tactics depends on multiple factors: which team possesses the ball, in what part of the pitch the ball and the teams are located, and how the players are arranged within their teams and in relation to the opponents. When a team possesses the ball, it aims at scoring a goal by offensive actions, although in some rarely occurring cases (such as a lead close to the end of the game) it can be a teams' solely objective to stay in ball possession. When the ball is possessed by the opponents, a team aims at preventing a goal and performs defense. There is an intermediate **turnover** stage between offensive and defensive actions. After winning the ball, a team can either **counter-attack** or **safeguard and build up**. After losing the ball, a team can either **fall back and defend** or perform **counter-pressing**. In some situations, e.g. after fouls or if the ball goes out of the pitch, the game is interrupted and resumes through **set pieces**, such as corners, free kicks, throw-ins, goalie kicks, and penalties.

Players in football teams have different roles. An established term **formation** means a way how 10 outfield players in a team generally position themselves relative to their teammates. Formations typically consist of three or four rows of players and are described, respectively, by three or four numbers specifying how many players are in each row from the most defensive to the most forward [8]. For example, formation 4-3-3 means that the team has 4 defenders, 3 midfielders and 3 forwards, or strikers. In some formations, intermediate lines appear denoting attacking or defensive midfielders or so-called second forwards playing slightly behind their partner.

Formations usually differ significantly depending on the ball possession, so that each team has an offensive formation and a defensive formation. Schematic figures in media usually show only the offensive formations. First investigations on comparing offensive and defensive formations of the same teams were made in [9] where the average line-ups of two teams were shown both in and out of ball possession. When the ball possession changes, teams strive to arrange themselves as fast as possible into the respective opposite (offensive or defensive) formation. Generally, formations as a major component of football tactics are carefully studied in literature [10]. There exist manuals for coaches (e.g. [11]) and catalogues of offensive formations (e.g. [12]) enumerating possible attacking styles and suggesting efficient defense.

However, not only the chosen formations are important. Football is a highly dynamic game where the players not just take fixed relative positions but constantly move in a coordinated manner, which does not simply mean moving in parallel and thereby keeping the same arrangement. Both the arrangement of the players and their relative movements depend on multiple factors, including which team and for how long possesses the ball, where on the pitch the teams are located, what are the distances to the opponents, what events happened recently, what is the current score of the game, etc. For understanding teams' tactics and their efficiency, it is necessary to see the spatial arrangements of the players and the character and dynamics of their changes in response to game events and other circumstances. In today's practice, this is a very time-consuming process done largely by analysts watching game videos and synthesizing information by reasoning. Several recent research prototypes [13], [14], [15], [16] support this activity by extracting formations and their changes from the data. However, changes of formations and, more generally, changes of movement behaviors do not happen instantly. What is still missing and challenging in supporting game analysis is a possibility to analyze the process of change in the context of game events and situation characteristics. Our paper intends to fill this very important gap.

2.2 Data acquisition, content, and structure

Today, detailed data are collected for almost every football game in major professional leagues. Usually positional data are extracted from video recordings. For this purpose, stadiums are equipped with stationary installations of multiple cameras that record games from different viewpoints. Video analysis software is used for extracting time-stamped positions of the players, referees, and the ball from video footage, usually with a sampling rate of 10-25Hz. Additionally, game events are extracted from video and positional data. Event data include the positions and times of the events and annotations, i.e., attributes describing the event types, involved players, outcomes, etc. The event extraction and annotation is done partly manually, though there exist implementations that facilitate manual annotation using machine learning approaches. Major companies doing data acquisition and processing are ChyronHego [17], OPTA [18], STATS [19], SecondSpectrum [20] and Track160 [21]. Smaller companies (e.g., FootoVision [22]) develop lightweight solutions for extracting data from a single video.

A typical data set for one game consist of general information (date and location, playing teams, names of referees), information about the teams (list of players with their intended positions on the pitch, list of reserve players for substitutions), positional data (coordinates of the ball, players and, sometimes, referees in 2D x,y or 3D x,y,z space with time references) and events (what happened, when and where, with event-specific characteristics). In average, about 140,000 positions for the ball and each player are recorded during one game, roughly 3,500,000 positions in total. In addition to automatically recorded positions, about 1,500 events are annotated manually and then validated using computational methods. As any real-world data, football data require assessment of data quality, plausibility checking, and evaluation of the coverage in space, time, and the set of moving objects [23]. Particularly, it is necessary to make queries with allowing certain tolerance to potential mismatch of times in positional and event data.

In addition to data collection, commercial companies provide basic analytical and visualization services. A typical menu of provided visuals includes depiction of individual events (e.g., positions of fouls and tackles, geometries of passes and shots) and aggregated representations of players' positions on the pitch such as density heat maps. Both types of visuals can be filtered by players and times. However, possibilities for exploration by connecting different aspects are not available.

2.3 Problem statement

The formulation of the research goals comes from the football experts. Basically, their question was: How team tactics can be understood from data reflecting the movements of the individual players and the ball (i.e., their trajectories) and the events that occurred during a game? All partners communicated to clarify the concept of team tactics and, on this basis, define and refine the research goals.

A team tactic can be defined as a general pattern of collective behavior in a group of situations with particular properties. This definition requires further clarification of the concepts of situation properties, collective behavior, and general pattern. Situation properties can be specified in terms of various attributes: which team possesses the ball, how much time has elapsed since the possession change, which team is winning, where on the pitch is the ball and the majority of team players, etc. Collective behavior means relative positions, movements, and actions of the players with respect to their teammates and the opponents. A general pattern means a synoptic representation integrating multiple specific instances of collective behavior that were practiced in similar situations. The patterns differ depending on the situation properties. Hence, understanding of team tactics requires consideration of groups of situations with different properties.

Based on this refinement, the research goals presented in the introduction section were formulated.

3 Related work

3.1 Major approaches to football analytics

Several groups of researchers managed to get access to game tracking data and developed interesting research prototypes. Often the starting point was an adaptation of methods and tools developed for other purposes (e.g. animal tracking or transportation) for football data. A prominent example is the famous Soccermatics book by D.Sumpter [4] that builds on his research on collective animal behavior [24].

A review [25] observes the state of the art, considering the following high-level tasks: playing area subdivision, network techniques for team performance analysis, specific performance metrics, and application of data mining methods for labelling events, predicting future event types and locations, identifying team formations, plays and tactical group movement, and temporally segmenting the game. Some of the considered methods actively use visualization components and thus fall into visual analytics (VA) approaches. Another review [9] takes a different perspective, emphasizing the works with substantial involvement of visualization and identifying the following major approaches:

Analysis of **game events**. A representative example is SoccerStories [26], which summarizes game episodes using visual primitives for game events such as long ball, turning the ball, cross, corner, shot etc.

Analysis of **trajectories** and **trajectory attributes**. A series of works from the University of Konstanz proposed methods for clustering trajectories of players during game episodes [27] and segmenting the game, finding interesting game situations [28], [29] and plays of particular configurations [30], analyzing multiple attributes along trajectories [31] and computing features of team coordination [32].

Analysis of **team formations** and derived features of them. Several papers from Disney Research target at reconstructing team formations and player roles from positional data. The proposed methods identify the role of each player at each time moment allowing the analyst to trace short- and longer-term roles and detect role swaps between players. This approach allowed characterization of team styles in several games [13], [14]. After enumerating offensive and defensive configurations of players, paper [15] evaluates pairwise success statistics. ForVizor [16] uses the dynamics of detected formations for segmenting the game. Another approach for game segmentation is clustering of time moments based on features reflecting relative positions of players or other team configuration indicators [33].

Computation of football-specific **constructs**, such as interaction spaces [31], and **indicators** such as scoring chances, pass options [34], [35], and pressure degrees [9], followed by visual representation of these in spatio-temporal displays. An interesting development is including visualization of computed features directly in video frames [36].

Apart from the research prototypes, there are also commercial systems and services that support coaches and match analysts in their work with positional data. We are aware of four such tools: STATS Edge Viewer [37], parts of the SAP Sports One [38], Second Spectrum [20], and the online match analysis portal offered to the Bundesliga clubs by Sportec Solutions [39]. The functionality provided by these tools can be divided into three categories: calculation of various statistics, which are visualized in business graphics, search for specific game episodes in video records, and replaying of selected episodes augmented with visual representation of calculated features, such as control zones and pass opportunities. Hence, it is possible either to analyze the overall statistics at the level of a whole game or to explore details of individual episodes. Our research fills the gap between these two extremes by developing approaches to extracting general movement patterns from multiple episodes with some common properties. Importantly, it is not limited to computing numeric statistics from selected parts of a game, but it produces more complex spatio-temporal constructs representing movements.

3.2 Relevant visual analytics approaches beyond football

Different visual analytics approaches proposed for analyzing spatio-temporal and movement data [40] are relevant to football analysis, although some of them have been developed for specific application domains such as transportation [41].

Querying and filtering. The structure of movement data suggests possibilities for selection of subsets based on the identities of the moving objects and their attributes, as well as dynamic data items including locations, times, and movement attributes, such as speed and direction [40]. A query may involve a combination of multiple heterogeneous aspects; thus, Weaver [42] discusses interactive cross-filtering across multiple coordinated displays by direct manipulation in the displays. There exist special query devices for temporal sequences of attribute values (e.g. TimeSearcher [43]) and for sequences of events.

The kind of analysis our group aimed to support requires selection of *groups of time intervals* containing game episodes with particular characteristics. Database researchers long ago proposed time query primitives [44], [45] suitable for such purposes. Recently, similar ideas were implemented within an interactive visual analytics environment in a tool called TimeMask [46]. Our approach extends this work by increasing query flexibility, see Section 4.1.

Transformation of space and time provide additional perspectives for looking at movement data. It may be useful to treat selected pairs of numeric attributes as coordinates in an abstract space [47]. A polar coordinate system may be used in such a space if the movement directions or cyclic time attributes are involved [48]. Research on group movement [49] introduces the idea of a group space consisting of relative positions in respect to a central trajectory of the group. This idea was successfully applied for analyzing the distribution of pressure over team formations in football [9].

Analysis of multiple asynchronous trajectories can benefit from transforming the time stamps of the positions to relative time references within relevant time cycles or within the individual lifetimes of the trajectories, which brings them to a common temporal reference system and thus supports comparisons and finding general patterns [50]. In this work, the ideas of space and time transformation have been further extended, taking into account the new ways of time filtering, see Section 4.2.

Aggregation is one of the most important tools for spatial abstraction and simplification of massive movement data [51]. Review [25] suggests spatial aggregation over Cartesian or polar grids or hand-designed polygons that reflect specific functions of pitch regions. Aggregation results can be automatically re-calculated in response to changes of query conditions.

The existing approaches to aggregation of movement data produce the following major types of aggregates: density fields [9], [52], place-related attributes reflecting various statistics of the appearances of moving objects in the places [53], flows between places [51], and a central trajectory of a set of similar trajectories [54]. The former two types represent the presence of moving objects rather than their movement while the latter two represent the movement of a whole group but not the movements of its members. In previous works, relative positions of group members were represented by density distributions [49] or by their average positions [9], but their movements within the group were not reflected. Hence, none of the previously existing aggregation methods is well-suited for representing collective movement patterns. Therefore, our group has devised a novel way of aggregation producing a novel kind of aggregate – a set of pseudo-trajectories, see Section 4.3. Aggregation operations are combined with time filtering and time transformation and enable assessment of variability within aggregates.

3.3 What is missing

Among the earlier works dealing with collective movement, some focus on detection of occurrences of specific relationships between moving objects, such as close approach, others search for overall patterns of collective behavior. However, the former result in multiple disjoint data pieces that do not make a general picture while the latter, on the opposite, tend to overgeneralize by neglecting essential differences between situations. Football is a dynamic phenomenon with high variability of situations, therefore it is necessary to understand the dynamics of patterns and differentiate individual and collective behaviors depending on the situational context. Hence, there is a need to develop methods for identifying classes of situations, detecting patterns of coordinated movement in subsets of similar situations, and comparing patterns across different subsets.

Another important aspect is a possibility to see relationships between individual behaviors in the overall context of coordinated movement. There are two important aspects of these relationships: (1) the spatial arrangement of individuals within a group and (2) how the arrangement changes in response to different circumstances. Our work aimed at developing appropriate methods for satisfying these needs.

4 Approach

4.1 Temporal queries for episode selection

We use the term *situation* to denote a particular combination of circumstances that may take place in the course of a game. The circumstances may include the ball status (in play or out of play) and possession (one of the two teams), the absolute or relative spatial positions of the players and/or the ball, their movement characteristics, such as direction and speed, the events that are happening currently or have happened before, the relative time with respect to the game start and end, etc. These circumstances dynamically change during the game. We refer to a sequence of changes happening during a continuous time interval as *situation development* and to the corresponding time interval as an *episode* of situation development.

To explore and generalize the behaviors of the players and teams in certain situations, one needs to be able to select all episodes when such situations happened and developed. The selection requires appropriate query facilities for (A) specification of the situations in terms of the circumstances involved and (B) specification of the relative time intervals in which the situation development will be considered. For example, the circumstances may be "team A gains the ball possession when the ball is in the opponents' half of the pitch", and the relative time interval may be from one second before to five seconds after the situation has arisen. An earlier proposed interactive query tool called TimeMask [46] supports (A) but not (B). To support both, we propose the following extended set of query primitives:

(A) Specification of situations
Result: set of target time intervals $T_1, T_2, ..., T_N$, where $T_i = [t_i^{start}, t_i^{end}]$
- Query conditions
  - Attribute-based: selection of value intervals for numeric attributes and particular values for categorical attributes
  - Event-based: selection of particular event categories
- Condition modifier: logical NOT
- Minimal duration of a situation

(B) Specification of relative intervals
Result: set of relative time intervals $R_1, R_2, ..., R_N$, where $R_i = [r_i^{start}, r_i^{end}]$
- Relative interval start r_i^{start}: reference (t_i^{start} or t_i^{end}) and time shift $\pm\delta$, i.e.,
$r_i^{start} = t_i^{start} \pm \delta$ or $r_i^{start} = t_i^{end} \pm \delta$.
- Relative interval end r_i^{end}: reference (t_i^{start} or t_i^{end}) and time shift $\pm\Delta$, i.e.,
$r_i^{end} = t_i^{start} \pm \Delta$ or $r_i^{end} = t_i^{end} \pm \Delta$.
- Relative interval duration D and reference r_i^{start} or r_i^{end}, i.e.,
$r_i^{end} = r_i^{start} + D$ or $r_i^{start} = r_i^{end} - D$.

The primitives for relative interval specification allow this to be done in one of two ways: to set both the start and the end (one of the time shifts δ or Δ may be zero), or to set either the start or the end and the interval duration. Here are examples of possible specifications of relative intervals with respect to a target T:

- select X sec after T:
$[R_{start}, R_{end}] \leftarrow [T_{end}, T_{end} + X]$
- select initial X sec of T:
$[R_{start}, R_{end}] \leftarrow [T_{start}, T_{start} + X]$
- add X sec before and Y sec after T:

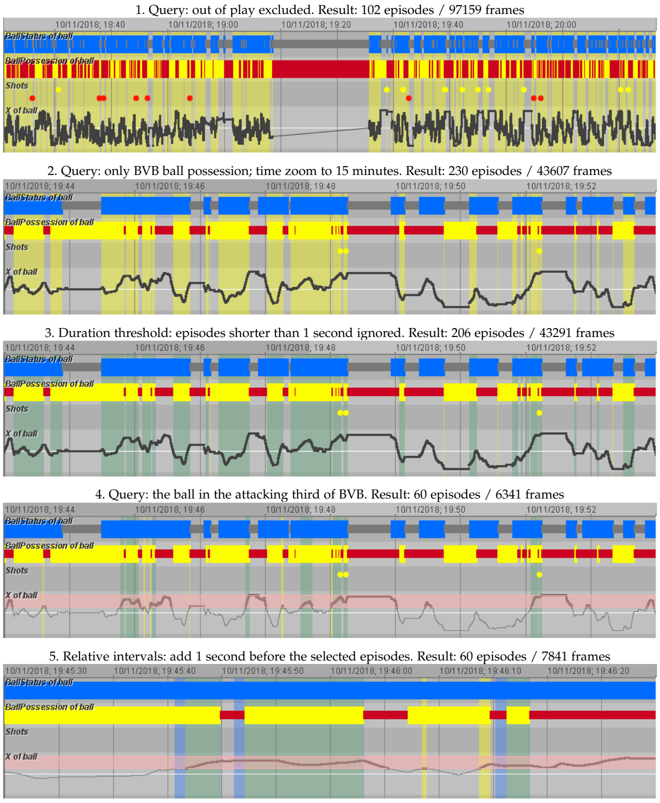

Fig. 1. A sequence of temporal query operations. Yellow vertical stripes mark time intervals selected by queries. Green is used for selected time intervals after ignoring short intervals. Blue shows interval extensions due to modifiers.

$$[R_{start}, R_{end}] \leftarrow [T_{start} - X, T_{end} + Y]$$

Figure 1 provides a visual illustration of a sequence of query operations for episode selection. Here and further on, we use data collected by a commercial service for the game of Borussia Dortmund and FC Bayern München [6], further called by the abbreviations BVB and FCB, respectively. BVB is usually shown in yellow and FCB in red. The data for this game span roughly over two hours (2 half times of 45+ minutes plus a break in between) with 25Hz resolution and include about 170,000 frames in total.

In the images shown in Fig. 1, the horizontal dimension represents time. In the vertical dimension, the images are divided into sections. Each section shows the variation of values of an attribute or a sequence of events. Categorical attributes are represented by segmented bars, the values being encoded in segment colors. Numeric attributes are represented by line charts. Events are represented by dots colored according to event categories. The yellow vertical stripes mark the target time intervals $T_1, T_2, ..., T_N$ selected according to the current situation specification. The blue vertical stripes mark the relative time intervals $R_1, R_2, ..., R_N$. The stripes are semi-transparent; so, the greenish color (a mixture of yellow and blue) appears where relative intervals overlap with target intervals.

The following sequence of query operations is shown: exclude the periods when the ball was out of play (Fig.1.1); select the episodes of BVB ball possession (Fig.1.2); exclude the episodes when BVB possessed the ball for less than 1 second (Fig.1.3); select the episodes in which the ball was in the attacking third of BVB (Fig.1.4); add 1-second intervals preceding the target situations (Fig.1.5).

Episode selection works as a temporal filter: only data from currently selected intervals $R_1, R_2, ..., R_N$ (or $T_1, T_2, ..., T_N$ if there is no relative interval specification) are treated as "active", being shown in visual displays and used in computational operations. This filter may be combined with various other filters applicable to movement data [40].

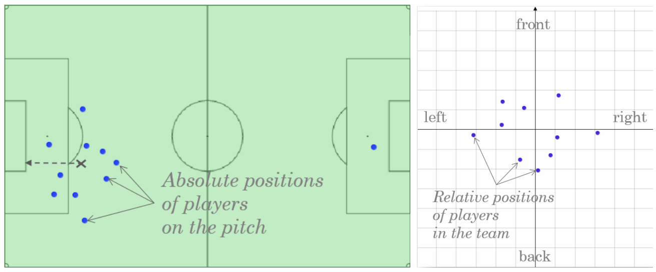

Fig. 2. A schematic illustration of a transformation from the pitch space (left) to the attacking team space (right). The coordinate grid in this and all further team spaces has 5m resolution.

4.2 Space transformation

The space transformation is based on determining the relative positions of points of all trajectories in respect to the corresponding point of a chosen or constructed reference trajectory and its movement vector [49]. Taking into account the nature of the football game, we usually assume the movement vector to be perpendicular to the opponent's goal line, see Fig. 2. This choice can be modified by, for example, treating differently situations when the players are very close to one of the goals, when teams/players often give up their preferred formations.

A reference trajectory may be chosen or constructed in different ways depending on the character of the collective movement and analysis goals. For our goals, none of the existing individual trajectories can be used as an adequate representative of the movement of a team as a whole. Instead, we generate a central trajectory of a team by applying an aggregation operator to the positions of all players of a team, excluding the goalkeeper, at each time moment. The operator may be the mean, median, or medoid (the medoid is the point having the smallest sum of distances to all others). We have extensively tested all three options using data from several games and found that the best is to take the team's mean after excluding the positions of two most outlying field players, i.e., the most distant from the mean of the whole team, excluding the goalkeeper. The central position computed in this way changes smoothly over time, while the medoid and median positions sometimes change abruptly, which leads to sharp kinks in the resulting central trajectory. Depending on the analysis task, it may be useful to compute the central trajectories of subgroups of players, such as the defenders or midfielders, and investigate the behaviors of these subgroups.

Figure 3 demonstrates how movements on the pitch translate to movements in a team space. It presents a short episode of a single attack of FCB on the pitch and in the BVB team space. The movements of the players and the ball are represented by lines, and the tiny square symbols at the line ends show the positions at the end of the selected time

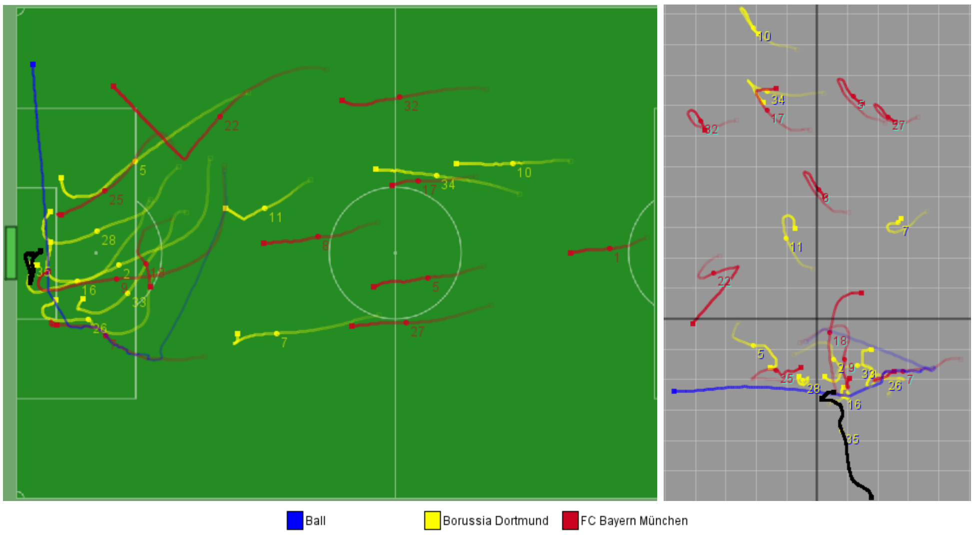

Fig. 3. One attack on the pitch (left) and in the BVB team space (right). In all illustrations, the rendering opacity varies along trajectories so that earlier segments are more transparent.

interval. The goalkeeper's trajectory is marked in black. The pitch map shows mostly parallel movements of all players towards the goal of BVB, but the player 22 of FCB, who initially ran in the direction of the goal, made a sharp turn to the right, and the players of BVB who were close to the goal or to the player 22 made similar movements. In the team space, the trajectories of 6 defence players of BVB located in the lower part of the team space are very short, which means high synchronization between them. In the upper part, the shapes of the trajectories belonging to the remaining 4 players of BVB and 5 players of FCB, show that the distance of these players from the team center originally increased (which means that they moved slower than the defenders) and then decreased (as the backward movement of the defenders slowed down). The diagonal orientation of these trajectories corresponds to the movement of the team center first to the left and then to the right.

While transformations to the team spaces are primarily meant for studying relative arrangements and movements of players, they create a useful by-product. By dividing a team space into meaningful zones and aggregating players' duration of presence in these zones, we obtain "fingerprints" of players' typical positions in the teams, which correspond to their roles in the game. We apply a division into a central zone (10m around the team center) and 8 areas around it. The fingerprints can be represented by *position glyphs*, as in Fig. 4. Thus, N.Süle in Fig.4, left, was present mostly on the back-left and back-center and sometimes in the central zone. Such glyphs facilitate identification of players and spotting their appearances in unusual positions and position swapping. Lines below some glyphs indicate the times the players were on the pitch (when it was not the full game) to give an idea of what changed after substitutions. Thus, N.Süle was a one-to-one substitution, which means that he exactly took M.Hummels' position after getting substituted for him. S.Wagner and R.Sanches were 1:1 substitutions to T.Müller and S.Gnabry, respectively, but interpreted their roles differently. This is visible from their aggregated positions in the team space and different distributions of the presence in their fingerprint glyphs. The display of position glyphs is optional. They were intensively used in our cases studies but are rarely present in the illustrations due to the length restrictions.

There is a possible additional use of the presence statistics by zones. The distance of a player to his average position in the team space can be treated as an indicator of how usual his current position is. These measures can be aggregated over the whole team or a selected group of players (e.g. the defence line). Based on the aggregated time series, it is possible to set query conditions for selecting episodes of unusual or usual team arrangements (Section 4.1). On the other hand, the presence statistics can be computed for different groups of selected episodes, and it is possible to choose which set of statistics to use for currently shown position glyphs and distance-to-usual-position calculations.

4.3 Aggregation

We propose a novel method for aggregating movements of an entity under different conditions. The output is a sequence of generalized positions organized in a pseudo-trajectory along an abstract timeline. Here we describe how we construct the positions and times of pseudo-trajectories.

4.3.1 *Obtaining generalized positions*

Aggregation is applied to positions selected by the current combination of data filters, including the episode selection (Section 4.1). It can be done in the pitch space and/or in the team spaces. The currently selected subset of positions of an entity is represented by a generalized position, which can be the mean, median, or medoid of the selected subset. The possibility to switch between these three options can serve as a means for checking the position variability. When the variability is small (i.e., the points are compactly clustered), the mean, median, and medoid positions are very close to each other, and switching between them does not change the general movement pattern obtained through the aggregation. Noticeable changes indicate presence of outliers. In this case, the analyst may look at the whole set of the original points and decide whether the outliers can be ignored.

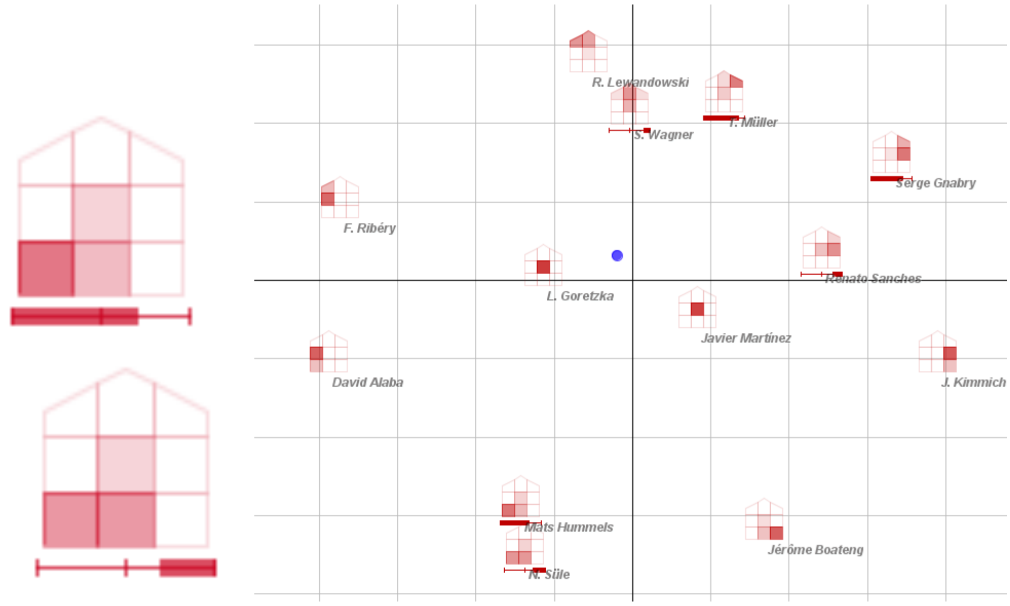

Fig. 4. Left: examples of position glyphs for N.Süle (below) who substituted M.Hummels (above). Right: glyphs shown at the average positions of the FCB players in their team space. The blue dot represents the average position of the ball.

This is possible when the outliers are few, their positions are randomly scattered, and the remaining points make a compact cluster. If the outliers are not negligible, they need to be examined in detail. To see the pattern formed by the remaining data, the analyst may switch to using the medoid, which is insensitive to outliers (but very costly to compute).

There is no one-size-fits-all rule for including or excluding outliers, but the decision may depend on specific contexts, such as temporary positional changes, different roles of the players in set plays, a change of the tactical system over time, etc. It is crucial that the soccer expert (match analyst, coach, assistant coach) is able to decide those things situation-specific and per interrogation of his/her own. Therefore, the user should be given a high degree of flexibility for investigations and the opportunities to exclude or include outliers and to change between more and less outlier-sensitive aggregates.

To represent the variation among the positions explicitly, we propose to build convex hulls outlining chosen percentages (e.g., 50 and 75%) of the set of original points ordered by their distances to the representative point. Examples can be seen in Fig.7 and Fig.9. Hence, a generalized position of an entity is a combination of a representative point and one or more *variation hulls*. Such a position is constructed for each entity being currently under analysis. The visual representation of the variation hulls can be controlled independently of that of the points; in particular, the hulls can be temporarily hidden for reducing the display clutter.

When the filter conditions change, aggregation can be applied to the new subset of selected positions, which produces a new set of generalized positions of all entities. The new generalized positions can be visualized together with the previous ones for comparing, as shown in Figs. 5 and 6. Figure 5 shows the aggregates in the pitch space, and Figure 6 shows the corresponding aggregates in the team space of BVB. In the left parts of the two figures, the first aggregation operation was performed using the episode filter with conditions $ballInPlay = true$ & $ballPossession = BVB$, the second one differs by $ballPossession = FCB$. To support the comparison, the corresponding points are connected by lines. The second points (FCB possession) are marked by dots. The differences and commonalities of players' arrangement depending on the ball possession can be easily seen. On the pitch, both teams move a bit towards the goal line. The most prominent behavioral differences happen with the wing defenders: they move wide under their own team's ball possession and narrow under the opponent's possession. In the team space, we see that the team without the ball gets more compact in both dimensions (all players tend to move towards the team center), while the team with the ball gets wider. The defenders of the attacking team move slower than the other players and thus increase team's covered area.

Another example is shown in the right parts of the Figs. 5 and 6. In the first half of the game, FCB scored a single goal at minute 26. We compare the mean positions of the players and the ball under the BVB possession before and after the goal (the latter are marked by dots), excluding the times when the ball was out of play. The pitch map shows us that both teams have shifted towards the FCB goal and a bit to the right. In the BVB team space, we see that all BVB players shifted synchronously except two central defenders (#2 and #16) who moved about 3m back from the team center. This increased the distance between the lines in the team and thus created opportunities for FCB counter-attacks. This fact may explain why the FCB player #9 moved about 5m aside in the BVB team space, searching for attacking opportunities.

According to the football experts, such representations can be very valuable to coaches by giving an overview of behavior changes after any significant game event. For seeing more than a single change and, in general, comparing generalized positions in more than two sets of situations, we construct abstract timelines for organizing multiple general-

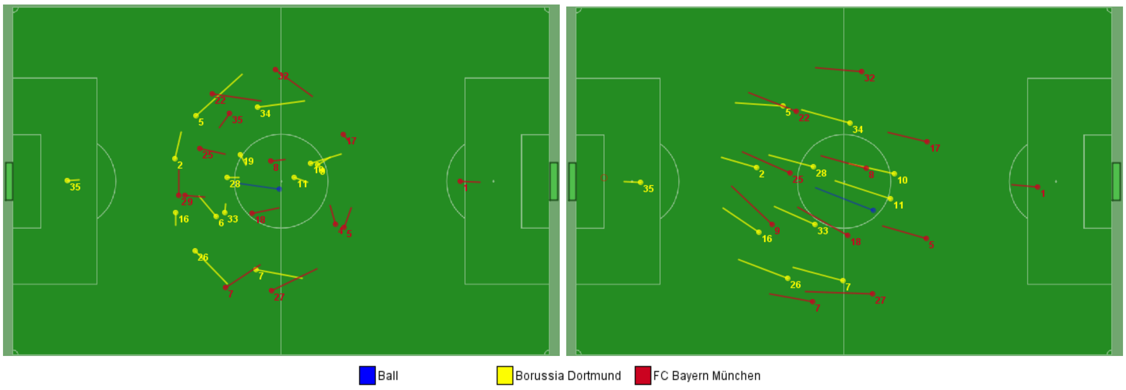

Fig. 5. Comparison of the generalized positions on the pitch in different groups of situations. Left: under the ball possession by the different teams; right: under the BVB possession before and after the goal in the first half of the game.

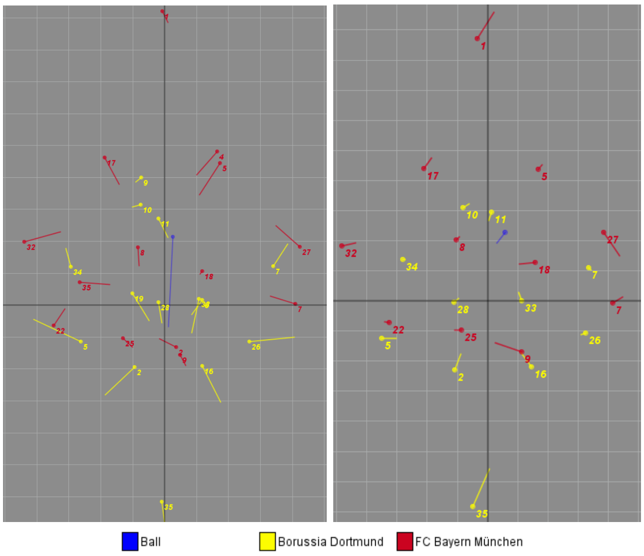

Fig. 6. Comparison of the generalized positions in the BVB team space for the same groups of situations as in Fig. 5.

ized positions in pseudo-trajectories.

4.3.2 Creating virtual times

A *pseudo-trajectory* consists of two or more linearly ordered generalized positions. A pseudo-trajectory is represented on a map by a line obtained by connecting consecutive positions (more precisely, their representative points). Each position of a pseudo-trajectory has an abstract numeric timestamp (1, 2, 3, ...) that equals the ordinal number of the position. Hence, a pseudo-trajectory has its internal abstract timeline made by the sequence of the position timestamps.

Pseudo-trajectories are generated by successively applying several query + aggregation operations. Each operation generates one position, which is appended after the previously generated position, if any. Hence, the order of the positions in pseudo-trajectories reflects the order of the query + aggregation operations by which they have been obtained. This very simple idea provides high flexibility for creating position sequences with different semantics, as demonstrated by an example below and further on in the case studies.

The queries used for generating pseudo-trajectories may include conditions of any kind, not necessarily time-based. Thus, the example in Fig. 7 demonstrates pseudo-trajectories of the players and the ball obtained by queries concerning the position of the BVB team center on the pitch. We made a sequence of 10 queries with a common condition $ballPossession = BVB$ and the differing conditions referring to the position of the BVB team center along the X-axis with respect to the pitch center: $x < -40m, -40m \leq x < -30m, ..., x \geq 40m$. The queries did not include explicit time constraints, but each query selected a set of time intervals when the x-coordinate of the team center was in a specific range.

Two upper images in Fig. 7 show the footprints of the pseudo-trajectories as lines on the pitch (left) and in the BVB team space (right). On top of the lines corresponding to the players, the position glyphs of the players are shown. The glyphs are drawn at the middle positions of the players' pseudo-trajectories, which correspond to the BVB team center position being in the interval $[-10..0)$ meters. The relative arrangement of the glyphs of each team reflect the formation used by BVB for preparing an attack and the defensive formation 4-4-2 of FCB. We can also see consistent monotonous changes of the player's positions from left to right and changes of the teams' widths along the pitch. Complementary to this, the team space demonstrates changes in the team compactness in both dimensions.

For illustrative purposes, the remaining images in Fig. 7 include the 50% variation hulls for selected 5 players of BVB. The hulls are shown in the pitch space and the BVB team space in 2D maps and 3D space-time cubes. The images of the space-time cubes are provided for merely illustrative purposes, to demonstrate that the pseudo-trajectories and the hulls are spatio-temporal objects, albeit the time domain in which they exist is abstract rather than real. These objects can be treated in the same ways as "normal" spatio-temporal objects existing in real time domains.

The colors of the variation hulls from violet through yellow to orange depict the virtual times of the positions. We can observe stable shapes and sizes of the hulls and their very stable locations in the team space, except for the two wing defenders who first moved outwards from the center and then back towards the center. Generally, such stability checks are necessary for all aggregates, and they were consistently performed during the analysis. However, the page limit does not permit having many such illustrations in the paper.

4.4 Interaction between the framework components

Since our paper aims at presenting the general framework, which can be implemented in different ways, rather than a specific implementation, we refrain from describing techniques for user-computer interaction, which can differ between possible implementations. What we describe here is how the components of the general framework are supposed to work together within the analysis process.

The process begins with creation of the team spaces (Section 4.2). Then, the following sequence of steps is repeatedly executed:

1. Temporal query (Section 4.1):
   1.1. Specify and find situations of interest (Section 4.1(A)).
   1.2. Specify a sequence of relative intervals for extracting the situation development episodes (Section 4.1(B)).
   Result: set of episodes.
2. Aggregation (Section 4.3):
   2.1. Automatically aggregate the query result and generate pseudo-trajectories in the pitch space and in the team spaces.
   2.2. Put the pseudo-trajectories as new information layers in the respective spaces.
   Result: set of pseudo-trajectories.
3. Visualization and comparative analysis:
   3.1. Represent the pseudo-trajectories within each space in an interactive visual display using techniques suitable for ordinary trajectories. The displays need to be linked through common visual encoding and by interactive techniques, such as brushing [55].
   3.2. Compare the new set of pseudo-trajectories with one or more of the previously obtained sets resulting from other temporal queries.

Our group has performed this analytical process in the case studies described in Section 5.

4.5 Evaluation of the framework

In our research project, it was not intended to design and develop software tools according to specific users' tasks and requirements. For developing and testing the components of the framework, the partners specializing in visual analytics and data science, further referred to as *the analysts*, utilized existing software tools. Among others, they used a research-oriented software system V-Analytics (http://geoanalytics.net/V-Analytics/). The analysts extended its base functionality by implementing new query, aggregation, and data transformation techniques. These software developments were necessary for achieving the research goals, but the key result of the project is the analysis methods and not the tools.

In the course of the research, the methods under development were constantly evaluated by the football domain experts according to the following criteria: (A) the possibility to select multiple situations with common properties and the flexibility in specifying the properties of interest; (B) the possibility to extract, visualize, and interpret general patterns of situation development. To assess the query facilities (A), the experts described the situations that were interesting for them, and the analysts translated the descriptions into queries, extracted the corresponding portions of the data, and provided the experts with tools for sampling some of the selected situations and reviewing them with the use of animated maps and corresponding fragments of the game video. To assess the situation generalization facilities (B), the experts were provided with visual displays of the pseudo-trajectories, which represented the extracted patterns.

The evaluation was carried out in a series of case studies, in which the experts set the analysis tasks, the analysts performed operations according to the framework, and the football experts interpreted and evaluated the results and posed further questions.

5 Case Studies

The objective of professional football is to win matches and entertain the public. To win matches, you have to score more goals than the opponent. This requires a good balance between the offensive and defensive strategies of a team. That is why it is important that a team has a tactical plan defining the desired team formations and behavior in different states of the game. Examples of states are own and opponent's ball possession, transitions between them after loosing the ball or recovering it, counter attacks, set pieces, such as corners etc. In the following, we consider several categories of situations that were interesting to analyze for our football experts. We use data from two Bundesliga games [6], [7] of the season 2018-19. The data from each game consist of two parts: (1) trajectories of the players and the ball, i.e., sequences of their positions in the pitch recorded every 40 milliseconds, and (2) records of the game events with their attributes, including the event type, time of occurrence, and players involved; see Section 2.2.

5.1 Ball possession change

The two switching moments between own and opponent's possession are getting more and more attention in football. The reason is that the desired field occupancy of the players and team tactics in situations where the team possesses the ball is completely different from the ideal field occupation and tactics in situations in which the opponent has the ball. As soon as a team loses the ball, the field occupation is often disorganized from a defensive perspective. It takes time for a team to adapt to the new situation ('switching cost') in which it has to apply its defensive tactics. This temporary 'chaos' is something the team with ball possession can take advantage of.

5.1.1 Checking the five-seconds rule

The famous P.Guardiola's 5 seconds rule for successful counterpressing says [56]: *"After losing the ball, the team has five seconds to retrieve the ball, or, if unsuccessful, tactically foul their opponent and fall back"*, that has been key to Manchester City conceding considerably fewer goals. The football experts were interested to see if BVB and FCB applied this tactics in the game [6].

To extract the episodes of interest, the analysts applied temporal query operations described in Section 4.1. They

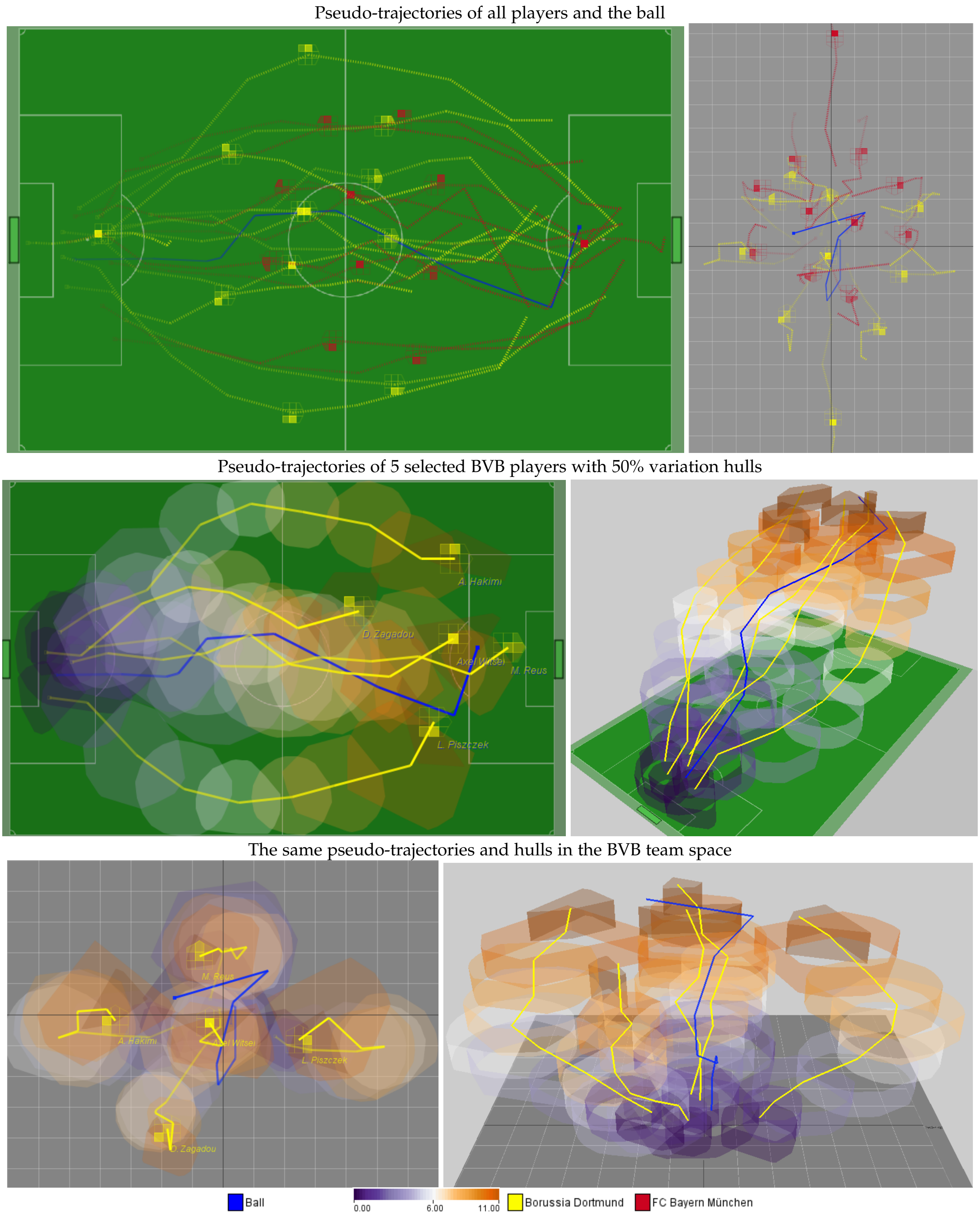

Fig. 7. Sequences of 10 generalized positions of the players and the ball under the BVB ball possession corresponding to different positions of the BVB's team center on the pitch along the X-axis.

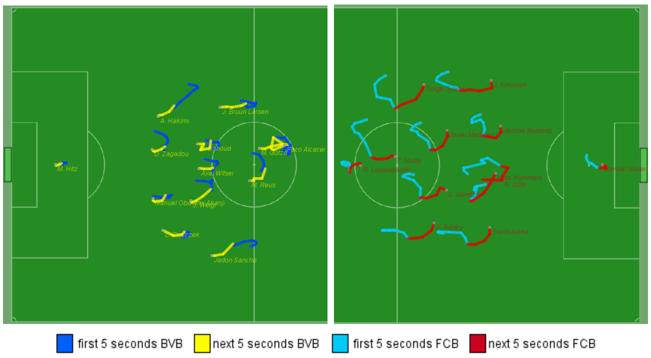

Fig. 8. Pseudo-trajectories of the players during 5+5 seconds after loosing the ball. Left: BVB, blue for the starting 5 sec and yellow for the following. Right: FCB, in cyan and red, respectively.

first selected the moments of change in the ball possession, excluding those when the ball got out of play. Next, they consecutively applied a series of operations for specification of relative intervals:
$[R_{start}, R_{end}] \leftarrow [T_{end} + X sec, T_{end} + (X + 1)\ sec]$,
$X = 0, 1, 2, ..., 9$
producing 10-steps long pseudo-trajectories from the mean positions and their 50% variation hulls. During the aggregation, irrelevant parts of the original trajectories that were shorter than 10 seconds (e.g. due to the ball going out of play or another change of possession) were discarded by an attribute-based query.

Figure 8 shows the results of the aggregation separately for BVB (left) and FCB (right). The pseudo-trajectories of the players are painted in two contrasting colors corresponding to the first 5 seconds (blue for BVB and cyan for FCB) and to the following 5 seconds (yellow for BVB and red for FCB). The images show that almost all BVB players do not move back during the first 5 seconds and gradually fall back in the next 5 seconds. This agrees with the fact that BVB is known for their pressing style of playing. The patterns of the FCB players are different. The players continue moving forward for the initial 2 seconds on the average and then start moving to the left and back.

Figure 9 shows the players' pseudo-trajectories and 50% variation hulls in the team spaces. The colors of the hulls encode their relative times. For BVB, the hull colors vary from dark blue to dark yellow, so that the shades of blue correspond to the first 5 seconds and the shades of yellow to the following 5 seconds. For FCB, the hull colors vary from dark cyan to dark red, respectively. The images show that BVB tended to reduce the team width and depth whereas FCB kept the width constant while slightly reducing the depth. The stacks of the hulls with the colors representing the relative times show that the hulls of the majority of the players were getting notably smaller over time. This means that the players were successfully reconstructing their planned defensive formations and then were keeping their intended relative positions. This is in accordance to the general football philosophy that suggests creativity while attacking and organization and structure while defending.

5.1.2 *Comparison of behaviors in two games*

For experts it is important to understand how a team adapts to different opponents and circumstances. To contrast the

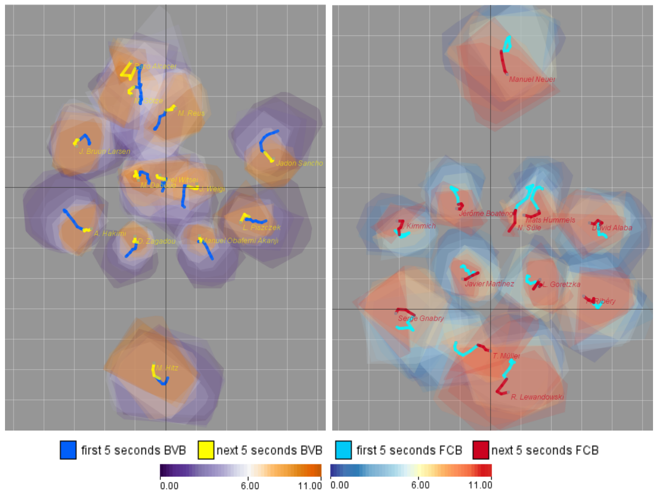

Fig. 9. The same aggregated data as in Fig.8 are presented in the team spaces (left: BVB, right: FCB) using the same colors together with the 50% variation hulls.

game 3:2 FC Bayern [6], they want to compare BVB's behavior with their 7:0 game against FC Nürnberg (FCN) [7]. Following the procedure described in Section 5.1.1, in each game for each player the analysts constructed two 10-steps long pseudo-trajectories summarizing the transitions to the attacking and defensive formations (Figs. 10 and 11).

The images on the top of Fig. 10 and on the left of Fig. 11 correspond to the game against FC Nürnberg, the other two images correspond to the game against FCB considered earlier. The BVB team tactics when loosing and regaining the ball are represented in black and yellow, respectively. The pitch images show that, while in the game against FCB (right) BVB players pressured for 5 seconds after loosing the ball, in the game against FCN they were falling back during the initial 5 seconds and only then put pressure on the opponents. The team was narrowing down in the FCB game but kept constant width in the FCN game. After ball regaining, BVB players attempted to perform fast counter attacks against FCB but preferred careful, slowly moving forward attack preparations against FCN.

The team space images are useful for seeing the synchronization among the players and changes of their arrangement. Game against FCN (left): synchronous movement in the transition to defense, except the central forwards, and widening of the team in the transition to offense. Game against FCB: increasing compactness of the team for defense and fast expansion in both direction for offense. The team depth in the game against FCN was about 30m in contrast to about 40m against FCB.

5.2 Long passes

One of the most important instruments for advancing the attack, finding open spaces on the pitch, and forcing opponents to make mistakes, is long passes. After examining the distribution of the pass lengths with respect to the X-axis (i.e., how much the ball moved in the direction to the opponents' goal) in the game [6], the experts and analysts jointly chose a query threshold of minimum 20m for selecting long passes. After inspecting the selected passes, the experts understood that the passes made from the own goal

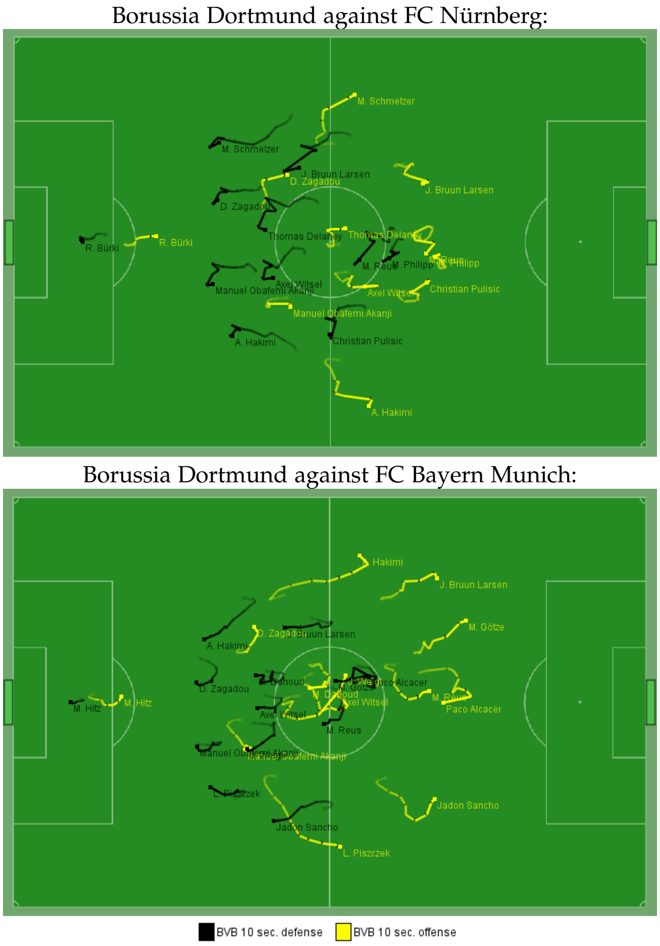

Fig. 10. Comparison of movements of Borussia Dortmund players in ball possession transition periods in games against FC Nürnberg [7] (top) and FC Bayern Munich [6] (bottom). Only players who were present on the pitch for at least 30 minutes are shown. Black lines show transitions to defense: from -1 to +10 seconds after loosing the ball control; yellow lines represent transitions to offense: from -1 to +10 seconds after gaining the ball. Different tactical patterns appear prominently (see explanations in section 5.1.2 for details).

box need to be excluded. Figure 12 shows the footprints of the remaining selected long balls. To produce them, the analysts used two queries based on the starting and ending moments of the selected passes:
$[R_{start}, R_{end}] \leftarrow [T_{start} - \delta, T_{start} + \delta]$ and
$[R_{start}, R_{end}] \leftarrow [T_{end} - \delta, T_{end} + \delta]$,
where $\delta = 0.2sec$ was applied for tolerating possible mismatches in the times between the position data and manually annotated passes. To include more information about the conditions in which these long balls were made, the analysts visualized the pass lines together with aggregated positions of all players (colored dots) on top of the density field summarizing the distributions of all players on the pitch during the execution of the selected passes.

An interesting diagonal configuration of the players during the long balls of BVB can be seen in Fig. 12, top. The passes were sent either along or across this dense concentration. Most of the long passes of both teams were directed to the left flank of BVB and right flank of FCB, so the same side of the pitch was used actively by both teams. These passes were very important in this game as they were involved in 3 out of 5 attacks that resulted in goal scoring. To consider them separately, the analysts added spatial filters by the pass destinations and obtained aggregates for the subset of the passes (Fig. 13).

It is interesting to observe that, although the pass targets were quite widely distributed on the pitch, they were compact in the spaces of the defending teams. All BVB passes were targeted in the area behind the FCB's right central defender J.Boateng, on the average 5 meters behind and 10 meters aside of him. He had to move back during these passes, breaking the last defensive line. The FCB passes targeted at a point about 25-30 meters aside of the BVB team center. These passes forced the defending team to shift left.

It can be concluded that the long forward passes of BVB were intended to make immediate danger to the goal. The attacking group of the BVB players moved far forward during these passes. The shape of the team became long but rather narrow. The long passes of FCB resulted in changes of the attacking direction with the players moving to the right. It should be noted that FCB striker R.Lewandowski was balancing around the offside line at the moment of the reception of the selected long passes, so it would be dangerous to pass to him immediately.

5.3 Building up for shots

Goals are scored after successful shots, which require not only high individual skills of a striker but also work of the whole team for reaching situations in which good shots become possible. To help the experts to investigate this teamwork in the BVB-FCB game, the analysts used queries to select, first, the moments of the shots and, second, the episodes preceding them. They made a series of queries
$[R_{start}, R_{end}] \leftarrow [T_{start} - X\ sec, T_{start} - (X-1)\ sec]$,
$X = 10, 9, 8, ..., 1$, where T_{start} is the moment of a shot, and obtained the corresponding sets of pseudo-trajectories of the ball, team centers, and all players separately for the shorts of BVB and FCB. As a measure of position variation, the analysts also computed the median distances of the representative points of the generalized positions to the original positions from which they had been derived. Since changes of the ball possession could happen during the 10-seconds intervals before the shots, the analysts applied attribute filters discarding irrelevant parts of the episodes in which the build up was shorter than 10 seconds.

The pseudo-trajectories of the players and the ball are shown in Fig. 14 and the pseudo-trajectories of the team centers in Fig. 15. These two figures demonstrate different levels of aggregation and abstraction applied to the same data. The position variation indicators associated with the points of the pseudo-trajectories are represented by proportional sizes of the dots thus marking 1-second segments. Please note that this is an abstract, symbolic representation using the visual variable 'size' to encode numeric values. The sizes of the symbols are not related to the map scale. This representation is essentially different from the representation of the variation by hulls, which are spatial objects. Unlike the dot sizes, the hulls occupy particular areas in space and have particular shapes.

The individual aggregates in Fig. 14 and team aggregates in Fig. 15 consistently show that the two teams tended to use

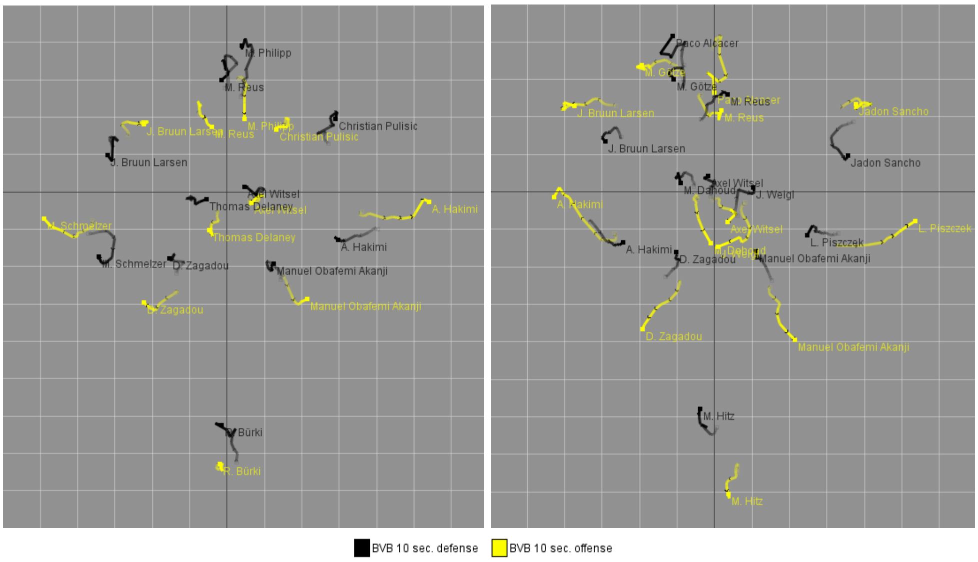

Fig. 11. Aggregated movements of BVB players in their team space after ball possession changes in games [7] (left) and [6] (right).

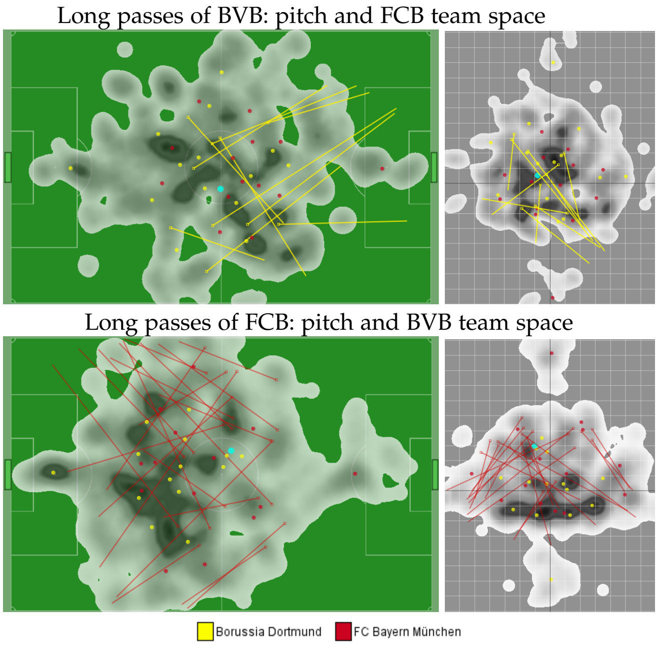

Fig. 12. Long forward passes of BVB (top) and FCB (bottom) and the average positions of all players (shown by colored dots) on top of the density fields of all players during the passes.

different ways to reach their opponents' goal. FCB mostly used the right flank and then turned towards the center of the penalty box. The overall shape of the BVB attacks looks like an arrow targeted straight at the FCB goal.

The position variation indicators (i.e., the median distances to the representative positions) can be compared along and across the pseudo-trajectories. For example, the variation of the positions of the FCB's right midfielder is much higher than that of the left midfielder. This can be related to the fact that the ball was transferred to the penalty box mostly from the right flank, and the opponents were putting more pressure on that side forcing the right midfielder to vary his positions.

Our group considered two options for studying further details for smaller subsets of similar shots. One option was grouping by the shot location. However, by inspecting the episodes preceding the shots, it was found that the variation of the shot positions does not match the variation of the trajectories of the ball, teams, and players. Similar build-ups do not necessarily lead to making shots from similar positions. Another option was to cluster the shots by similarity of the last preceding passes or by similarity of particular trajectories, e.g., of the ball and/or the team centers. We evaluated several variants of grouping using clustering of trajectories by relevant parts [54]. They produced either heterogeneous clusters with too small differences between them or homogeneous clusters that were too small for valid generalization. This procedure, however, appears to have a good potential when applied to a larger number of situations extracted from multiple games of the same team.

5.4 Conclusions from the use cases

Even if the top leagues and clubs are aware of the necessity of acquiring event and tracking data, the potential of using them in the team's daily business is not yet tapped. Since the game is very complex and the interpretation of situations is very subjective even for experts, a lot of data-driven projects fail in terms of communication between data-science and soccer experts. The football experts concluded that the proposed VA approach is a great step towards making the complex spatio-temporal tracking and event data understandable and so usable for professionals.

Application of visual analytics approaches allowed our group to find many interesting patterns that would be very difficult or even impossible to detect by means of watching game footage on TV or an animated visual representation of the data. We were able to identify patterns, compare them, and synthesize further higher-level patterns. Moreover, obtaining interesting and meaningful results motivated many further analysis scenarios such as identifying formations during long periods of ball possession, assessing efficiency of counter-pressing efforts, comparing evolution of playing style of the same team over a season and across multiple

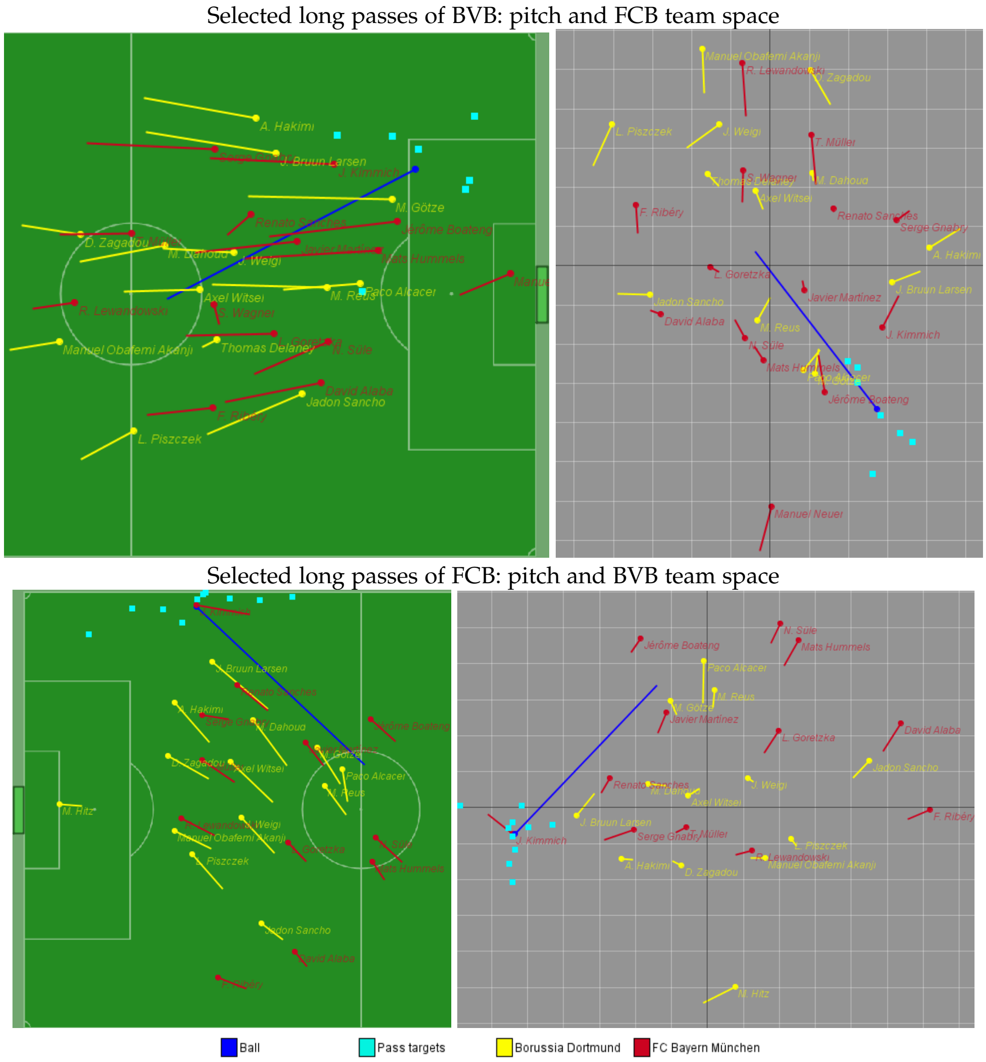

Fig. 13. Changes in aggregated positions of players and the ball during the selected long passes of BVB (top) and FCB (bottom). Pass targets are marked by cyan dots.

seasons, searching for impact of changes of team coaches and/or key players. Considering such scenarios would be out of question if only the state-of-the-art techniques were available.

6 Discussion

This paper presents results of a research project involving visual analytics researchers, data scientists, and football domain experts with the aim to find effective approaches to gaining practicable knowledge from real complex data. The football experts were impressed by the capabilities of the visual analytics techniques that were developed. They said that the selection of similar game situations based on underlying data and extraction of general patterns of the teams' and players' behaviors in such situations has been so far an unsolved challenge. Hence, appropriate query and generalization techniques would bring a big benefit for experts, especially for scouts and match analysts. The proposed framework has a very high potential to bring all the data-insights finally on the pitch and thus produce a substantial impact on professional football.

In the experts' opinion, the power of the framework can be further increased by involving Key-Performance-Indicators (KPI) for the soccer game metrics such as expected goals, dangerosity, pass options based either on measuring the difficulty of performing passes or representing possible gain if a pass is successful (e.g. packing rate), indicators of team compactness and structure such as space occupation, team shape damage, stretch index etc., and pressure indicators. These KPIs are on the rise and could be utilized in making episode queries to both validate and improve the metrics and to select game situations of interest, enabling further application scenarios.

To put the results of this research into practice, it is necessary to develop software tools that can be easily utilized by end users. As there exist different categories of potential users, as discussed below, user-centered design and development may need to be done specifically for each category, taking into account the specific tasks, requirements, and capabilities of the target users. Different classes of users need different interfaces with different levels of interactivity and visual complexity, but, irrespective of these, achieving high level of automation in making queries, constructing

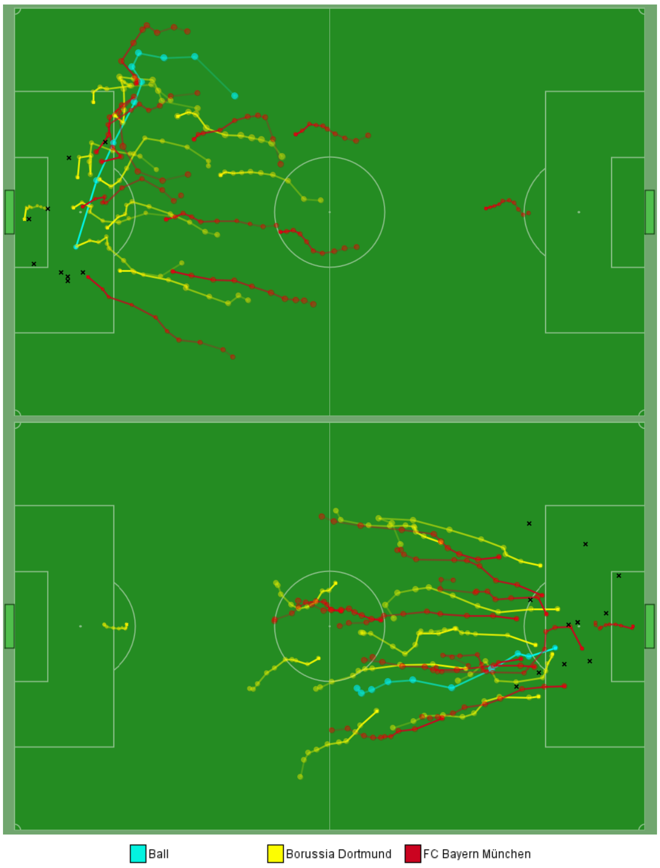

Fig. 14. Build-up for the shots by FCB (top) and BVB (bottom). The shot positions are marked by black crosses. The cyan line corresponds to the ball. The dots with the sizes representing the variation mark the generalized positions, which are separated by 1-second intervals.

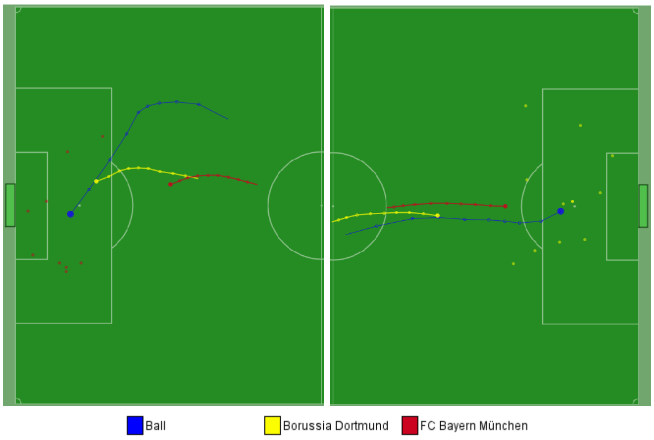

Fig. 15. Build-ups for shots: pseudo-trajectories of the team centers before the FCB shots (left) and BVB shots (right).

pseudo-trajectories, and putting them in visual displays is of great importance. There exists a technical possibility to implement the presented framework in the form of automated procedures oriented to specific analysis tasks. Such automated procedures can be used for extracting patterns from large databases containing data from many games.

We anticipate that the following user categories could benefit from specific applications based on the framework. **Match analysts** could evaluate the efficiency of their own team in previous games and create automatically a catalogue of tactical schemes of opponents over a big set of their previous games. Such automatically acquired tactical schemes conditioned over different classes of situations could be a great hint-giving and decision-supporting means. **Medical staff** of clubs could examine movements of players during episodes characterized by fast running at different times in the games. **Scouts** could evaluate players' movements and actions in different classes of situations and spot their strong and weak abilities. **Journalists** could present tactical schemes and compare them in pre-game and post-game articles or TV shows or even during game breaks. **Leagues** could provide services to clubs and also enhance their media products. Some user categories (particularly, the latter two) require tools not only for analysis but also for communication of the insights gained to certain audiences, including the general public. This requires specific approaches for synthesizing audience-targeted stories from results of tactical analyses [57].

While the presented framework have been developed with an orientation to football data and analysis tasks, it is potentially generalizable to various kinds of coordinated movements of multiple objects in applications where the task of extracting general behaviour patterns under different circumstances is relevant. Examples are movements of players in other team sports, such as ice-hockey or basketball, behaviors of animal groups, or movements of people in crowded environments. We also envision potential applications in domains of air and sea traffic management. The main components of the approach, i.e., the query facilities for episode selection, the method for generalization and aggregation, the data structure for representing generalization results, and the visualization techniques, can be adjusted to the specifics of various application domains.

7 Conclusion

Our contribution can be summarized as proposing an analytical framework involving interactive **queries, generalization and aggregation** of query outcomes, and **comparative visual exploration** of resulting general patterns. The framework makes use of an interesting and fruitful interplay of physical and constructed spaces (pitch and team spaces) and times (absolute and relative times). The query primitives enable selection of sets of time intervals containing situations with specified characteristics and, moreover, further selection of sets of intervals having particular temporal relationships to the previously selected intervals. This can be used, in particular, for considering the episodes of situation development step-wise or for studying what happened before or after them. The aggregation method produces a novel type of movement aggregate, pseudo-trajectory, consisting of generalized positions arranged along an abstract timeline. The aggregation results are visualized in ways enabling exploration, comparison, and assessment of the variation of the original movements summarized in the aggregates. The techniques proved useful for discovery of general patterns of collective movement behavior in diverse classes of situations.

ACKNOWLEDGMENTS

This research was supported by Fraunhofer Cluster of Excellence on "Cognitive Internet Technologies" and by EU in project SoBigData.

REFERENCES

[1] Wikipedia, *Association football*, 2019 (accessed February 14, 2019), https://en.wikipedia.org/wiki/Association_football.

[2] D. Memmert and D. Raabe, *Revolution im Profifußball. Mit Big Data zur Spielanalyse 4.0*. Springer, 2017.

[3] D. Link, *Data Analytics in Professional Soccer*. Springer, 2018.

[4] D. Sumpter, *Soccermatics: mathematical adventures in the beautiful game*. Bloomsbury Publishing, 2016.

[5] J. Ladefoged, *How Data (and Some Breathtaking Soccer) Brought Liverpool to the Cusp of Glory*, 2019, https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html (accessed June 18, 2019).

[6] Bundesliga, *Borussia Dortmund - FC Bayern München 3:2*, 10.11.2018 (accessed February 14, 2019). [Online]. Available: https://www.bundesliga.com/de/bundesliga/spieltag/2018-2019/11/borussia-dortmund-vs-fc-bayern-muenchen/stats

[7] ——, *Borussia Dortmund - 1. FC Nürnberg 7:0*, 26.09.2018 (accessed February 14, 2019). [Online]. Available: https://www.bundesliga.com/de/bundesliga/spieltag/2018-2019/5/borussia-dortmund-vs-1-fc-nuernberg/stats

[8] Wikipedia, *Formation (association football)*, 2019 (accessed February 14, 2019), https://en.wikipedia.org/wiki/Formation_(association_football).

[9] G. Andrienko, N. Andrienko, G. Budziak, J. Dykes, G. Fuchs, T. von Landesberger, and H. Weber, "Visual analysis of pressure in football," *Data Mining and Knowledge Discovery*, vol. 31, no. 6, pp. 1793—-1839, 2017.

[10] J. Wilson, *Inverting the pyramid: the history of soccer tactics*. Nation Books, 2013.

[11] A. Zauli, *Soccer: Modern Tactics*. Reedswain Inc., 2002.

[12] M. Lucchesi, *Attacking soccer: A tactical analysis*. Reedswain Inc., 2001.

[13] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. A. Matthews, "Large-scale analysis of soccer matches using spatiotemporal tracking data," in *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014*, R. Kumar, H. Toivonen, J. Pei, J. Z. Huang, and X. Wu, Eds. IEEE, 2014, pp. 725–730.

[14] A. Bialkowski, P. Lucey, G. P. K. Carr, Y. Yue, S. Sridharan, and I. A. Matthews, "Identifying team style in soccer using formations learned from spatiotemporal tracking data," in *2014 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2014, Shenzhen, China, December 14, 2014*, Z. Zhou, W. Wang, R. Kumar, H. Toivonen, J. Pei, J. Z. Huang, and X. Wu, Eds. IEEE, 2014, pp. 9–14.

[15] J. Perl, A. Grunz, and D. Memmert, "Tactics analysis in soccer– an advanced approach," *International Journal of Computer Science in Sport*, vol. 12, no. 1, pp. 33–44, 2013.

[16] Y. Wu, X. Xie, J. Wang, D. Deng, H. Liang, H. Zhang, S. Cheng, and W. Chen, "Forvizor: Visualizing spatio-temporal team formations in soccer," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 65–75, Jan 2019.

[17] ChyronHego, *https://chyronhego.com/*, 2019 (accessed June 18, 2019).

[18] OPTA, *https://www.optasports.com/*, 2019 (accessed February 14, 2019).

[19] STATS, *https://www.stats.com/*, 2019 (accessed February 14, 2019).

[20] S. Spectrum, *https://www.secondspectrum.com/*, 2019 (accessed June 18, 2019).

[21] Track160, *https://track160.com/*, 2019 (accessed June 18, 2019).

[22] FootoVision, *https://www.footovision.com/*, 2019 (accessed February 14, 2019).

[23] G. Andrienko, N. Andrienko, and G. Fuchs, "Understanding movement data quality," *Journal of Location Based Services*, vol. 10, no. 1, pp. 31–46, 2016.

[24] D. J. Sumpter, *Collective animal behavior*. Princeton University Press, 2010.

[25] J. Gudmundsson and M. Horton, "Spatio-temporal analysis of team sports," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, p. 22, 2017.

[26] C. Perin, R. Vuillemot, and J.-D. Fekete, "Soccerstories: A kick-off for visual soccer analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2506–2515, 2013.

[27] D. Sacha, F. Al-Masoudi, M. Stein, T. Schreck, D. Keim, G. Andrienko, and H. Janetzko, "Dynamic visual abstraction of soccer movement," *Computer Graphics Forum*, vol. 36, no. 3, pp. 305–315, 2017.

[28] D. Sacha, M. Stein, T. Schreck, D. A. Keim, O. Deussen *et al.*, "Feature-driven visual analytics of soccer data," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2014, pp. 13–22.

[29] L. Shao, D. Sacha, B. Neldner, M. Stein, and T. Schreck, "Visual-interactive search for soccer trajectories to identify interesting game situations," *Electronic Imaging*, vol. 2016, no. 1, pp. 1–10, 2016.

[30] M. Stein, H. Janetzko, T. Schreck, and D. A. Keim, "Tackling Similarity Search for Soccer Match Analysis: Multimodal Distance Measure and Interactive Query Definition," in *Symposium on Visualization in Data Science (VDS) at IEEE VIS 2018*, 2018.

[31] M. Stein, H. Janetzko, T. Breitkreutz, D. Seebacher, T. Schreck, M. Grossniklaus, I. D. Couzin, and D. A. Keim, "Director's cut: Analysis and annotation of soccer matches," *IEEE Computer Graphics and Applications*, vol. 36, no. 5, pp. 50–60, 2016.

[32] M. Stein, J. Häußler, D. Jäckle, H. Janetzko, T. Schreck, and D. A. Keim, "Visual soccer analytics: Understanding the characteristics of collective team movement based on feature-driven analysis and abstraction," *ISPRS International Journal of Geo-Information*, vol. 4, no. 4, pp. 2159–2184, 2015.

[33] G. Andrienko, N. Andrienko, G. Budziak, T. von Landesberger, and H. Weber, "Coordinate transformations for characterization and cluster analysis of spatial configurations in football," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 27–31.

[34] J. Gudmundsson and T. Wolle, "Football analysis using spatio-temporal tools," *Computers, Environment and Urban Systems*, vol. 47, pp. 16–27, 2014.

[35] M. Horton, J. Gudmundsson, S. Chawla, and J. Estephan, "Automated classification of passing in football," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2015, pp. 319–330.

[36] M. Stein, H. Janetzko, A. Lamprecht, T. Breitkreutz, P. Zimmermann, B. Goldlücke, T. Schreck, G. Andrienko, M. Grossniklaus, and D. A. Keim, "Bring it to the pitch: Combining video and movement data to enhance team sport analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 13–22, Jan 2018.

[37] STATSEdgeViewer, *https://www.stats.com/edge/*, 2019 (accessed June 18, 2019).

[38] SAPSportsOne, *https://www.sap.com/germany/products/sports-one.html*, 2019 (accessed June 18, 2019).

[39] SportecSolutions, *http://www.bundesliga-datenbank.de/*, 2019 (accessed June 18, 2019).

[40] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel, *Visual Analytics of Movement*. Springer, 2013.

[41] G. Andrienko, N. Andrienko, W. Chen, R. Maciejewski, and Y. Zhao, "Visual analytics of mobility and transportation: State of the art and further research directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 2232–2249, Aug 2017.

[42] C. Weaver, "Cross-filtered views for multidimensional visual analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 2, pp. 192–204, March 2010.

[43] H. Hochheiser and B. Shneiderman, "Dynamic query tools for time series data sets: Timebox widgets for interactive exploration," *Information Visualization*, vol. 3, no. 1, pp. 1–18, Mar. 2004.

[44] S. K. Gadia, "A homogeneous relational model and query languages for temporal databases," *ACM Trans. Database Syst.*, vol. 13, no. 4, pp. 418–448, Oct. 1988.

[45] C. S. Jensen, J. Clifford, S. K. Gadia, A. Segev, and R. T. Snodgrass, "A glossary of temporal database concepts," *SIGMOD Rec.*, vol. 21, no. 3, pp. 35–43, Sep. 1992.

[46] N. Andrienko, G. Andrienko, E. Camossi, C. Claramunt, J. M. C. Garcia, G. Fuchs, M. Hadzagic, A.-L. Jousselme, C. Ray, D. Scarlatti, and G. Vouros, "Visual exploration of movement and event data with interactive time masks," *Visual Informatics*, vol. 1, no. 1, pp. 25 – 39, 2017.

[47] T. von Landesberger, S. Bremm, T. Schreck, and D. W. Fellner, "Feature-based automatic identification of interesting data seg-

ments in group movement data," *Information Visualization*, vol. 13, no. 3, pp. 190–212, 2014.
[48] N. Andrienko, G. Andrienko, J. M. C. Garcia, and D. Scarlatti, "Analysis of flight variability: a systematic approach," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 54–64, Jan 2019.
[49] N. Andrienko, G. Andrienko, L. Barrett, M. Dostie, and P. Henzi, "Space transformation for understanding group movement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2169–2178, Dec 2013.
[50] G. Andrienko, N. Andrienko, H. Schumann, and C. Tominski, *Visualization of Trajectory Attributes in Space–Time Cube and Trajectory Wall*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 157–163.
[51] N. Andrienko and G. Andrienko, "Spatial generalization and aggregation of massive movement data," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 2, pp. 205–219, 2011.
[52] N. Willems, H. Van De Wetering, and J. J. Van Wijk, "Visualization of vessel movements," *Computer Graphics Forum*, vol. 28, no. 3, pp. 959–966, 2009.
[53] G. Andrienko and N. Andrienko, "Spatio-temporal aggregation for visual analysis of movements," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST) 2008*, 2008, pp. 51–58. [Online]. Available: https://doi.org/10.1109/VAST.2008.4677356
[54] G. Andrienko, N. Andrienko, G. Fuchs, and J. M. Cordero-Garcia, "Clustering trajectories by relevant parts for air traffic analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 34–44, Jan 2018.
[55] A. Buja, J. A. McDonald, J. Michalak, and W. Stuetzle, "Interactive data visualization using focusing and linking," in *Proceedings of the 2Nd Conference on Visualization '91*, ser. VIS '91. Los Alamitos, CA, USA: IEEE Computer Society Press, 1991, pp. 156–163. [Online]. Available: http://dl.acm.org/citation.cfm?id=949607.949633
[56] J. Candil, *Pep's five-second rule, the key to City's success*, 2018, https://en.as.com/en/2018/07/26/football/1532614241_079674.html (accessed February 14, 2019).
[57] S. Chen, J. Li, G. Andrienko, N. Andrienko, Y. Wang, P. H. Nguyen, and C. Turkay, "Supporting story synthesis: Bridging the gap between visual analytics and storytelling," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

Gennady Andrienko is a lead scientist responsible for visual analytics research at Fraunhofer Institute for Intelligent Analysis and Information Systems and part-time professor at City University London. Gennady Andrienko was a paper chair of *IEEE VAST* conference (2015–2016) and associate editor of *IEEE Transactions on Visualization and Computer Graphics* (2012–2016), *Information Visualization* and *International Journal of Cartography*.

Natalia Andrienko is a lead scientist at Fraunhofer Institute for Intelligent Analysis and Information Systems and part-time professor at City University London. Results of her research have been published in two monographs, "*Exploratory Analysis of Spatial and Temporal Data: a Systematic Approach*" (2006) and "*Visual Analytics of Movement*" (2013). Natalia Andrienko is an associate editor of *IEEE Transactions on Visualization and Computer Graphics*.

Gabriel Anzer is the lead data scientist at Sportec Solutions GmbH, a subsidiary of the Deutsche Fußball Liga (DFL). He holds a M.Sc. in Financial Mathematics and Actuarial Sciences. His research focuses on using spatio-temporal positional data to analyze individual and team based performances of soccer players.

Pascal Bauer has a background in mathematics and data-science in applied research at Fraunhofer and worked as a coach/speaker at Fraunhofer Big Data & Artificial Intelligence Alliance. He holds a UEFA A-level coaching license with almost nine years of experience as a head coach and has a passion for soccer. At the DFB Academy, he is responsible for a wide field of data analysis & machine learning applications in soccer, including talent identification, injury prediction models, tactical match analysis based on positional data, and much more.

Georg Fuchs is a senior research scientist heading the Big Data Analytics and Intelligence division at Fraunhofer IAIS. His research work is focussed on visual analytics and Big Data analytics, with a strong emphasis on spatio-temporal and movement data analysis. His further research interests include information visualization in general, Smart Visual Interfaces, and computer graphics. He has co-authored 55+ peer-reviewed research papers and journal articles, including a best short paper award at Smart Graphics 2008 and a VAST challenge award in 2014.

Guido Budziak holds a master's degree in Computer Science. He played professional football for 10 years in the Netherlands, including Dutch national youth teams. He founded Connected.Football, a football technology company aimed at making expert football knowledge accessible to youth academies and amateur football clubs. His research work focuses on communicating football tactics and tactical performance analysis.

Dirk Hecker is vice-director of the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS and a member of the board of directors of Fraunhofer Academy. His current topics of work include spatial analytics, deep learning and trustworthy AI. He has authored multiple publications on these topics and works as expert and auditor in several boards concerned with artificial intelligence.

Hendrik Weber leads the area of innovation and sports technology for the German Football League (DFL) and is managing director of the subsidiary Sportec Solutions responsible for Sport Tech Operations. He holds a PhD in business administration and is frequent lecturer and author for performance analysis in sports.

Stefan Wrobel is Professor of Computer Science at University of Bonn and Director of the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS. His work is focused on questions of the digital revolution, in particular intelligent algorithms and systems for the large-scale analysis of data and the influence of Big Data/Smart Data on the use of information in companies and society. He is the author of a large number of publications on data mining and machine learning, is on the Editorial Board of several leading academic journals in his field, and is an elected founding member of the "International Machine Learning Society".

B Appendix—Study II: 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$^{\text{nd}}$ 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$^{\text{nd}}$ 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 2$^{\text{nd}}$ 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

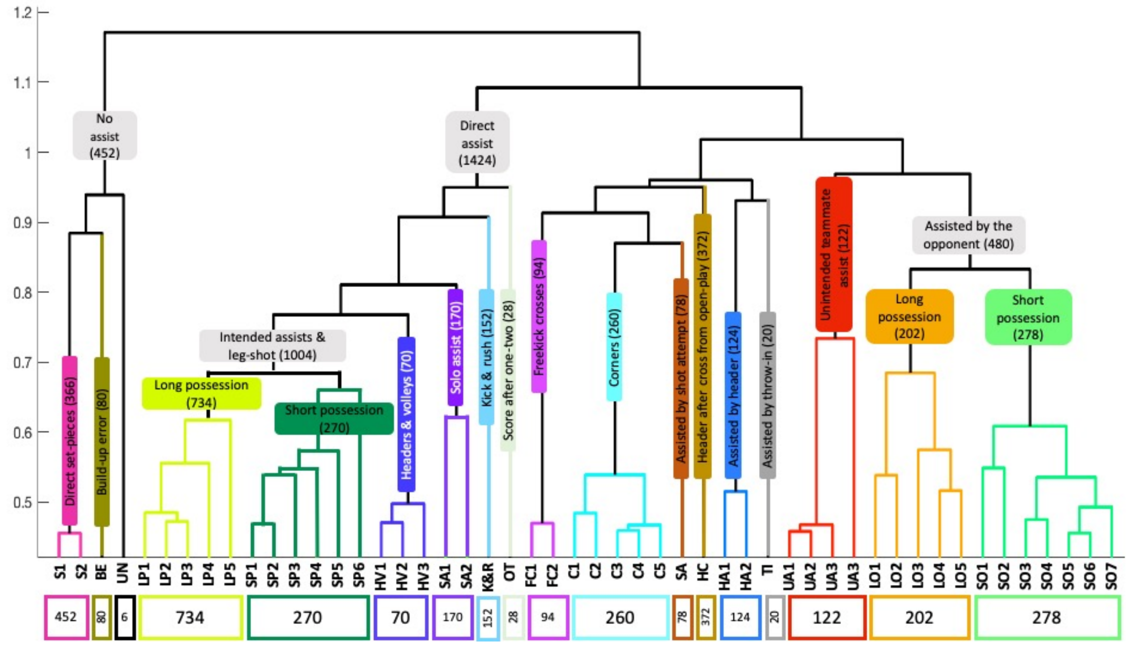

Figure 1: The resulting dendrogram with contextual annotations. Numbers show the amount of goals in the respective branches.

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

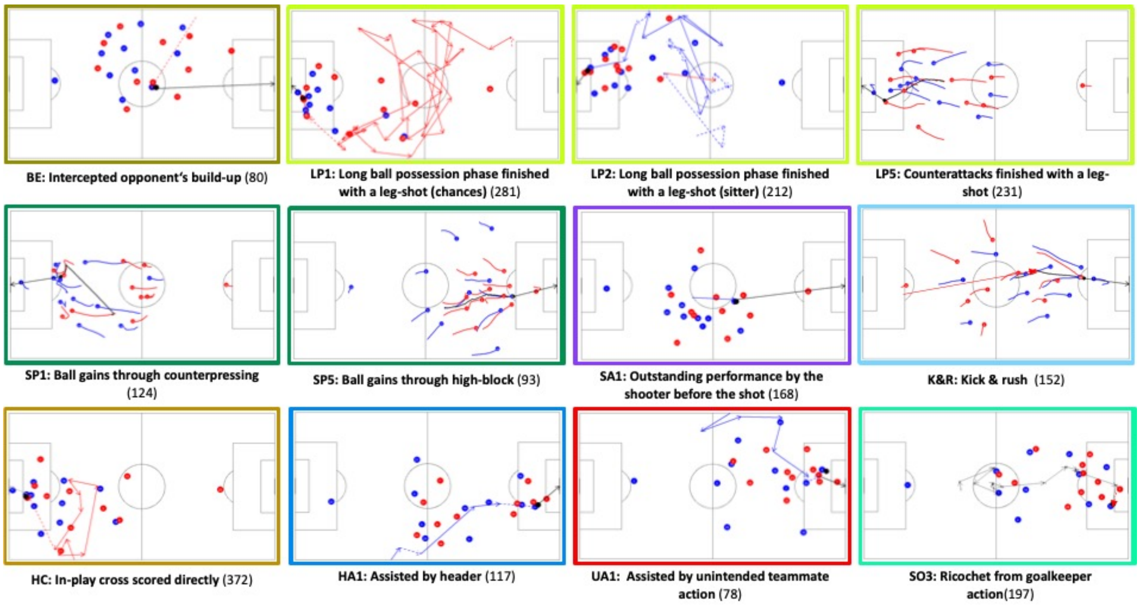

Figure 2: Exemplary in-play goals and their clusters. Arrows show the path of the ball leading up to the successful shot, positions of players at the time of shot are indicated in blue and red, respectively.

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

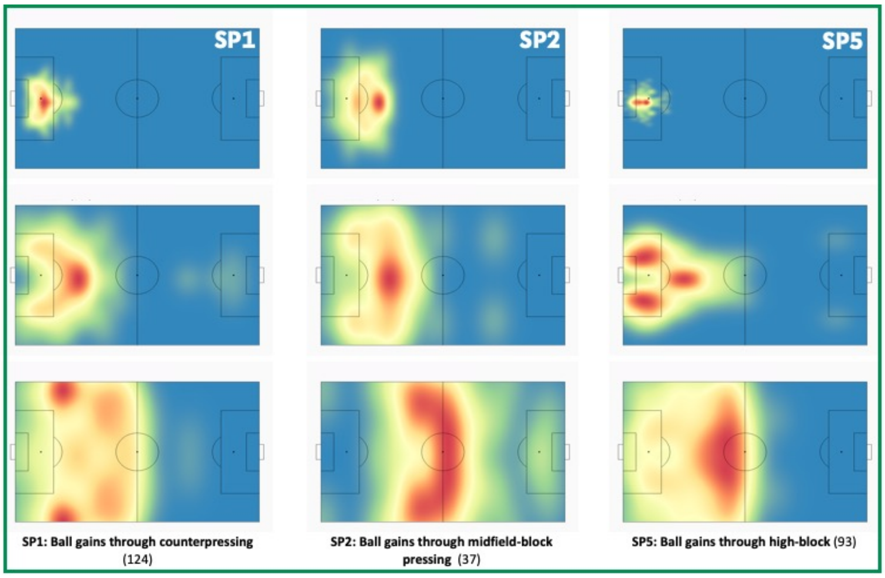

Figure 3: Goals originating from strategic ball gains in **SP1**, **SP2**, and **SP5**. The figure shows heat maps for shot (top row), assist location (center row) and start of the ball possession phase (bottom row). In total 270 goals (7.8% of all) were scored this way.

(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*.

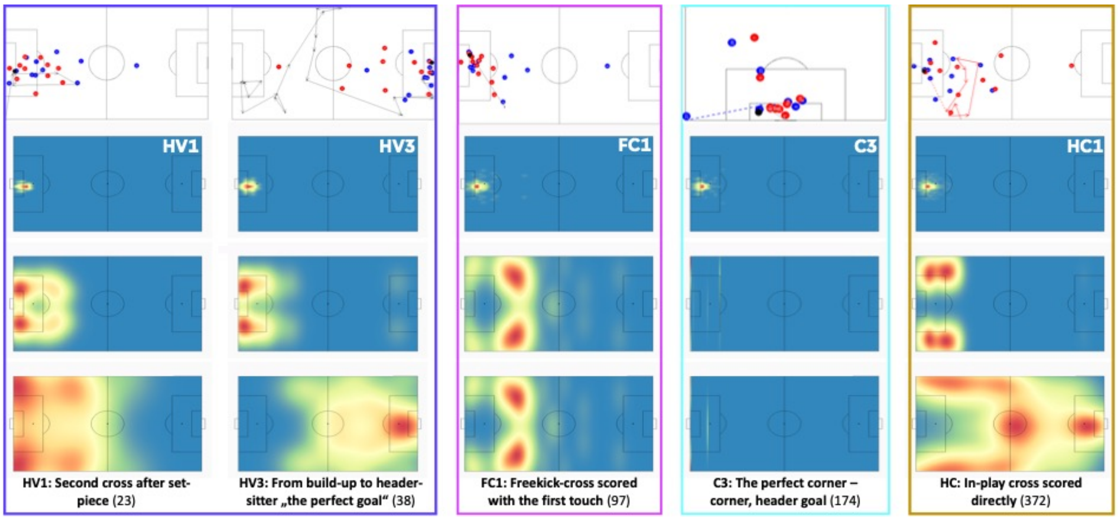

Figure 4: Visualizations of goals scored by headers: example goal (top row), scorer position (2nd row), assist location (3nd row), begin attacking phase (bottom row).

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

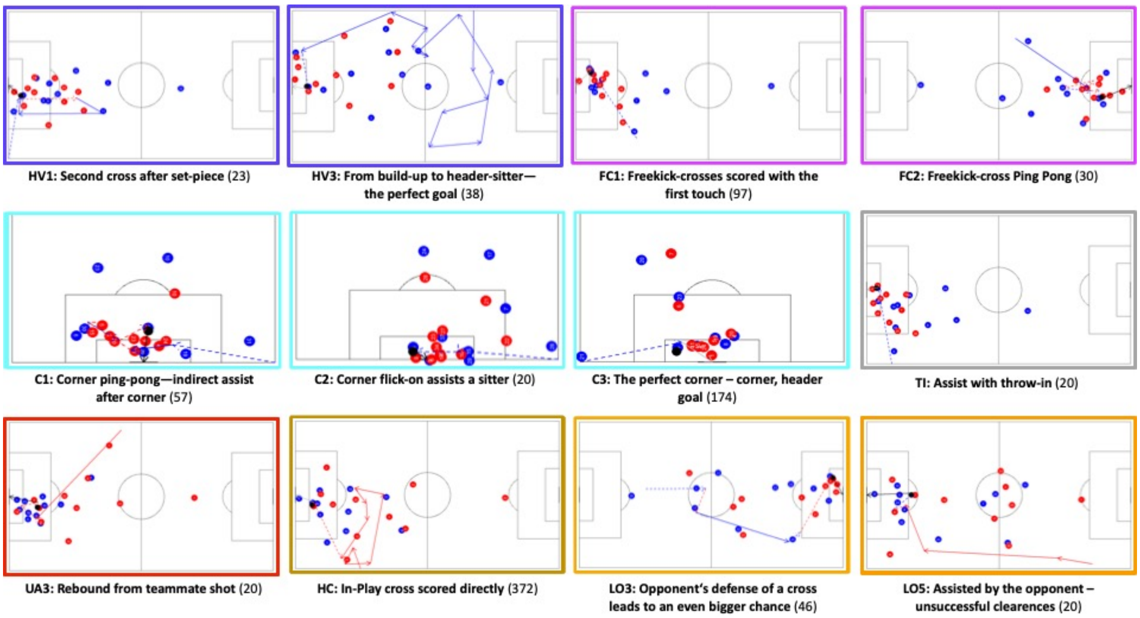

Figure 5: Exemplary visualizations of all goals occurring by corners and crosses. Both goals from set-pieces and from open-play are included representing a total amount of 490 goals (14.17%).

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

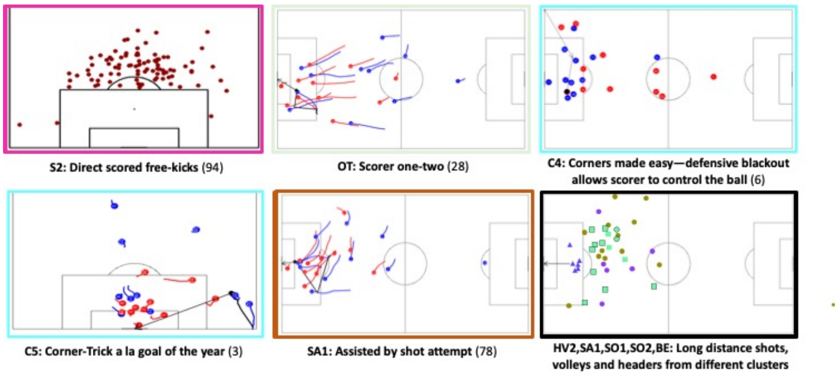

Figure 6: Selected special goals. Top left: shot chart for **S2**. Bottom right: aggregation of extraordinary shots (circles), volleys (squares), and headers (triangles) from different clusters: shots are plotted as circles, volleys as quarters and headers as triangles - all in the respective cluster-color.

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

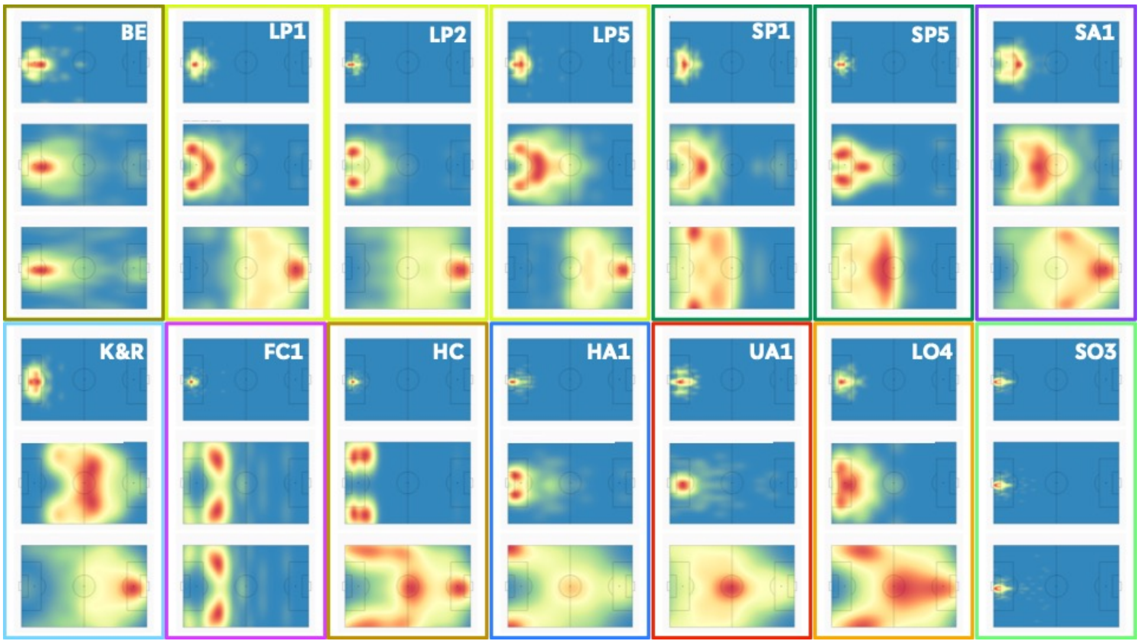

Figure 7: The largest 14 in-play clusters shown by three heatmaps each: start of the ball possession phase (bottom), the assist location (middle) such as the shot location (top).

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

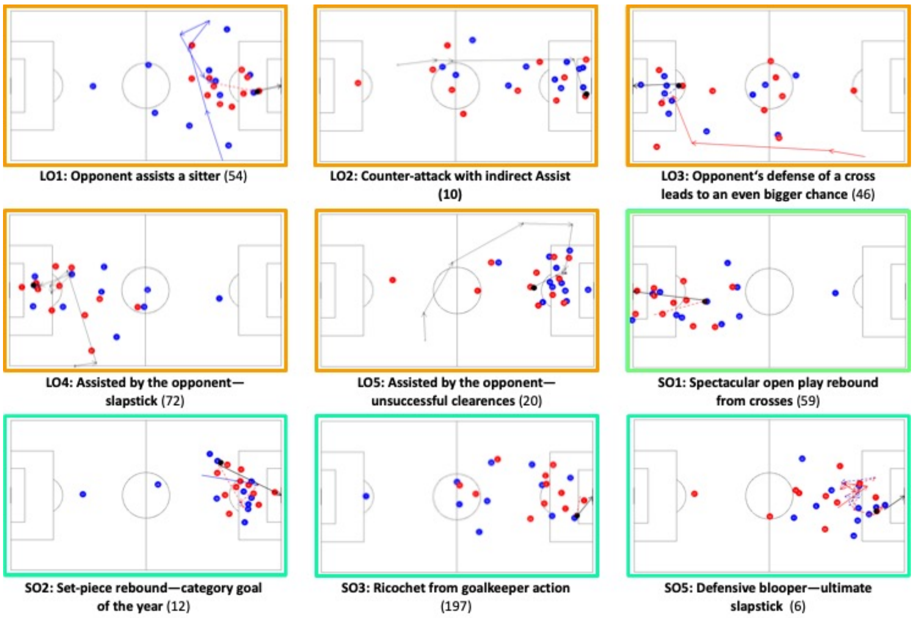

Figure 8: Visualizations of goals assisted by the opponent. The 278 goals in this category present 7.8% of all goals.

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.

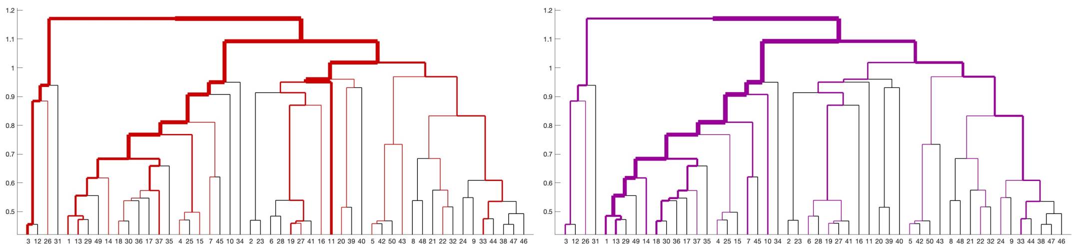

Figure 9: Footprints of Robert Lewandowski (left) and Timo Werner (right) in the dendrogram.

5 Conclusions

We studied the origin of goals in the German Bundesliga and 2$^{\text{nd}}$ 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.

	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
Scored goals	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	3	4	13
Hannover 96	74	7	2	0	15	3	0	3	4	0	3	8	2	9	4	0	1	8	5
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	6	7	4
1. FC Heidenheim 1846	101	6	7	1	21	4	4	4	4	0	3	7	0	16	5	0	7	7	5
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	2	6	11
Eintracht Braunschweig	34	0	0	0	12	5	0	1	3	0	0	1	1	1	2	0	4	0	4
SSV Jahn Regensburg	103	7	4	3	19	7	1	7	5	2	5	5	3	12	5	0	4	8	6
DSC Arminia Bielefeld	99	8	5	5	16	10	0	5	3	0	6	5	1	15	7	1	4	5	3
SV Darmstadt 98	81	9	4	0	7	2	2	5	4	0	4	4	1	13	7	0	4	5	10
FC St. Pauli	81	6	2	2	16	8	2	3	5	0	4	4	2	8	1	1	4	7	6
FC Erzgebirge Aue	76	6	1	1	18	4	1	9	5	0	1	6	1	7	1	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	6	7	7
SC Paderborn 07	73	3	4	4	23	9	3	4	2	0	2	4	0	4	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	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	→ 4	→ 4	↗ 5
Borussia Mönchengladbach	93	→ 7	↑ 0	↗ 2	↓ 28	→ 6	↑ 1	↘ 6	↑ 2	↗ 1	→ 4	↗ 5	↗ 1	→ 9	→ 5	↑ 0	→ 4	→ 5	↗ 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

C Appendix—Study III: Data-Driven Detection of Counterpressing in Professional Football

In the following, Bauer and Anzer (2021) is reproduced with permission from Springer.

Data Mining and Knowledge Discovery
https://doi.org/10.1007/s10618-021-00763-7

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

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

Received: 17 August 2020 / Accepted: 4 May 2021

Abstract
Detecting counterpressing is an important task for any professional match-analyst in football (soccer), but is being done exclusively manually by observing video footage. The purpose of this paper is not only to automatically identify this strategy, but also to derive metrics that support coaches with the analysis of transition situations. Additionally, we want to infer objective influence factors for its success and assess the validity of peer-created rules of thumb established in by practitioners. Based on a combination of positional and event data we detect counterpressing situations as a supervised machine learning task. Together, with professional match-analysis experts we discussed and consolidated a consistent definition, extracted 134 features and manually labeled more than 20, 000 defensive transition situations from 97 professional football matches. The extreme gradient boosting model—with an area under the curve of 87.4% on the labeled test data—enabled us to judge how quickly teams can win the ball back with counterpressing strategies, how many shots they create or allow immediately afterwards and to determine what the most important success drivers are. We applied this automatic detection on all matches from six full seasons of the German Bundesliga and quantified the defensive and offensive consequences when applying counterpressing for each team. Automating the task saves analysts a tremen-

Responsible editor: Albrecht Zimmermann.

Pascal Bauer
pascal.bauer@dfb.de

Gabriel Anzer
gabriel.anzer@sportec-solutions.de

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

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

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

Published online: 08 July 2021

dous amount of time, standardizes the otherwise subjective task, and allows to identify trends within larger data-sets. We present an effective way of how the detection and the lessons learned from this investigation are integrated effectively into common match-analysis processes.

Keywords Sports analytics · Football (Soccer) · Tactical performance analysis · Applied machine learning · Positional and event data

1 Introduction

Acquiring accurate and high frequency positional and event data is common in most of the world's top professional football (soccer) leagues. Manually annotated *event data* provides information about the one player carrying the ball at the time of a game relevant action only, whereas so called *positional data* can capture highly accurate positions of all 22 players up to 25 times a second.

Every professional football team spends a substantial amount of time analyzing and monitoring strategies such as *counterpressing*.—a complex team strategy for transition situations—of their own and opposing teams. Navarro and Javier (2018) defines counterpressing as simple as "*[..] pressure after losing the ball*". Related to this, Pep Guardiola made the '*five second rule*' for counterpressing famous.[1] Another coach to experience tremendous success in recent seasons is Liverpool FC manager Jürgen Klopp. He is generally accepted as the originator of the term '*Gegenpressing*', which is well-known in both its German version and English translation.[2] It is apparent to football experts that Klopp's counterpressing concept is closely related to Guardiola's strategy of regaining the ball within the first five seconds.

There are significant differences in team's defensive and offensive tactical line-ups (Bialkowski et al. 2014, 2015; Andrienko 2019; Shaw and Mark 2019). The *transition phase* describes the period following a win or loss of possession in which the team transitions between its offensive and defensive tactical line-ups and vice versa. When a team is in possession for at least a certain amount of time, we can assume that generally its tactical formation is optimized for offensive play and, consequently, suboptimal in terms of defending its own goal (Andrienko 2019; Shaw and Mark 2019). Therefore, the first seconds after losing the ball are critical from a defensive perspective. Several studies proved that transition phases are a substantial factor for a team's overall performance: As early as in 1968, Reep and Benjamin (1968) demonstrated in the first known football analytics study, that 30% of all regained possessions lead to shots on goal and 25% of all goals came from regained possessions in the attacking quarter. Grant et al. (1999) confirmed these findings for the 1998 World Cup. Both outcomes align perfectly with Jürgen Klopp's statement that regaining the ball immediately after loosing it, potentially through successful counterpressing, *"...is the best*

[1] "*[..] after losing the ball, the team has five seconds to retrieve the ball, or, if unsuccessful, tactically foul their opponent and fallback*", Pep Guardiola; https://www.theblizzard.co.uk/article/peps-four-golden-rules, accessed 06/20/2020.

[2] https://www.sueddeutsche.de/sport/premier-league-bei-klopps-liverpoolern-klemmt-das-gaspedal-1.2695408-2, accessed 06/20/2020.

playmaker".[3] Klopp hereby claims that counterpressing can also be seen as an offensive strategy. Recent studies show that regaining the ball in open play likelier leads to a goal than a save build-up from a team's own half (Vogelbein et al. 2014; Hobbs et al. 2018). Based on tracking data from the English Premier League, Hobbs et al. (2018) detected possession regains close to the opponent's goal—potential counterpressing situations—highlighting their relevance once more. Even though many coaches and clubs affected the development of this sophisticated strategy, neither an objective proof of its efficiency, nor an analysis on its usage in top leagues is presented in the literature.

Hughes and Ian (2015) point out that team sports performance analysis tends to be operationalized on the basis of *notation systems*, described as a replicable and consistent method of recording sport performance. Recent literature explained a framework, where coaches‘ decisions are supported by several performance analysis reports from games, teams and players (Travassos et al. 2013) and pointed out that team tactics in football refer to both a priori decisions made before the match, and also real-time adaptions during the game (Rein and Daniel 2016). In accordance to that it is described as a complex process resulting from a network of inter-dependent parameters (Kempe et al. 2014). These processes are conducted in a time-critical set-up, especially when it comes to the world's top leagues and competitions where teams need to encounter different opponents several times a week. Although many clubs extended their match-analysis departments considerably within the past years, the limited amount of time and resources during matches forced teams to seek ways to automate processes and gain insights faster in order to obtain a competitive edge.

These recent developments—the availability of accurate performance data and the need for a quick detailed tactical analysis—signifies a huge potential for the application of sophisticated machine learning techniques to football data and requires an efficient collaboration of computer-science and domain experts (Herold et al. 2019; Goes et al. 2020; Rein and Daniel 2016). Many recent scientific investigations aimed to establish new *key performance indicator (KPI)*—metrics quantifying certain aspects of the game: pass evaluation metrics were examined (Steiner et al. 2019; Goes et al. 2019), metrics to quantify controlled space were defined (Kim 2004; Fernandez and Bornn 2018; Brefeld et al. 2019) and several studies evaluated shot metrics (Lucey et al. 2014; Rathke 2017; Fairchild et al. 2018; Anzer and Bauer 2021)[4] and goal scoring opportunities through possession values (Link et al. 2016; Spearman 2018; Fernandez and Bornn 2018; Decroos et al. 2020). Additionally, there are many approaches for measuring the defensive behavior of teams (Santos et al. 2018; Andrienko 2019; Goes et al. 2019), and even approaches aiming to quantify pressing (Bojinov and Luke 2016; Andrienko 2017; Robberechts 2019). Although pressing and counterpressing are closely related, they are two different phenomena. An interesting conference proceeding describes how specific counterpressing situations can be derived from detected general pressing scenes.[5] Several approaches also showed, that analyzing these KPI's or even aggregating simple statistics (e.g. the pass completion rate) over one or several

[3] https://tactical-times.com/the-history-and-evolution-of-jurgen-klopp/, accessed 06/20/2020.

[4] Often referred to as *expected goals (xG)* values.

[5] Will Gürpinar-Morgen (2018). "How StatsBomb data helps measure counterpressing", *Statsbomb Innovation in Football Conference 2018;* https://statsbomb.com/2018/05/how-statsbomb-data-helps-measure-counter-pressing/, accessed 11/11/2020.

seasons provides a helpful indication to practitioners (Power et al. 2018; Pappalardo et al. 2019). The primary goal of all these approaches is to derive new insights by processing vast amounts of information. Decroos et al. (2018) presented a first approach to detect interesting match-phases based on event data. To the best of our knowledge no peer-reviewed study focused on automating parts of the performance analysts everyday life by detecting complex tactical patterns based on positional and event data. However, a noteworthy approach aiming to detect counterattacks was presented in an established football analytics conference.[6]

With this practical need for process optimization in mind, it is the pivotal issue of this study to detect counterpressing situations without human-support and provide several ad-hoc reports for match analysts in near real-time. The outcome is optimized to fulfill their practical requirements and fit seamlessly into their tool-ecosystem. Additionally, the automated detection allows us to analyze large amounts of data that would exceed manual processing capacities. Consequently, our approach enables us to perform impartial long-term analysis of the German Bundesliga's latest seasons investigating the following research questions:

– Can we differentiate between varying regaining strategies and determine reasons for a short defensive reaction time (definition in Sect. 2.1.2), i.e. to which extent is a fast ball regain actually caused by counterpressing (*RQ1*)?
– Can we set objective benchmarks to quantify counterpressing strategies (amount and effectiveness) on a match- and season-level and point out their correlation with a team's overall success (*RQ2*)?
– Do the established rules of thumb agree with the data (i.e. counterpressing is more effective close to the sideline) (*RQ3*)?
– To what extent do team's counterpressing strategies differ in the German Bundesliga (*RQ4*)?

All together, answering these research questions helps us to define the baseline for a qualitative discussion with experts, and thus allows them to formulate requirements for the practical application (*PA*) set-up.

The remainder of this paper is structured as follows: Sect. 2 provides a detailed description of the used data, the underlying rules and definitions, the labelling process and the extracted features. The outcomes in Sect. 3 are split into three parts: First in Sect. 3.1, we describe a statistical evaluation of the detection models. Section 3.2 presents a subject-specific evaluation by interpreting our results on six seasons of German Bundesliga. Lastly, in Sect. 3.3, we demonstrate how this approach can be operationalized in the performance analysis process. This application is based on two matches of the German national teams.[7,8] All parts of this study were developed in close cooperation with the professional match-analysts and coaches (see Acknowledgements).

6 Karun Singh (2020). "Learning to watch football: self-supervised representations for tracking data" In *OptaPro Analytics Forum, London;* https://www.youtube.com/watch?v=H1iho17lnoI, accessed 11/11/2020.

7 Germany against Northern Ireland; *19th of November 2019, Commerzbankarena Frankfurt*.

8 Germany U21 against Belgium U21; *17th of November 2019, Schwarzbaldstadion Freiburg*.

2 Methods

2.1 Data and definitions

2.1.1 Data collection

The present study uses *positional* and *event data* collected in more than six seasons (4118 matches) of the German Bundesliga and 2nd Bundesliga, as well as the above mentioned two matches of the German national teams. Positional data is captured by optical tracking systems[9] and event data consists of manual annotations based on a dedicated *official match data catalogue*,[10] defining around 30 events with more than 100 attributes. The event data can be seen as a log of the ball relevant actions (e.g. passes, shots, tacklings or fouls), however, it does not cover complex team-tactical behaviors such as counterpressing.

Since the two data sources are collected independently of each other, they need to be synchronized before they can be processed together. Even though several steps of quality management from independent institutions are conducted on the manually collected event data, the assigned timestamp of a given event can differ significantly to one in the positional data. The synchronization of positional and event data is conducted by dedicated rules (per event) that extracts the exact timestamp and the exact location on the pitch from tracking data given a manually tagged event. For example, when synchronizing a pass, the sudden increase in the distance between the passing player and the ball, captured by the optical tracking, can be used to align both location and timestamp of the pass. The positional data is collected at a frequency of 25 Hz and includes the longitudinal, latitudinal, and in case of the ball, also the altitudinal positions of the players, ball and referees related to the pitch markings.

The information about which team is currently in possession of the ball (hereafter referred to as *ball possession*) and whether the game is running or currently stopped (hereafter referred to as *ball status*) are crucial for our survey. Both values are collected live in the stadium for every frame of the match by a skilled operator focused exclusively on this task.

2.1.2 Definitions

Since there are conflicting definitions of ball possession in the literature (Kempe et al. 2014), we decided to adopt published definitions with expert feedback. The above mentioned operators, dedicated to acquire information about ball possession and status, are briefed to mark *ball possession for one team*, if and from that time point a player of that team touches the ball with ball control, until the ball is out of play, or an opponent player touches the ball with ball control. *Ball control* is defined in this context, as the ability to conduct a contrived action with the ball. Whenever a pass is played between two players of one team, the ball possession belongs to that team as long as no opposing player intercepted that pass or won the ball within an

9 https://chyronhego.com/products/sports-tracking/tracab-optical-tracking/, accessed 06/20/2020.

10 https://www.sportec-solutions.de/en/index.html, accessed 06/20/2020.

individual duel. According to the definition from Link and Hoernig (2017) we also compute ball possession on a player level (*individual ball possession*). In the case of an interception, the ball possession change is detected exactly at the time of the first ball touch of the intercepting player. We use the term *defensive reaction time*—the time it takes to regain ball possession after losing it—as defined in Vogelbein et al. (2014). All situations where either the ball is beyond the pitch markings or the play is stopped by the referee (e.g. because of a foul) are labeled as out-of-play. Hence, if the ball goes out of bounds there must typically be a change in ball possession. Situations in which the touch of the player carrying the ball outside the markings is not declared as a ball possession due to missing control (e.g. a deflected shot), or when the individual possession model disagrees with the team possession flag are excluded. For all further investigations only the *effective playing time* (also referred to as *net playing time*)—defined as all the situations while the game is running—are considered. Shots, for example, always represent the end of a ball possession phase per definition. Ball possession phases that end with the halftime-, or final-whistle or a referee ball are excluded from our analysis.

In addition to these general rules, we developed the following transition-related definitions in consultation with match-analysis experts: A *defensive transition phase* is defined as the time-window when a team loses ball possession, but is not yet into their ideal defensive formation. Within these defensive transition phases,

a team conducts *counterpressing* if at least one player exerts (spatio and/or temporal) pressure on the ball carrier, or on the opponents close to the ball.

Note that there exist many different definitions for pressing: StatsBomb[11] defines pressing as a defensive player being within a five-yard radius of the ball-carrying opponent.[12] Very similarly, a more granular and non-binary definition, aggregating the pressure of several defensive players, is presented by Andrienko (2017). Based on these pressing definitions, counterpressing could be defined as situations where pressing is exerted immediately after a ball possession change (Navarro and Javier 2018). Both of these rule-based definitions are used as a baseline model for our investigation.

However, according to the match-analysts involved in this project, being close to the player in ball possession is not the only way to exert pressure. Attacking or blocking the easiest pass options could, for instance, also be seen as applying pressure.

To quantify the *success of counterpressing*, we consider it as successful if the ball is regained within five seconds and shots and goals, scored or received, are accredited to the previous counterpressing phase if they occur within the following 20 seconds.

From hereon the game is split into ball possession phases which could start and end either with an in-play ball possession change or a stoppage such as a set-piece. Note that the set-up of the in-play ball possession change, such as the defensive transition, might not be the only influence factor on the defensive reaction time—it can also occur due to short, uncontrolled ball possessions or risky passes of the opponent. Any ball possession phases that either start with a set-piece or end with a stoppage in play will not be considered further. Fig. 1 shows a heatmap displaying the occurrences of

[11] Statsbomb is a football event data provider based in the UK, https://statsbomb.com/, accessed 12/17/2020.

[12] StatsBomb event data, including the pressing tag, are can be accessed for many professional leagues https://statsbomb.com/data/.

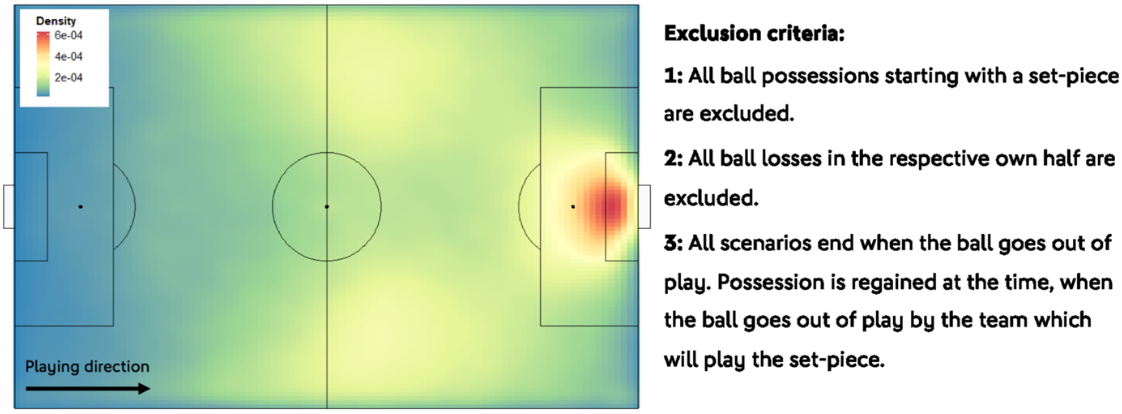

Fig. 1 Overview of where on the pitch turnovers happen most frequently (from the perspective of a team playing from left to right)

transition situations related to the pitch. It indicates, that most turnovers happen in the opposing half, especially near the sidelines. Ball possession changes due to a ball going out of bounds are added to the area touching the sideline. Easily identified is the high proportion of turnovers in the opponent's six-yard box. This is likely because both saved shots and shots missing the goal wide are counted as a change in possession as soon as the goalkeeper receives the ball.

2.2 Supervised machine learning set-up

2.2.1 Hand-crafted labeling of defensive transition situations

Since the rule-based approaches to detect counterpressing we investigated lead us to an insufficient accuracy (see Sect. 3.1), we conducted a manual tagging procedure with trained student-analysts. It was their task to label situations with a detectable counterpressing strategy. In total, out of 11, 108 relevant defensive turnovers, 3, 196 situations were labeled as counterpressing. The labeling was conducted for the first eleven Bundesliga-matchdays of the 2018/2019 season from the perspective of the home team. The percentage of counterpressings detected per transitions differs significantly per team. Borussia Mönchengladbach presented the highest value (40.07%), whereas only 21.80% of Hannover 96's transitions have been labeled as counterpressing. The aggregated outcome of the labeling process per team of the German Bundesliga is displayed in Table 7 in the Appendix A.

To quantify the inter-labeler reliability, 20 matches were labeled by three different students. To compute the pairwise accuracy for each defensive turnover, we checked if both students had identified countpressing in the following two seconds. This yielded a pairwise accuracy of 82.01%, i.e. in 82.01% of defensive turnovers both students agreed on the nature of the actions following a turnover.

As additional information, the experts tried to detect the exact start and end-frame of the respective transition situation. The average duration of all transitions phases is 9.34 s, 9.89 s for counterpressing, whereas all non-counterpressing turnovers took in average 9.11 s.

2.2.2 Expert-based feature extraction

We defined a list of 134 features that aim to characterize the transition. The features describe the location of a ball possession change, several relevant factors describing both teams' exact positioning at the time of turnover and their movements in the first two seconds immediately after the ball loss. A time-window longer than two seconds was problematic, because it would cut off too many situations where the ball possession changed within that time.

A teams' decision to conduct counterpressing is heavily influenced by the situation of the ball possession itself. To take this into consideration, all features are also calculated at the moment of the ball possession change. According to football experts, turnovers without the chance to counterpress are often characterized by immediate clearances or aerial duels. Therefore, we included the ball position, the ball height, and the individual ball possession time (Link and Hoernig 2017)—describing the time a player of the ball possessing team was in direct control over the ball. The involved football experts suggested, that counterpressing is often characterized by achieving a local compactness close to the ball. We aimed to cover this with several metrics measuring the regaining team's positioning around the ball. For instance, we use the team's covered area, global and local stretch indices (Bourbousson and Carole Sève 2010; Santos et al. 2018) as features in our model. A team primarily aiming to defend their own goal after losing possession does this usually with high-speed towards their own goal, whereas counterpressing requires often only players close to the ball to attack their opponents with a high speed towards the ball carrier. This is addressed by calculating several speed-values and considering each team's average position, the so-called team-center (Bourbousson and Carole Sève 2010; Andrienko 2017). In contrast to a more conservative transition strategy, counterpressing's primary objective is not to place many players in a compact unit behind the ball quickly, but rather to defend more aggressively up the pitch. Therefore, we calculate both the number of players in front and behind the ball, as well as their respective compactness. Although the pressing definition from Andrienko (2017) was not sufficient as a stand-alone rule-based counterpressing detection criteria (see Sect. 3.1), it is incorporated in various features of our model.

All features were discussed, consolidated and steadily improved within workshops and based on several steps of evaluation of the detection. A detailed list and description of the features is presented in Table 1, a video describing some of the features can be accessed here.

2.3 Model training

2.3.1 Detection of counterpressing as a supervised machine learning task

We trained several classification algorithms based on the 11, 108 labeled defensive turnover situations from 97 matches fulfilling our inclusion criteria (see Fig. 1).

Table 1 The extracted features that are used for counterpressing detection. Features used in both dimensions of pitch coordinates (horizontally and vertically) and for different time points after the initial ball possession change are listed only once

Feature	Definition
Turnover position	Location (x,y-coordinate) of the ball at the timepoint of the ball possession change (BPC), hereafter 0 s.
Ball height	z-coordinate of the ball tracked at several timepoints after (BPC) (0 s, 1 s, 2 s)
Distance team	Distance closest player to the ball calculated at different timepoints after BPC (0 s, 1 s, 2 s) and for each team (opposing, regaining).
Number players close to ball	Number of players in several circles (10 m, 20 m, 30 m) around the ball counted at different timepoints after BPC (0 s, 1 s, 2 s); calculated for both teams separately.
Team center	Average position of all players of each team (goalkeeper excluded) calculated for both dimensions (x,y) and at several time points after BPC (0 s, 1 s, 2 s); calculated for both teams separately.
Covered area team	Team's covered area measured as the widest distance between two players in two dimensions (x,y) and at several time points after BPC (0 s, 1 s, 2 s); calculated for both teams separately.
Players closer to ball	Number of players from a team closer to the ball than next player from the other team calculated at different time points after BPC (0 s, 1 s, 2 s); calculated for each team separately.
Speed closest player	Speed of player closest to the ball calculated at different time points after BPC (0 s, 1 s, 2 s) and for each team separately.
Duration previous ball possession	Duration of the previous ball possession phase; sequences where the ball was out-of-play (*ball status*) have been excluded.
Speed team	Average speed of each team at different time points after BPC (0 s, 1 s, 2 s); calculated for both teams.
Individual ball possession (absolute)	Total individual ball possession time of any player from the regaining team within up to the first 3 seconds after BPC as defined in Link and Hoernig (2017).
Individual ball possession (relative)	Total individual ball possession divided by the length of the considered time window, so either divided by 3 seconds or by the duration of the ball possession change if it is shorter.
Players in front of the ball	Number of players for each team that are further away from the own goal center than the ball at different time points after BPC (0 s, 1 s, 2 s); calculated for both teams separately.
Players behind the ball	Number of players for each team that are closer to the own goal center than the ball at different time points after BPC (0 s, 1 s, 2 s); calculated for both teams separately.
Global compactness team	Normalized stretch index of each team team (excluding goalkeeper) based on definition in Bourbousson and Carole Sève (2010) calculated at several time points after BPC (0 s, 1 s, 2 s); calculated for both teams separately.

Table 1 continued

Feature	Definition
Local compactness	Normalized local stretch index of different player groups (3,4 and 5 player closest to the ball) at different time points after BPC (0 s, 1 s, 2 s) as shown in Santos et al. (2018); calculated for both teams separately.
Compactness in front of the ball	Normalized local stretch index of all players of each team that are further away from the own goal center than the ball at different time points after BPC (0 s, 1 s, 2 s); calculated for both teams separately.
Compactness behind the ball	Normalized local stretch index of all players of those players of each team that are closer to the own goal center than the ball at different time points after BPC (0 s, 1 s, 2 s); calculated for both teams separately.
Pressure regaining team on ball	Normalized pressure (definition Andrienko (2017)) the regaining team conducts on the ball at different time points after BPC (0 s, 1 s, 2 s).
Pressure on ball possessing player	Normalized pressure (definition Andrienko (2017)) the regaining team conducts on the ball possessing player at different time points after BPC (0 s, 1 s, 2 s).
Effective playing space	Defines the space covered by the convex hull of all players of a team (excluding goalkeeper) calculated at different time points after the BPC (0 s, 1 s, 2 s) as shown in Santos et al. (2018); calculated for each team separately.

We split the labeled data-set (75% training data, 25% test data) by taking randomly 25% of all transitions out of every match to avoid over-representing teams, scores, or results.

We used the above defined features (section 2.2.2 or Table 1) and evaluated the best performing models on our set of test data.

Among different basic-models, we applied extreme gradient boosting (hereafter referred to as XGBoost), a scalable tree boosting system, introduced by Chen and Carlos (2016), which outperformed traditional machine learning algorithms in numerous applications (Li et al. 2019; Liu et al. 2020; Zhang et al. 2020). For our investigation, we want to point out three major advantages of XGBoost: (a) To make use of our wide set of features without taking the risk of overfitting, an additional regularization term is added to the loss function. Additionally, (b) XGboost is a scalable machine learning model, which can be extended seamlessly with more data or more features being available. Furthermore, (c) no normalization is required.

Before training our model, a set of hyperparameters (shown in Table 2) has to be defined. As presented in Wang (2019), we applied Bayesian tree-structured Parzen Estimator hyperparameter optimization approaches to obtain the highest possible accuracy and avoid overfitting. By using tree-structured Parzen Estimators (Bergstra 2011)

Table 2 Hyperparameter-selection of the XGBoost models

	Hyperparameter	Description	Range	XGB-M1	XGB-M2
1	Learning rate	Controls the step size used per update	[0, 1]	0.045	0.05
2	Max depth	Limits the depth of the tree	$[0, \infty)$	5	7
3	Subsample	Controls number samples applied to the tree	(0, 1]	0.419	0.279
4	Min child weight	Controls instance weight of a node	$[0, \infty)$	0.855	1
5	nrounds	Limits number of iterations	1–400	400	400
6	Class balancer	Controls the balance of negative and positive weights (Number of negative cases / Number of positive cases)	$(0, \infty)$	1	2.5

as the surrogate model in Bayesian optimization (Dewnacker et al. 2016), we reduce the running time of hyperparameter tuning and achieve better scores on the testing set. Note that the hyperparameter *nrounds* was set to a maximum of 400 iterations.

To further guarantee the stability of our model and avoid overfitting, we applied five-fold cross-validation on the training data.

As described in Sect. 2.1.2, we also implemented two rule-based baseline models that serve as a benchmark for our detection. A naive approach defines counterpressing as follows: whenever one or more players are within a five-yard radius around the ball carrier during the first individual ball possession phase following a turnover, it is classified as a counterpress (hereafter referred to as *naive rule-based approach)*.[13] The second approach (hereafter referred to as *Andrienko-approach*) defines counterpressing as all turnovers whenever the first player in ball possession receives pressure exceeding a certain threshold according the pressure-definition in Andrienko (2017), whereby the final threshold of 0.74 was obtained by maximizing the F_1-score on the training set.

2.3.2 Effectiveness for counterpressing and fast possession regains

In order to define some success metric of a transition phase, the low scoring nature of football causes us to examine the following actions more granularly, rather than just checking whether they are followed by a goal. For both cases—successful and unsuccessful ball recoveries through counterpressing—we extracted taken shots, expected goals[14] and actual goals following a transition phase. To investigate this issue, two definitions had to be made: Which ball recovery latency of a counterpressing strategy should be considered as successful, and for how long a defensive and an offensive action would be accredited to the previous ball recovery (strategy). As a starting point for a potential threshold for successful counterpressing, a first indicator is given by Pep Guardiola's five second rule. We queried relevant video scenes with possession regains after 3, 4, 5, 6, 7 and 8 seconds and discussed them with a group of professional match-analysts. The same procedure was conducted to investigate the follow-up goal-scoring opportunities. Here scenes with shots 10, 15, 17, 20 and 25 seconds after the initial ball loss were discussed. Through this procedure, we finally agreed on the definitions described in Sect. 2.1.2.

[13] This approach has been suggested at the *Statsbomb Innovation in Football Conference* by Will Gürpinar-Morgen (https://statsbomb.com/2018/05/how-statsbomb-data-helps-measure-counter-pressing/), accessed 11/11/2020.

[14] The "Expected Goal" (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. Details regarding the xG-model used can be found in Anzer and Bauer (2021).

Table 3 Statistical evaluation of the counterpressing outcome

	Model	Precision	Recall	F_1 -score	AUC
1	XGBoost (Model 1)	0.72	0.63	0.67	0.874
2	Logistic regression	0.69	0.51	0.59	0.841
3	Random forest	0.74	0.55	0.63	0.867
4	XGBoost with class balancer (Model 2)	0.60	0.80	0.69	0.865
5	Naive rule-based approach	0.31	0.87	0.46	0.602
6	Andrienko-approach	0.37	0.37	0.37	0.568

3 Results

3.1 Statistical evaluation

3.1.1 Detection of counterpressing

With the above described supervised machine learning set-up we are able to detect counterpressing situations with sufficient accuracy for practical applications (see also Sect. 3.3). Table 3 shows a statistical evaluation of the different models, from which XGBoost performed the best. Per team and per match, we detect around 20 to 30 counterpressing situations, out of around 90 to 200 transition situations.

With the highest overall area under the curve (AUC) the above presented optimization (Table 3, row 1) is best suited for the long term analysis of several seasons with the goal to identify trends and underpin practitioner rules (*RQ1*-*RQ4*, Sect. 3.2). The XGBoost model with a class balancer (Table 3, row 4) has a higher recall of 80% with still an acceptable false positive rate. Thus, it can be applied for specific performance analysis of either the own match or several matches of the next opponent (PA, Sect. 3.3)—where match-analysts spend a lot of time analyzing video footage either way. The optimal hyperparameters used for both models can be found in Table 2. When examining the results of the two baseline approaches, they exhibit a very low overall accuracy. The naive rule-based approach (Table 3, row 5) classifies 72.41% of all turnovers as counterpressing, which leads to a high recall but also a large number of false positives. For the Andrienko-approach, selecting the threshold by optimizing the F_1-score lead to a more realistic percentage of predicted counterpressing situations in the test set (25.65%), but are, nevertheless, significantly outperformed by either machine learning approach.

Another advantage of the XGboost approach is that the individual influences of our rich feature set can be somewhat quantified and interpreted by analyzing the respective SHAP-values.[15] The naming was both coined by their originator Lord Shapley, who introduced them in the context of cooperative game theory (Roth and Thomson 1988), but also by Lundberg and Su (2017), who used the concept to interpret the features for machine learning models. In comparison to traditional feature importance models (e.g. gain or Saabas method), SHAP-values present a consistent and locally accurate

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

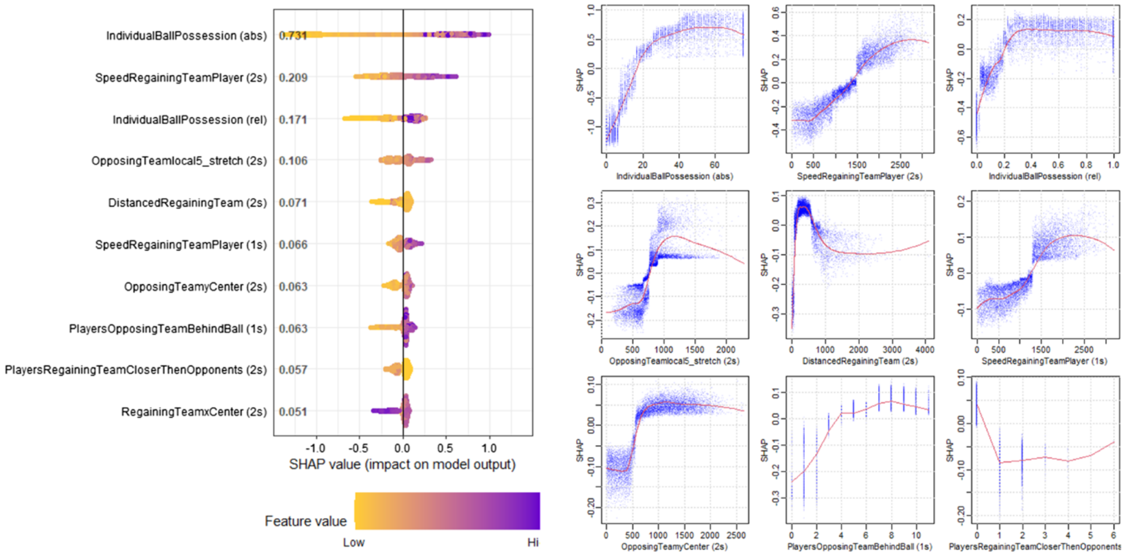

Fig. 2 Feature influence to the counterpressing prediction based on SHAP-values

method to identify the individualized feature contribution to machine learning models. This method has been effectively used in different applications (Antipov et al. 2020; Meng et al. 2020; Ibrahim et al. 2020; Anzer and Bauer 2021).

Figure 2 displays the most influential features according to the SHAP values in two different representations. In the left Fig., each dot represents the contribution of the feature to the model, whereas the color-coding describes the value of that feature. Both, the absolute individual ball possession time (*IndividualBallPossession (abs)*) and the speed of the regaining team two seconds after the change in ball possession (*SpeedRegainingTeamPlayer (2s)*), have a very strong and linear impact on the predictions. Besides the fact, that both features have the highest overall influence on the prediction (widest dispersion of the dots on the left part of Fig. 2, the interpretation of the SHAP-values can be expressed as follows: the higher the absolute amount of individual ball possession time within the first three seconds and the higher the speed of the regaining team two seconds after the turnover, the more likely a defensive turnover is classified as counterpressing. The number of opposing players behind the ball after one second (*PlayersOpposingTeamBehindBall (1s)*) influences the prediction in a different way: A high number of players behind the ball increases the chances for a classification of counterpressing, but the relation is clearly non-linear. To get a better idea of the influence, we will have a look at the right part of Fig. 2. The value per feature is now displayed by the x-axis, whereas the model influence is shown on the y-axis. If less than four players are behind the ball, this feature on average decreases the chance of a counterpressing classification. The number of four defenders—almost half of the team—seems to be a decisive threshold. If four or more players are behind the ball, this feature has a positive contribution to the prediction. This not only aligns with the expectation of the practitioners, but also led to a very valuable discussion among the professional analysts. A more complex relation is shown by the local stretch index of the five closest player to the ball of the opposing team two seconds after the ball possession change (*OpposingTeamlocal5 stretch (2s)*). After a steep increase starting at 600 cm, the influence of that feature reaches its maximum at roughly $1,000$ cm (see

Table 4 For all 4118 considered matches of the German Bundesliga and 2nd Bundesliga, this table shows the probability of shots and goals per team following counterpressing situations

Strategy	Goals scored %	Goals conceded %	Shots for %	Shots against %
Counterpressing (all)	0.35	0.78	3.22	5.15
Unsuccessfull counterpressing	0.16	1.02	1.83	6.7
Successfull counterpressing	0.75	0.25	6.27	1.76

right plot in Fig. 2) but decreases afterwards. This indicates that a higher stretch index (lower compactness) of the opposing team after two seconds increases the chances for counterpressing.

Excluding features with little to no influence according to the SHAP-values, did not improve neither the F_1-score nor the AUC of our prediction on the test data-set.

3.1.2 Effects of counterpressing

Table 4 shows the outcome of counterpressing regarding goals and shots scored or conceded within 20 seconds. If one is successful, i.e. wins back the ball within five seconds the chance of scoring increases tremendously, but if unsuccessful one is far likelier to concede a goal.

It is no surprise, that the chances to either shoot or score are significantly higher when counterpressing was applied successfully, since it implies the crucial attacking advantage of having gained the possession of the ball.

While using goals scored versus conceded would theoretically be ideal to measure success, the low scoring nature often prevents us from doing so. Therefore, we use shots to compare teams and coaches. Nevertheless, according to the experts, looking at both shot- and goal- (or even expected goal-)balance is a very valuable key-performance-indicator for counterpressing.

3.2 Subject-specific evaluation of six seasons of German Bundesliga data

In the following section, we use our quantitative results as a baseline for a qualitative, subject-specific evaluation and interpretation with the involved experts.

3.2.1 Lessons learned about defensive transitions (*RQ1*)

A common procedure when analyzing a team's transition strategy is looking at the easily acquirable defensive reaction time (Vogelbein et al. 2014). This, however, comes with the drawback that it is not able to distinguish between situations with intentional counterpressing behavior and noise. Note that in non-trivial defensive turnover situations typically a team can choose between falling back or conduct counterpressing. But for the purpose of this study, the distinction between fallback and other non-counterpressing situations was removed for the sake of simplicity. An additional analysis of the expert-based labeling (see Appendix A) showed, that around 62.5%

of all ball losses fulfilling the inclusion criteria cannot be assigned to either defensive strategy (counterpressing or fallback). The sheer number of these situations—with a very short defensive reaction time (on average 7.83 s) and without any defensive tactical choices being detectable—significantly influences the defensive reaction time when applied to all transitions. Further analysis (see Appendix A) shows that specifically turnovers with very short individual ball possession times fall in this category. Their exclusion presents a crucial step for a better understanding of transition situations.

In general, counterpessing is not always advantageous and needs to be executed well. Although, the expectation of some practitioners (invigorated by Jürgen Klopp's statement *"counterpressing is the best playmaker"*) may be different, it is still intuitive that the team in possession of the ball has a higher chance to perform a successful offensive action (e.g. a shot or goal; see Table 4, row 1). However, this is an average over all teams independent of their skill level and as we will later see, there are some teams/coaches that were able to apply counterpressing so successfully that they even ended up with a positive shot balance by creating more shots after counterpressing than conceding. In order to properly assess the risk versus reward nature of counterpressing, one would ideally compare it to its strategic counterpart falling back. But even then, one would need to carefully address potentially confounding variables describing the original situations, since, as it seems from the feature importance discussion, the situation typically dictates the strategic response. This goes, however, beyond the scope of this study, but could be the ground for interesting future work. Additionally, since all non-counterpressing situations consist of myriad of different circumstances, they do not serve as reasonable baseline to effectiveness of counterpressing.

3.2.2 Define and statistically underpin objective Benchmarks (*RQ2*)

Based on the above explained definitions and trained prediction models, several quotients and ratios were discussed with the experts. Aggregated on a season level, we analyzed the correlation with a team's final ranking. Our detection provides several different metrics with a significant correlation to success. These performance indicators can be calculated per match, per match phase (e.g. one halftime) or even per turnover, which allows practitioners to objectively compare their teams performance with pre-defined benchmarks. With a negative Pearson correlation, the ratio of successful counterpressing-situations to the total number of transitions predicts a team's final ranking the best ($r = -0.44$). Another metric, correlating with the final ranking that was very valuable to the experts due to its direct interpretability, was shot and goal balance ($r = -0.36$ shots, $r = -0.42$ goals)—describing, whether more shots or goals are taken after successful counterpressing, than are conceded after failed attempts. However, these metrics should be used carefully since they are based on small sample size and could contain confounding effects with the overall offensive or defensive qualities of a team. Loosing the ball, increases the probability of an opponent conducting an offensive action. Thus we present an effective strategy to monitor the outcome of counterpressing strategies, such as several performance indicators that enables coaches to objectively benchmark a team's defensive transition behavior.

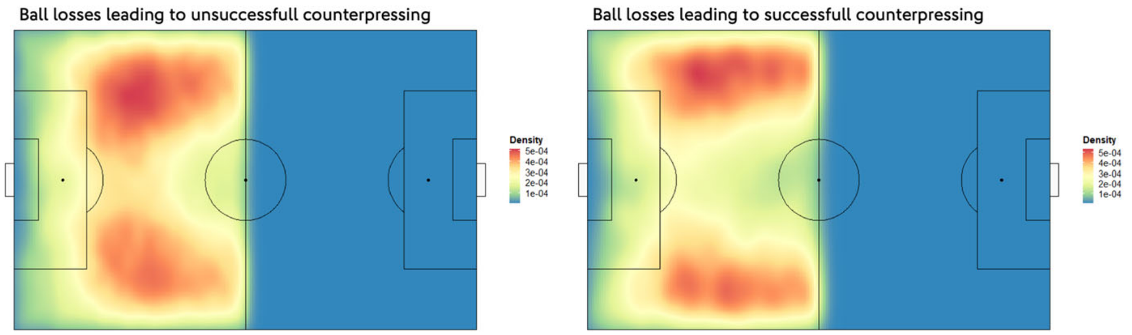

Fig. 3 Heatmap of ball possession change locations before unsuccessful counterpressing situations (left) and sucsessfull situations (right)

3.2.3 Approve established rules of thumb (*RQ3*)

A widely spread rule of thumb is that counterpressing is ideal after ball losses close to the sideline or close to the corners in the opponent's half. Fig. 3 presents two heatmaps that underpin this statement. Secondly, we want to examine, whether a numerical superiority of players close to the ball increases the chance of a successful counterpress as assumed by many experts. For that we examine the 109, 852 detected counterpressing situations satisfying the inclusion criteria. Whenever the team out of possession has a numerical superiority within a 10 m radius around the ball, at the time of the turnover, they regain ball possession within 5 seconds 36.2% of the time, compared to only 30.2%, when the other team has more players in that area. This indicates that the rule of thumb has some truth to it, but is far from the only influencing factor, deciding whether a counterpress will be successful.

3.2.4 Compare Team's and Coach's Counterpressing in German Bundesliga (*RQ4*)

Further investigations highlight to which extent teams use completely different defensive transition strategies. We investigated the coaches that were expected to have a pronounced counterpressing-behavior by the experts. The respective ranking among all coaches finishing a full season can be found in the Appendix C.

Table 5 also gives a first indication of which further aspects could be considered. Teams ending up in the top 5 of a season perform above average in almost all defined metrics. No consistent tendency over the considered seasons is detectable: games do not get more intensive in terms of total in-play transitions per match, nor are there significant changes in teams average counterpressing behavior. FC Bayern München has the best goal balance after counterpressing, which might be heavily influenced by their offensive efficiency. Jürgen Klopp and Ralf Ragnick performed extraordinary well in terms of ending up with more created than received shots after attempting counterpressing. Given that Jürgen Klopp ended only seventh place in one of his two considered seasons, this should be seen as an outstanding performance. Over the course of a whole season only nine coaches achieved a non-negative shot balance within 20 s after their own counterpressing—the average final ranking of the respective teams was four. Note that teams playing at home tend to conduct counterpressing slightly more often than away teams. Considering only home-teams, our model classifies 27.24%

Table 5 Comparison of counterpressing-related performance indicators in Bundesliga. The colored arrows display the rankings within the respective groups compared to the average (teams by ranking; all teams by seasons; teams and coaches). For the columns presenting the shot and goal balance we used colored dots for the teams and coaches, showing whether the total outcome per match is positive, neutral or negative

Team	Coach/Match	Matches	All Transitions: per Match	All Transitions: %CP	All Transitions: %Successfull CP	Off. / Def. Balance: Goals	Off. / Def. Balance: Shots	Succesful Counterpressing: All per CP	Succesful Counterpressing: %Shots	Succesful Counterpressing: %Goals	Unsuccessful Counterpressing: All per CP	Unsuccessful Counterpressing: %Shots	Unsuccessful Counterpressing: %Goals
All Teams	All Coaches	4118	115.60	23.08%	7.25%	-0.11	-1.15	31.42%	2.65%	1.10%	68.58%	7.51%	1.13%
Top 5 Teams	All Coaches/Season	1142	114.99	23.74%	7.80%	-0.04	-0.28	32.85%	10.94%	1.42%	67.15%	6.85%	0.94%
Bottom 5 Teams	All Coaches/Season	1145	116.62	22.42%	7.03%	-0.16	-0.65	31.35%	9.78%	0.95%	68.65%	8.10%	1.32%
All Teams	Season 2013/2014	610	98.72	24.50%	7.47%	-0.12	-0.47	30.50%	10.94%	1.07%	69.50%	7.59%	1.16%
All Teams	Season 2014/2015	612	126.33	21.80%	7.19%	-0.07	-0.28	32.99%	10.36%	1.10%	67.01%	6.61%	0.90%
All Teams	Season 2015/2016	612	121.58	22.56%	7.33%	-0.09	-0.50	32.49%	8.97%	1.10%	67.51%	7.01%	1.00%
All Teams	Season 2016/2017	612	121.79	22.28%	7.39%	-0.08	-0.49	33.16%	8.75%	1.09%	66.84%	7.02%	0.99%
All Teams	Season 2017/2018	612	116.32	23.46%	7.16%	-0.09	-0.59	30.51%	10.56%	1.22%	69.49%	7.74%	1.03%
All Teams	Season 2018/2019	612	112.42	23.97%	7.15%	-0.18	-0.64	29.81%	11.68%	1.04%	70.19%	8.32%	1.41%
All Teams	Season 2019/2020	448	110.65	23.53%	7.04%	-0.20	-0.69	29.91%	11.12%	1.09%	70.09%	8.53%	1.55%
FC Bayern München	All coaches	229	112.21	24.63%	8.02%	0.06	-0.24	32.56%	11.98%	2.04%	67.44%	7.10%	0.66%
FC Bayern München	Pep Guardiola	102	109.47	25.24%	8.61%	0.02	-0.21	34.10%	11.34%	1.46%	65.90%	7.00%	0.65%
FC Bayern München	Jupp Heynckes	26	109.73	22.43%	6.98%	-0.08	-0.31	31.09%	12.56%	1.51%	68.91%	7.48%	1.13%
FC Bayern München	Carlo Angelotti	40	118.80	23.84%	7.60%	0.13	-0.68	31.86%	7.48%	2.22%	68.14%	6.99%	0.39%
FC Bayern München	Hansi Flick	15	123.87	26.53%	8.56%	0.07	0.00	32.25%	17.61%	2.52%	67.75%	8.38%	0.90%
Borussia Dortmund	All Coaches	229	119.57	24.87%	8.51%	-0.04	-0.16	34.20%	10.56%	1.29%	65.80%	6.29%	0.87%
Borussia Dortmund	Jürgen Klopp	68	129.00	23.51%	8.54%	-0.04	0.26	36.32%	11.62%	0.80%	63.68%	5.26%	0.69%
Borussia Dortmund	Thomas Tuchel	68	121.59	24.73%	8.84%	0.00	-0.16	35.75%	9.58%	1.64%	64.25%	6.16%	0.91%
Borussia Dortmund	Lucien Favre	58	105.07	26.98%	8.34%	-0.10	-0.34	30.90%	10.24%	1.38%	69.10%	6.34%	1.14%
Mainz 05	Thomas Tuchel	34	95.82	24.31%	7.98%	-0.18	-0.50	32.83%	11.15%	1.15%	67.17%	8.65%	1.69%
All Teams	Thomas Tuchel	102	113.00	24.61%	8.60%	-0.06	-0.27	34.93%	9.99%	1.51%	65.07%	6.88%	1.14%
RB Leipzig	All Coaches	127	126.69	22.13%	7.84%	-0.06	-0.02	35.41%	10.71%	1.43%	64.59%	6.00%	1.13%
RB Leipzig	Julian Nagelsmann	25	110.60	22.21%	7.88%	-0.12	-0.20	35.50%	11.47%	0.92%	64.50%	7.58%	1.26%
RB Leipzig	Ralf Ragnick	34	131.47	21.72%	7.00%	-0.03	0.56	32.23%	14.38%	1.28%	67.77%	3.95%	0.76%
TSG Hoffenheim	Julian Nagelsmann	116	111.01	23.98%	7.46%	-0.03	-0.53	31.09%	13.44%	1.77%	68.91%	8.98%	0.99%
All Teams	Julian Nagelsmann	141	110.94	23.67%	7.53%	-0.05	-0.48	31.82%	13.07%	1.61%	68.18%	8.76%	1.03%

of all included defensive transitions as counterpressing, which is roughly in line with the labeled training data, that was conducted only on home teams (28.77%).

Figure 4 shows a shortlist of teams' counterpressing outcomes. Note that, since on average more shots/goals are conceded than created when counterpressing, all four axis are scaled differently. For both sub-figures, teams on the upper left, including, for example, 1. FC Nürnberg[16], Hannover 96[17] and SV Darmstadt 98[18] in the left figure, perform worse. Teams on the bottom right tend to generate more shots/goals and allow fewer while counterpressing. Teams with high values in the top right quadrant like Borussia Mönchengladbach[19], FC Augsburg[20] or TSG Hoffenheim[21] seem to employ risky defensive transition strategy, by both creating and also receiving many shots after their own counterpressing.

As a general recap of the Bundesliga analysis, we would like to point out that teams use significantly different transition strategies (*RQ4*). The experts' expectations of which coaches use counterpressing more often and/or more efficiently were underpinned by the results.

3.3 Proof of concept

The central objective of this study is to automate the detection of counterpressing situations. This helps match-analysts in their daily processes by saving them time, but also by providing objective and comparable benchmarks. First, we describe the general

16 Highest in the left figure.

17 Green circle with the inscription *96*.

18 Blue circle around a white lily.

19 Black and white hatched diamond logo roughly in the center of both plots.

20 Fifth highest in the left figure.

21 Second from the right in the right figure.

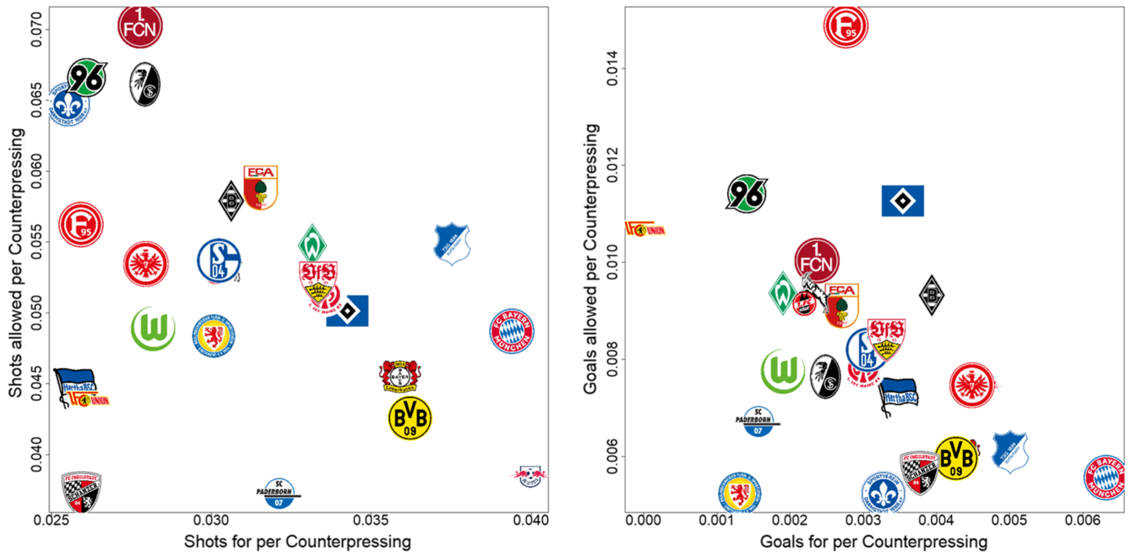

Fig. 4 For both shots (left figure) and goals (right figure) teams are displayed depending on their percentage of counterpressings leading to either shots or goals for (x-axis) or against (y-axis). Team values are aggregated across all considered seasons

set-up, whereas a prototypical application for two exemplary matches of German national teams is conducted.[22,23] Based on the results described above, we are now able to provide match-analysts with two files fully automatically in virtually real-time: First, they receive a list of all detected counterpressing situations. To integrate this efficiently into their ecosystem, the files are produced in different file-formats, which can be imported into their video-analysis tool of choice (e.g. Hudl Sportscode,[24] Stats Edge Viewer[25]). Such tools basically help to handle tags or labels in combination with the video footage. Figure 5 shows how this eliminates the usual process of the match-analysts labeling the videos manually in an exemplary tool (here Hudl's Sportscode). Usually, match-analysts use these tools to tag important situations live during the match but also in detail post-match for opponent analysis. Once a match or parts of it are tagged, the tool allows the analyst to output the tags either as a video-playlist or as an xml-file, containing the category and the time-frame of each tag. Depending on the coaches needs, the outcome is either presented as a video-playlist or a quantitative report giving an aggregated overview—which also is typically produced manually. An automatically generated counterpressing-playlist for the U21 match based on our prediction can be viewed here.

Second, we automatically provide coaches and analysts a counterpressing match-report with visualization after the respective match or entire season reports. For the U21 match an excerpt of the automated match report is presented in Fig. 6. Only two shots occurred within 20 seconds after either team counterpressed. However, these two shots after unsuccessful counterpressing attempts by the German team lead to two goals, which were decisive for the total match-outcome (Germany vs. Belgium, final score

[22] Germany against Northern Ireland; *19th of November 2019, Commerzbankarena Frankfurt*.

[23] Germany U21 against Belgium U21; *17th of November 2019, Schwarzbaldstadion Freiburg*.

[24] https://www.hudl.com/products/sportscode, accessed 06/20/2020.

[25] https://www.statsperform.com/team-performance/football/stats-edge/, accessed 06/20/2020.

Fig. 5 Integration of the counterpressing-analysis into the match-analysts and coaches daily business. The yellow colored part shows the traditional process, the grey part displays how the automated analysis assimilates seamlessly

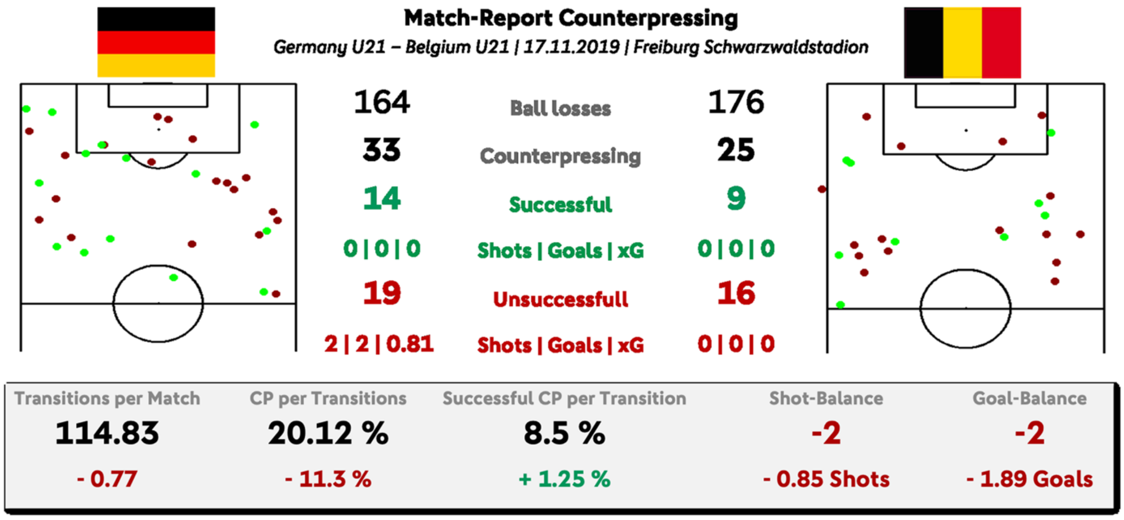

Fig. 6 Automated match report for the match Germany U21 against Belgium U21 in Freiburg. The pitch visualizations shows the positions of ball losses leading to a successful (green) or (unsuccessful) counterpressing. Whereas the absolute values are presented in the middle block, a benchmark of the central performance indicators against Bundesliga average is shown in the lower grey box

2–3). Whereas situations leading to goals are analyzed on a highly detailed level by coaches and analysts either way, they arrived at the same conclusion. Nonetheless, this report helps coaching staffs to evaluate whether the German team had a general bias in their defensive transition behavior or whether the two goals happened as a consequence of extraordinary opposing actions or defensive errors from the German team. In this case, bad counterpressing behavior had a significant stake in the origin of both goals. Another outcome was that many successful counterpressing situations (1.23% above Bundesliga average) did not end up in a single shot.

The practical implementation of this study was prototyped by the example of two recent matches of the German national teams. After each match, both the Sportscode-xml file and the automated match-report sheets were produced and shared with the experts for their post-match-analysis process. Additionally, to validate the report, the

Table 6 Statistical expert-evaluation of the counterpressing outcome. For the defensive transition column, the number in brackets displays the number of scenes excluded by our criteria (see Fig. 1)

Match	Team	Defensive transitions	Counterpressing detected	Manually excluded	Additionally detected
GER-NIR	Germany	153 (85)	25	6	0
GER-NIR	Northern Ireland	155 (95)	24	3	2
GER-BEL	Germany	164 (123)	33	5	0
GER-BEL	Belgium	176 (87)	25	2	1

matches were manually analyzed and provided a ground truth to compare our results with. The overall results are shown in Table 6.

For the first game 153 relevant defensive transition situations of the German national team were queried by the above defined rules. Out of these, 25 scenes were detected as counterpressing from which 6 were manually excluded from the final counterpressing playlist. For 164 defensive transitions from the second match 33 situations were detected as counterpressing and analysts ruled out 5 manually. The manually excluded scenes were discussed with all experts and it turned out that different definitions and interpretations lead to different labels. In this case study, ten of the eleven manually excluded situations consisted of only one player exerting pressure. Although this fulfills our definition, some of the experts would classify situations with only one player defending actively towards the ball as fallback (definition in Appendix A). All scenes that were additionally labeled by the respective match-analysts (in total five) contained a clearance closely after the initial ball loss and where thus not clearly related to the counterpressing strategy.

Both the Sportscode-xml and the automated match report turned out to be valuable for the coaching staff. Due to the interpretability of the inaccuracies, the experts trust the outcome for further applications. They deemed the results to be sufficient in terms of a practical usage and noted that it saved vast amounts of resources in the pre- and post-match-analysis. The automated match-report (see Fig. 6) allows us to provide an objective comparison, with a flexible benchmark (here Bundesliga average) and therefore provide a new way to approach complex tactical strategies. The shot- and goal-balance are very intuitive and present a direct monitoring of the efficiency of conducted strategies.

4 Discussion

This paper shows that complex tactical strategies, such as counterpressing, can be detected automatically based on synchronized positional and event data. Comparing team's counterpressing strategies objectively and on longer periods of time creates insights that could not have been achieved with traditional methods.

The interdisciplinary cooperation turned out to be a very beneficial factor for this study. In our opinion, such a set-up of competencies is necessary to obtain relevant results. Machine learning techniques are required to detect complex strategies from spatio-temporal data, but also tactical football expertise are inevitable to determine definitions, extract features and evaluate and interpret the resulting outcomes. A key lesson we learned through this study is that both definitions of complex strategies and their reading vary between football experts—this became apparent during an intensive process of expert-supported evaluations. One of our most meaningful key-performance-indicator is the shot-balance after counterpressing. Here, shots are used as a proxy for a successful attack. This is a common procedure in football analytics, however, the approach could be extended by using expected goals (e.g. Anzer and Bauer (2021)) or expected possession values (e.g. Spearman (2018)).

Since there does not exist a comparable approach for detecting counterpressing in the literature, we implemented two naive baseline approaches to benchmark our model against (see Table 3). While the approach based on Andrienko (2017) was originally not designed to quantify counterpressing but rather pure pressure, we build this rule based approach on Fernandez and Bornn (2018), who defined counterpressing as immediate pressure after losing the ball. Hence, it may not be the ideal approach, but due to complete the lack of alternatives in the literature, we use it as a benchmark model. Even though our model outperforms different rule-based baseline models (see Table 3) and the prediction accuracy is sufficient for practical application (see section 3.3), the basic limitation to achieve further accuracy is the inter-labeler reliability of 82.01%. After discussing the definitions, no further steps of consolidation between the labelers were conducted—but we would highly recommend such a step including the strict monitoring of the inter-labeler reliability for similar investigations. However, data labeling is a time-consuming process which cannot be conducted for each occurring philosophy and definition. Furthermore, methodologies that reduce labeling efforts, such as weak supervision, should be implemented on top of general detections as the one presented here to adjust definitions to the specific needs and to improve both the accuracy and the degree of individualization (Ratner et al. 2016, 2017). With an even larger amount of labeled data, one could consider using continuous features, or even the raw positional data of all players instead of features at discrete time points. The application of labeling-support methods could lead to more individualized and accordant labels and thus to a better prediction.

With an even more accurate model using one of the above described approaches, an improved and team-individual success prediction model for counterpressing could support the reflection of teams' decision making processes significantly. Also, the adaption of the counterpressing detection itself to team-specific definitions, provide a huge potential for further investigations.

Vogelbein et al. (2014) evaluated 306 matches of the 2010/2011 Bundesliga season and showed the time it takes to regain the ball also depends of the score at the time. They pointed out that teams with a lead tend to regain the ball slower than the ones that are trailing, and that teams finishing their season in the top third of the table regain ball possession significantly faster than the other teams—especially in drawn and loosing match states. We found that the defensive reaction time, which serves as a baseline for our success definition, typically includes many noisy situations, where no clear strategy is observable. Further lessons learned regarding defensive transitions are described in section 3.2.1 and extended in Appendix A. The high influence of individual ball possession times on the predictions (see Fig. 2) can be attributed to uncontrolled situations without the possibility for either defensive strategy (counterpressing, fallback).

This also shows a limitation of our work, that some of the model's most important features focus on the situation itself and fewer on the strategy conducted thereafter. A possible explanation for this is that besides filtering out noisy situations, the model found that most often the situation dictates the defensive response.

Nevertheless, a tendency of the opponent to play counterattacks, or especially risky passes could lead to many fast ball recoveries independent from the defensive transition strategy. This issue could be considered by either combining this approach with an equivalent offensive transition strategy detection as shown for example in Hobbs et al. (2018) or by including more features of the opposing team or even the raw data of all 22 players and the ball. Not only the question in which situations counterpressing induces a high chance for a possession regain, but furthermore—given a situation where counterpressing is conducted—how likely is it to take/receive a shot when conducting that strategy are of high interest for practitioners. Regaining the ball fast might not be the only objective of counterpressing. Consequently, future investigations should also consider quantifying alternative success definitions, e.g. slowing down the opposing attack, forcing back-passes etc.

Another missing piece which should be investigated further is an accurate selection of fallback situations. Comparing their risk-reward structure to counterpressing situations, could lead to crucial insights by evaluating a teams' decision to counterpress versus falling back objectively. Since different teams may have their own club-specific definitions, our experimental set-up could be applied to arbitrary counterpressing-definitions or even other tactical patterns, as long as they gather sufficient labeled data. An interesting follow-up study, could investigate how many labeled matches would be necessary to achieve a sufficient accuracy depending on the definition. In our case we found 100 labeled matches to be sufficient, but also stress that a high inter-labeler reliability is necessary.

Note that counterpressing is only one example of a complex tactical pattern, that is of interest to match-analysts, but not covered in typical event level data. The needs of match-analysis departments combined with the growing availability and accuracy of positional and event data present a huge potential for task automating approaches.

5 Conclusions

Based on expert-evaluated definitions and hand-crafted labels, we are able to detect counterpressing strategies automatically with a sufficient accuracy in a supervised machine learning set-up. By producing both an understandable match-report and tagging-files suitable for conventional video-analysis software, the integration of the process into a match-analyst's daily business saves a significant amount of time. The outcome helps to analyze the own team's performance and provides helpful information about the next opponent's defensive transition behavior (*PA*).

We can differentiate between intended counterpressing strategies and the many uncontrolled transition situations with short defensive reaction times. This provides not only a better understanding of transitions but also several more granular performance indicators describing defensive transitions (*RQ1*). The respective performance indicators, consolidated by statistical influences and expert opinions, derived interpretable and intuitive metrics (*RQ2*), such as the goal- or shot balance presenting an effective efficiency quantification for the counterpressing strategy—that were not used before but seem to have a huge potential according to the experts. Two of the proven rules of thumb are that counterpressing is more likely to succeed closer to the sidelines and a numerical superiority close to the ball increases the chance of winning it back (*RQ3*). Through analyzing different facets over several seasons we are also able to quantify trends over a large period of time: teams within the German Bundesliga follow appreciably different transition strategies (*RQ4*). Furthermore, successful teams—measured against their final ranking—tend to use the counterpressing strategy more efficiently, giving credence to the notion of declaring it as an offensive strategy (*RQ2*).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10618-021-00763-7.

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

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

Appendix

Appendix A gives additional information regarding the labeling conducted for this study including further tactical explanation of defensive transitions. Appendix B, C, and D show the outcome of the defined counterpressing performance indicator per team, per coach, and per season on the full Bundesliga data-set.

For all columns in Appendix B, C and D, T stands for turnovers, M for matches, CP for counterpressing, S for Shots and G for Goals. Whereas + indicates successful or positive offensive actions and − the other way around, +/− points out the offensive/defensive balance as described above. The first table in Appendix B compares all teams playing in Bundesliga between the 2013/2014 and 2019/2020 (until matchday 26) seasons. The teams are ordered by their average final ranking, the number of matches considered are shown in brackets. In a second table in Appendix C all coaches playing in Bundesliga between the 2013/2014 and 2018/2019 seasons are presented, whereas a third table shows all teams per season ordered first by their final ranking and secondly by the column CP+/T. Only full seasons with 34 matches one and the same coach where considered. The succession is made based on the highest correlating feature with teams final ranking, number of successful counterpressings per transition. Appendix D compares teams on a season-level and is sorted by the respective final ranking which is shown in brackets.

Appendix A

As discussed in Sect. 3.2, the actual labeling was simplified for our analysis. During the expert labeling process (see Sect. 2.2.1), not only counterpressing, but also *fallback* was labeled by the experts. This strategy is said to be the alternative to counterpressing and was defined as a defensive transition phase, where all players' intention is to either react inactively, or move backwards to their defensive line-up without exerting pressure on the ball. The terms *forward-* and *backward-defending* are often used in this context interchangeably to counterpressing and fallback. Figure 7 shows two exemplary situations for both strategies and lists some typical observations for each strategy.

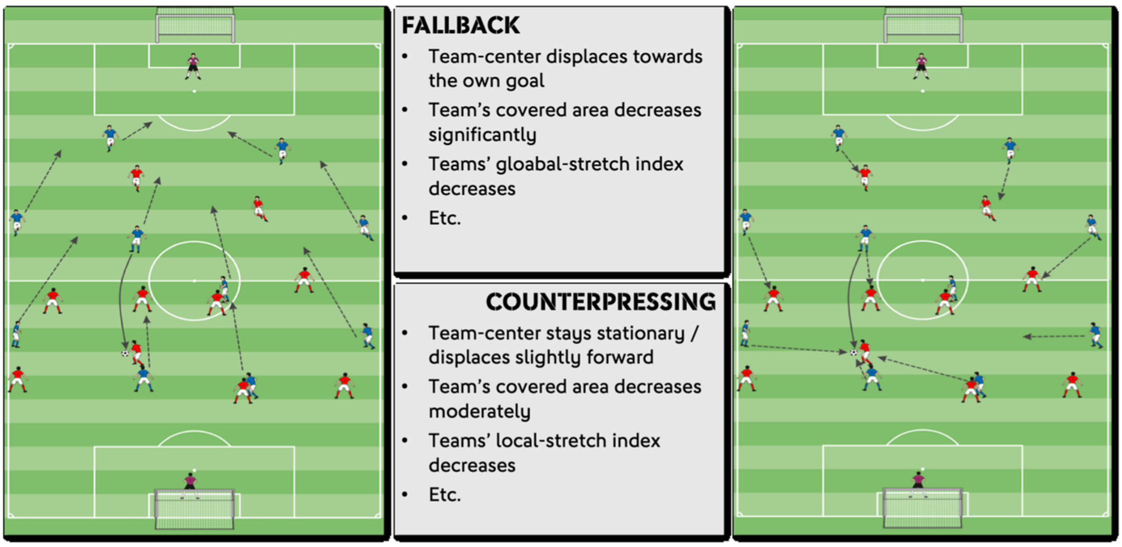

Fig. 7 This figure shows an exemplary ball loss of the blue team, who conducts counterpressing in the right figure and a fallback in the left picture. The pass leading to the turnover is displayed as a solid arrow and the dotted arrows show the expected player movements in the seconds after the ball loss. The metrics *stretch-index* and *team-center* are defined in Table 1

Table 7 Outcome of the manual expert labeling. The first number per entry shows the number of scenes that fulfill the criteria defined in Sect. 2.1.2 and are used for the model building. The total number of tagged scenes is displayed in brackets

Team (Matches)	Defensive Turnovers	Counterpressing	Fallback	Undefined	% CP
All (97)	11,108 (20,928)	3,196 (4,523)	970 (1,268)	6,942 (15,137)	28.77
1. FC Nürnberg (6)	666 (1,167)	178 (228)	94 (111)	394 (828)	26.72
1. FSV Mainz 05 (6)	689 (1,253)	214 (276)	76 (88)	399 (889)	31.06
Bayer 04 Leverkusen (5)	622 (1,092)	175 (262)	57 (63)	390 (767)	28.14
Borussia Dortmund (6)	656 (1,227)	210 (340)	38 (53)	408 (834)	32.01
Borussia Möchengladbach (5)	534 (963)	214 (289)	46 (50)	274 (624)	40.07
Eintracht Frankfurt (5)	656 (1,092)	206 (273)	61 (75)	389 (744)	31.40
FC Augsburg (5)	583 (1,075)	130 (188)	49 (59)	404 (828)	22.30
FC Bayern München (4)	443 (684)	133 (165)	46 (53)	264 (466)	30.02
FC Schalke 04 (5)	623 (1,054)	149 (190)	47 (56)	427 (808)	23.91
Fortuna Düsseldorf (6)	660 (1,233)	186 (233)	74 (96)	400 (904)	28.18
Hannover 96 (5)	610 (1,079)	133 (156)	44 (55)	433 (868)	21.80
Hertha BSC (5)	548 (1,010)	178 (258)	35 (41)	335 (711)	32.48
RB Leipzig (6)	752 (1,298)	196 (246)	43 (47)	513 (1005)	26.06
Sport-Club Freiburg (5)	556 (1,044)	128 (167)	34 (45)	394 (832)	23.02
SV Werder Bremen (6)	667 (1,176)	224 (304)	60 (66)	383 (806)	33.59
TSG Hoffenheim (6)	677 (1,187)	226 (313)	52 (59)	399 (815)	33.38
VfB Stuttgart (5)	488 (946)	143 (192)	50 (63)	295 (691)	29.30
VfL Wolfsburg (6)	678 (1,181)	173 (215)	64 (77)	441 (889)	25.52

Turnovers in which neither of the two strategies can be identified are labeled as *undefined* or *uncontrolled transition situations*. As pointed out in this study, this usually occurs due to short or very uncontrolled ball possession phases (e.g. headers after a corner situation). These situations typically exhibit a very low (relative) individual ball possession time in the three seconds after the turnover. Therefore, it has a high impact on the resulting XGBoost model (see Fig. 2). In total, $20,928$ defensive transition situations from the first eleven matchdays of the 2018/2019 Bundesliga season were labeled based on the definitions formulated above (and in Sect. 2.1.2). The task was to label situations with a detectable strategy with either counterpressing or fallback and not to label any defensive transitions where no strategy was noticeable. Through this procedure the students with a background in football tactics excluded on average 62.50% of all turnover situations by implicitly labeling them as undefined. In total, out of $11,108$ relevant defensive turnovers (after the inclusion criteria), $3,196$ situations were labeled as counterpressing, 970 as fallback, and $6,942$ were explicitly dropped as uncontrolled. Table 7 shows the outcome of the full labeling.

Further evaluation of the labeling outcome showed that more than 95% of all as fallback detected situations start with a goalkeeper catching the ball, what can also be queried solely rule-based on the event data. Due to this fact and because it was our focus to investigate counterpressing situations we decided to exclude the distinction between fallbacks and undefined ball possession changes for all further investigations. Note that fallback situations last on average 18.30 s, whereas undefined situations were the shortest with 7.83 s on average, which again highlights the tremendous influence on the defensive reaction time (counterpressing: 9.89 s; all turnovers average: 9.34 s).

Appendix B

Table 8 Counterpressing per team for Bundesliga seasons 2013/2014 to 2019/2020

Team (Games)	T/M	CP+/T (%)	%CP	S+/ −	G+/ −	% CP+	%CP −	% S+	% G+	% S−	% G−
FC Bayern München (229)	112.21	8.02	24.63	− 0.24	0.03	32.56	67.44	3.90	0.66	7.10	0.80
Borussia Dortmund (229)	119.57	8.51	24.87	− 0.16	− 0.04	34.20	65.80	3.61	0.44	6.29	0.87
RB Leipzig (127)	126.69	7.84	22.13	− 0.02	− 0.06	35.41	64.59	3.79	0.51	6.00	1.13
FC Bayer 04 Leverkusen (229)	125.59	8.13	22.91	− 0.24	− 0.04	35.48	64.52	3.66	0.46	6.96	0.92
Borussia Mönchengladbach (229)	104.44	7.27	24.60	− 0.69	− 0.12	29.54	70.4	3.18	0.41	8.32	1.23
Schalke 04 (229)	113.48	7.06	23.27	− 0.59	− 0.11	30.33	69.67	3.11	0.31	7.64	1.07
TSG 1899 Hoffenheim (228)	115.36	7.48	23.27	− 0.43	− 0.04	32.16	67.84	3.82	0.47	8.02	0.94
VfL Wolfsburg (229)	111.14	6.96	23.88	− 0.49	− 0.15	29.14	70.86	2.93	0.20	6.76	1.07
AVERAGE (4118)	**115.60**	**7.25**	**23.08**	**− 0.51**	**− 0.11**	**31.42%**	**68.58%**	**3.22%**	**0.35%**	**7.51%**	**1.13%**
Hertha BSC (229)	113.14	7.36	23.67	− 0.50	− 0.10	31.09	68.91	2.58	0.31	6.46	1.02%
1. FC Köln (161)	115.12	6.46	22.45	− 0.54	− 0.16	28.79	71.21	2.91	0.24	7.02%	1.18%
1. FC Augsburg (229)	111.49	6.75	22.53	− 0.62	− 0.14	29.95	70.05	3.25	0.28%	8.14%	1.19%
Eintracht Frankfurt (228)	119.78	7.44	23.50	− 0.68	− 0.10	31.66	68.34	2.77	0.37	7.61	1.07%
SV Werder Bremen (228)	115.01	6.87	23.06	− 0.58	− 0.19	29.79	70.21	3.31	0.20	7.84	1.32%
1. FSV Mainz 05 (229)	120.28	7.48	22.95	− 0.45	− 0.13	32.58	67.42	3.40	0.32	7.46	1.15%
Sport-Club Freiburg (195)	111.33	6.14	20.87	− 0.83	− 0.11	29.40	70.60	2.85	0.26	9.07	1.06%
Fortuna Düsseldorf (59)	103.44	6.83	24.56	− 0.78	− 0.37	27.82	72.18	2.94	0.33	8.32	2.50%
1. FC Union Berlin (25)	114.96	5.50	18.65	− 0.48	− 0.24	29.48	70.52	2.99	0.00	7.41	1.59%
VfB Stuttgart (170)	113.80	7.44	23.41	− 0.50	− 0.14	31.80	68.20	3.33	0.33	7.64	1.23%
ING (68)	147.91	7.55	21.10	− 0.35	− 0.06	35.77	64.23	2.59	0.38	5.80	0.88
Hannover 96 (170)	114.62	6.43	22.05	− 1.02	− 0.25	29.14	70.86	2.61	0.14	9.40	1.61%
Hamburger SV (170)	120.76	7.10	22.05	− 0.42	− 0.21	32.21	67.79	3.42	0.35	7.40	1.66%
SV Darmstadt 98 (68)	118.96	6.21	18.90	− 0.88	− 0.04	32.83	67.17	2.55	0.33	9.64	0.78%
1. FC Nürnberg (68)	102.16	6.36	24.37	− 1.06	− 0.19	26.11	73.89	2.78	0.24	9.51	1.36%
Eintracht Braunschweig (33)	104.42	6.99	22.20	− 0.42	− 0.09	31.50	68.50	3.01	0.13	7.06	0.76%
SC Paderborn 07 (59)	115.51	7.31	23.15	− 0.12	− 0.12	31.56	68.44	3.36	0.19	5.56	0.93%

Appendix C

Table 9 Counterpressing per coach for Bundesliga seasons 2013/2014 to 2019/2020. Only full seasons with in total 34 matches per coach are taken into consideration

Coach (Games)	T/M	CP+/T (%)	%CP	S+/−	G+/−	% CP+	%CP−	% S+	% G+	% S−	% G−
Pep Guardiola (102)	109.47	8.61	25.24	− 0.21	0.02	34.10	65.90	3.87	0.50	7.00	0.65
Thomas Tuchel (102)	113.00	8.60	24.61	− 0.27	− 0.06	34.93	65.07	3.49	0.53	6.88	1.14
Jürgen Klopp (68)	129.00	8.54	23.51	0.26	- 0.04	36.32	63.68	4.22	0.29	5.26	0.69
Roger Schmidt (34)	156.29	8.49	18.84	0.91	0.09	45.05	54.95	5.29	0.80	4.00	0.91
Jos Luhukay (34)	104.97	8.27	24.99	− 0.12	− 0.09	33.07	66.93	2.80	0.11	4.86	0.67
Adi Hütter (34)	128.94	8.05	23.27	− 0.59	− 0.09	34.61	65.3	3.63	0.39	8.55	1.05
Markus Gisdol (34)	141.03	8.03	22.34	0.24	0.12	35.95	64.05	4.30	0.56	5.54	0.29
Ralph Hasenhttül (102)	136.35	8.00	21.99	− 0.27	− 0.03	36.38	63.62	2.91	0.56	6.01	1.03
Sandro Schwarz (68)	120.76	7.93	24.05	− 0.26	− 0.18	32.96	67.0	3.70	0.25	6.87	1.28
Carlo Ancelotti (34)	120.03	7.69	23.52	− 0.56	0.15	32.71	67.29	2.60	0.83	6.81	0.46
Heiko Herrlich (34)	115.71	7.68	23.64	− 0.44	− 0.03	32.47	67.53	3.76	0.54	7.96	0.96
Martin Schmidt (68)	128.29	7.62	21.92	− 0.46	− 0.06	34.78	65.22	2.98	0.37	7.06	0.88
Domenico Tedesco (34)	113.94	7.56	23.88	− 0.12	− 0.09	31.68	68.32	3.35	0.22	5.54	0.79
Niko Kovac (102)	118.18	7.46	23.92	− 0.28	− 0.03	31.18	68.82	3.09	0.55	5.95	0.96
AVERAGE (2176)	**116.22**	**7.40**	**23.06**	**- 0.39**	**- 0.08**	**32.07**	**67.93**	**3.28%**	**0.38%**	**6.99%**	**1.01%**
Julian Nagelsmann (102)	110.10	7.36	23.70	− 0.38	0.03	31.03	68.97	4.43	0.64	8.55	0.76
Thomas Schaaf (34)	127.44	7.29	22.62	− 0.24	− 0.12	32.24	67.76	2.96	0.31	5.57	1.05

Table 9 continued

Coach (Games)	T/M	CP+/T (%)	%CP	S+/−	G+/−	% CP+	%CP−	% S+	% G+	% S−	% G−
Dieter Hecking (170)	106.00	7.27	24.13	− 0.41	− 0.14	30.13	69.87	3.40	0.39	7.18	1.32
Florian Kohfeldt (34)	111.24	7.22	23.43	− 0.71	− 0.26	30.81	69.19	3.61	0.11	9.14	1.63
Markus Weinzierl (136)	109.07	7.21	23.53	− 0.46	− 0.10	30.66	69.34	3.30	0.32	7.31	0.99
Pal Dardai (136)	113.18	7.15	23.36	− 0.69	− 0.08	30.62	69.38	2.50	0.33	7.37	0.92
Bruno Labbadia (68)	114.91	7.06	24.75	− 0.63	− 0.19	28.54	71.46	2.95	0.21	7.24	1.23
Robin Dutt (34)	106.94	7.04	23.93	− 0.56	− 0.24	29.43	70.57	3.45	0.23	7.98	1.63
Ralf Rangnick (34)	131.47	7.00	21.72	0.56	− 0.03	32.23	67.77	4.63	0.41	3.95	0.76
Lucien Favre (68)	94.10	6.95	25.38	− 0.87	− 0.03	27.40	72.60	2.65	0.43	8.65	0.76
Armin Veh (34)	93.38	6.93	25.98	− 0.79	− 0.24	26.67	73.33	2.91	0.00	8.43	1.32
Andr Breitenreiter (102)	117.86	6.82	22.20	− 0.67	− 0.18	30.72	69.28	3.07	0.26	8.11	1.35
Viktor Skripnik (34)	125.35	6.69	21.73	− 0.32	− 0.18	30.78	69.22	3.02	0.32	6.08	1.40
Friedhelm Funkel (34)	100.44	6.62	24.77	− 0.59	− 0.47	26.71	73.29	3.43	0.35	7.90	3.06
Peter Stöger (102)	116.09	6.60	21.81	− 0.35	−0.12	30.25	69.75	2.48	0.19	5.55	0.94
Manuel Baum (34)	121.62	6.41	21.45	− 0.65	0.00	29.88	70.12	2.71	0.34	7.40	0.48
Christian Streich (136)	114.29	6.26	20.34	− 0.68	− 0.07	30.77	69.23	2.66	0.25	8.04	0.78
Dirk Schuster (34)	125.82	5.82	16.88	− 0.65	− 0.03	34.49	65.51	2.35	0.28	8.25	0.63

Appendix D

Table 10 Counterpressing of Bundesliga teams on a season level (1/2). The number in brackets shows the respective position in the final ranking

Team (Ranking)	T/M	CP+/T	%CP	S+/−	G+/−	% CP+	%CP−	% S+	% G+	% S−	% G−
FCB - 2013/2014 (1)	97.29	8.98%	26.18%	− 0.71	0.03	34.30%	65.70%	3.58%	0.46%	9.67%	0.53%
FCB - 2014/2015 (1)	116.65	8.50%	24.79%	0.00	0.06	34.28%	65.72%	3.87%	0.51%	5.88%	0.46%
FCB - 2015/2016 (1)	114.47	8.40%	24.90%	0.09	− 0.03	33.75%	66.25%	4.13%	0.52%	5.76%	0.93%
FCB - 2019/2020 (1)	116.68	8.40%	26.02%	− 0.32	− 0.04	32.28%	67.72%	4.87%	0.66%	8.75%	1.17%
FCB - 2016/2017 (1)	120.03	7.69%	23.52%	− 0.56	0.15	32.71%	67.29%	2.60%	0.83%	6.81%	0.46%
FCB - 2018/2019 (1)	111.15	7.33%	24.21%	0.32	0.12	30.27%	69.73%	5.14%	1.20%	5.64%	1.10%
FCB - 2017/2018 (1)	110.41	7.03%	23.39%	− 0.56	− 0.06	30.07%	69.93%	3.30%	0.46%	7.82%	0.98%
BVB - 2015/2016 (2)	120.79	9.40%	24.71%	− 0.09	0.03	38.03%	61.97%	3.15%	0.59%	5.56%	0.79%
BVB - 2013/2014 (2)	113.12	8.71%	24.60%	− 0.12	− 0.09	35.41%	64.59%	3.59%	0.21%	6.22%	0.82%
BVB - 2018/2019 (2)	105.82	8.59%	26.57%	− 0.32	− 0.12	32.32%	67.68%	2.41%	0.21%	5.26%	0.93%
RBL - 2016/2017 (2)	134.97	8.35%	21.16%	− 0.24	− 0.06	39.44%	60.56%	2.68%	0.72%	5.78%	1.53%
BVB - 2019/2020 (2)	104.00	8.19%	27.65%	− 0.32	− 0.08	29.62%	70.38%	4.17%	0.70%	7.51%	1.38%
S04 - 2017/2018 (2)	113.94	7.56%	23.88%	− 0.12	− 0.09	31.68%	68.32%	3.35%	0.22%	5.54%	0.79%
WOB - 2014/2015 (2)	118.44	7.18%	22.00%	0.15	0.00	32.62%	67.38%	4.63%	0.45%	6.03%	0.67%
LEV - 2015/2016 (3)	143.85	8.79%	22.33%	− 0.24	0.15	39.38%	60.62%	2.66%	0.55%	5.59%	0.15%
BVB - 2016/2017 (3)	122.38	8.29%	24.75%	− 0.24	− 0.03	33.50%	66.50%	3.69%	0.58%	6.72%	1.02%

Table 10 continued

Team (Ranking)	T/M	CP+/T	%CP	S+/−	G+/−	% CP+	%CP−	% S+	% G+	% S−	% G−
RBL - 2019/2020 (3)	110.60	7.88%	22.21%	− 0.20	− 0.12	35.50%	64.50%	4.07%	0.33%	7.58%	1.26%
HOF - 2017/2018 (3)	109.15	7.49%	23.31%	− 0.18	0.12	32.14%	67.86%	4.74%	0.81%	8.01%	0.51%
RBL - 2018/2019 (3)	131.47	7.00%	21.72%	0.56	− 0.03	32.23%	67.77%	4.63%	0.41%	3.95%	0.76%
S04 - 2013/2014 (3)	96.74	6.87%	23.44%	− 0.03	0.03	29.31%	70.69%	4.02%	0.65%	5.87%	0.73%
BMG - 2014/2015 (3)	103.12	6.85%	24.02%	− 0.62	− 0.03	28.50%	71.50%	2.85%	0.36%	7.48%	0.66%
LEV - 2014/2015 (4)	156.29	8.49%	18.84%	0.91	0.09	45.05%	54.95%	5.29%	0.80%	4.00%	0.91%
LEV - 2019/2020 (4)	114.76	8.16%	23.60%	− 0.64	− 0.12	34.56%	65.44%	3.10%	0.00%	8.35%	0.68%
BVB - 2017/2018 (4)	121.85	7.89%	24.79%	− 0.71	0.00	31.84%	68.16%	3.51%	0.49%	8.57%	0.71%
BMG - 2015/2016 (4)	117.32	7.60%	24.34%	− 0.47	− 0.15	31.20%	68.80%	3.30%	0.41%	7.19%	1.35%
HOF - 2016/2017 (4)	107.94	7.52%	23.98%	− 0.35	− 0.06	31.36%	68.64%	3.64%	0.34%	7.28%	0.83%
LEV - 2013/2014 (4)	101.26	7.44%	26.08%	− 0.38	− 0.06	28.51%	71.49%	3.01%	0.56%	6.23%	1.09%
LEV - 2018/2019 (4)	114.68	7.28%	24.26%	− 0.29	− 0.15	30.02%	69.98%	4.65%	0.21%	8.16%	1.06%
LEV - 2017/2018 (5)	115.71	7.68%	23.64%	− 0.44	− 0.03	32.47%	67.53%	3.76%	0.54%	7.96%	0.96%
AUG - 2014/2015 (5)	121.35	7.54%	22.95%	− 0.85	− 0.21	32.84%	67.16%	2.75%	0.21%	8.65%	1.42%
WOB - 2013/2014 (5)	94.82	7.44%	24.88%	− 0.41	− 0.12	29.93%	70.07%	2.24%	0.12%	5.69%	0.89%
BMG - 2018/2019 (5)	99.29	7.35%	26.04%	− 0.85	− 0.29	28.21%	71.79%	3.98%	0.34%	10.14%	2.06%
KOE - 2016/2017 (5)	113.41	6.95%	22.15%	− 0.26	− 0.12	31.38%	68.62%	2.34%	0.00%	4.95%	0.68%
S04 - 2015/2016 (5)	115.44	6.80%	22.39%	− 0.76	− 0.12	30.38%	69.62%	3.30%	0.34%	8.99%	1.14%

Table 10 continued

Team (Ranking)	T/M	CP+/T	%CP	S+/−	G+/−	% CP+	%CP−	% S+	% G+	% S−	% G−
BMG - 2019/2020 (5)	110.48	6.59%	22.95%	− 0.56	− 0.24	28.71%	71.29%	2.84%	0.32%	7.08%	1.77%
RBL - 2017/2018 (6)	125.44	8.14%	23.56%	− 0.26	− 0.06	34.53%	65.47%	3.88%	0.50%	7.29%	1.06%
M05 - 2015/2016 (6)	123.71	7.75%	22.92%	− 0.50	0.03	33.82%	66.18%	3.01%	0.31%	7.21%	0.31%
HBER - 2016/2017 (6)	115.29	7.73%	22.65%	− 0.47	− 0.03	34.12%	65.88%	2.48%	0.34%	6.50%	0.68%
WOB - 2018/2019 (6)	110.38	7.49%	27.15%	− 0.79	− 0.21	27.58%	72.42%	2.85%	0.29%	7.59%	1.36%
BMG - 2013/2014 (6)	85.09	7.09%	27.03%	− 1.12	− 0.03	26.21%	73.79%	2.43%	0.51%	9.88%	0.87%
WOB - 2019/2020 (6)	123.36	6.74%	23.09%	− 0.40	− 0.12	29.21%	70.79%	3.23%	0.14%	6.55%	0.79%
S04 - 2014/2015 (6)	123.56	6.19%	20.97%	− 1.03	0.00	29.51%	70.49%	2.50%	0.34%	9.18%	0.48%
BVB - 2014/2015 (7)	144.88	8.40%	22.66%	0.65	0.00	37.10%	62.90%	4.75%	0.36%	4.42%	0.57%
FRA - 2018/2019 (7)	128.94	8.05%	23.27%	− 0.59	− 0.09	34.61%	65.39%	3.63%	0.39%	8.55%	1.05%
M05 - 2013/2014 (7)	95.82	7.98%	24.31%	− 0.50	− 0.18	32.83%	67.17%	3.66%	0.38%	8.65%	1.69%
STU - 2017/2018 (7)	114.26	7.54%	24.53%	− 0.71	− 0.06	30.75%	69.25%	2.31%	0.31%	6.97%	0.76%
HBER - 2015/2016 (7)	111.53	7.30%	22.97%	− 0.56	− 0.09	31.80%	68.20%	2.18%	0.34%	6.40%	1.01%
SCF - 2016/2017 (7)	122.44	5.79%	19.10%	− 0.76	− 0.03	30.31%	69.69%	2.26%	0.25%	7.94%	0.54%
SCF - 2019/2020 (7)	106.48	5.48%	22.46%	− 1.84	− 0.44	24.41%	75.59%	3.18%	0.17%	14.38%	2.65%
HOF - 2014/2015 (8)	141.03	8.03%	22.34%	0.24	0.12	35.95%	64.05%	4.30%	0.56%	5.54%	0.29%
BRE - 2018/2019 (8)	111.24	7.22%	23.43%	− 0.71	− 0.26	30.81%	69.19%	3.61%	0.11%	9.14%	1.63%
S04 - 2019/2020 (8)	113.68	7.14%	24.77%	− 0.88	− 0.24	28.84%	71.16%	2.84%	0.14%	8.38%	1.40%

Table 10 continued

Team (Ranking)	T/M	CP+/T	%CP	S+/−	G+/−	% CP+	%CP−	% S+	% G+	% S−	% G−
AUG - 2013/2014 (8)	93.50	7.14%	25.01%	− 0.12	0.00	28.55%	71.45%	3.65%	0.63%	5.81%	0.88%
FRA - 2017/2018 (8)	120.18	7.12%	25.01%	− 0.97	− 0.12	28.47%	71.53%	2.25%	0.39%	7.66%	1.09%
BRE - 2016/2017 (8)	110.15	6.92%	23.68%	− 0.76	− 0.24	29.20%	70.80%	3.16%	0.34%	8.60%	1.75%
WOB - 2015/2016 (8)	109.85	6.53%	23.51%	− 0.35	− 0.26	27.79%	72.21%	2.62%	0.34%	5.52%	1.89%
BMG - 2017/2018 (9)	107.59	7.90%	24.69%	− 0.59	0.00	32.00%	68.00%	3.43%	0.66%	8.31%	0.98%
HOF - 2013/2014 (9)	108.24	7.61%	23.91%	− 0.33	− 0.18	31.85%	68.15%	3.28%	0.23%	6.70%	1.37%
FRA - 2014/2015 (9)	127.44	7.29%	22.62%	− 0.24	− 0.12	32.24%	67.76%	2.96%	0.31%	5.57%	1.05%
BMG - 2016/2017 (9)	109.79	7.26%	23.36%	− 0.59	− 0.12	31.08%	68.92%	3.21%	0.23%	7.99%	1.00%
AVERAGE (9)	**115.60**	**7.25%**	**23.08%**	**− 0.51**	**− 0.11**	**31.42%**	**68.58%**	**3.22%**	**0.35%**	**7.51%**	**1.13%**
HOF - 2018/2019 (9)	113.21	7.07%	23.82%	− 0.62	0.03	29.66%	70.34%	4.91%	0.76%	10.23%	0.93%
KOE - 2015/2016 (9)	115.47	6.11%	21.88%	− 0.59	− 0.12	27.94%	72.06%	2.10%	0.12%	6.14%	0.81%
HOF - 2019/2020 (9)	102.44	5.66%	22.73%	− 1.28	− 0.12	24.91%	75.09%	2.75%	0.69%	10.98%	1.60%
S04 - 2016/2017 (10)	114.38	7.17%	23.68%	− 0.47	− 0.09	30.29%	69.71%	3.04%	0.43%	6.85%	1.09%
BRE - 2014/2015 (10)	127.21	6.68%	21.39%	− 0.44	− 0.12	31.24%	68.76%	3.57%	0.00%	7.55%	0.63%
HSV - 2015/2016 (10)	119.44	6.67%	22.53%	− 0.47	− 0.18	29.62%	70.38%	3.06%	0.11%	6.83%	1.09%
DUE - 2018/2019 (10)	100.44	6.62%	24.77%	− 0.59	− 0.47	26.71%	73.29%	3.43%	0.35%	7.90%	3.06%
H96 - 2013/2014 (10)	98.32	6.61%	23.24%	− 0.38	− 0.15	28.44%	71.56%	4.12%	0.26%	8.09%	1.26%

Table 10 continued

Team (Ranking)	T/M	CP+/T	%CP	S+/−	G+/−	% CP+	%CP−	% S+	% G+	% S−	% G−
KOE - 2019/2020 (10)	106.52	6.57%	23.73%	− 0.96	− 0.28	27.69%	72.31%	3.16%	0.00%	9.63%	1.53%
HBER - 2017/2018 (10)	112.94	6.25%	23.41%	− 0.62	− 0.09	26.70%	73.30%	2.89%	0.33%	7.13%	0.91%
HBER - 2013/2014 (11)	104.97	8.27%	24.99%	− 0.12	− 0.09	33.07%	66.93%	2.80%	0.11%	4.86%	0.67%
FRA - 2016/2017 (11)	123.21	7.90%	22.58%	− 0.21	− 0.09	34.99%	65.01%	2.01%	0.11%	4.23%	0.65%
ING - 2015/2016 (11)	148.65	7.58%	21.43%	− 0.32	0.03	35.36%	64.64%	2.22%	0.46%	5.00%	0.57%
HBER - 2019/2020 (11)	104.04	7.38%	25.03%	− 0.52	− 0.24	29.49%	70.51%	2.76%	0.31%	6.75%	1.74%
HBER - 2018/2019 (11)	112.94	7.32%	24.43%	− 1.12	− 0.12	29.96%	70.04%	2.45%	0.32%	9.28%	1.07%
BRE - 2017/2018 (11)	116.06	6.79%	24.00%	− 0.68	− 0.12	28.30%	71.70%	2.75%	0.11%	7.22%	0.74%
M05 - 2014/2015 (11)	129.59	6.13%	21.56%	− 0.59	− 0.09	28.42%	71.58%	3.47%	0.21%	7.79%	0.74%

Table 11 Counterpressing of Bundesliga teams on a season level (2/2). The number in brackets shows the respective position in the final ranking

Team (Ranking)	T/M	CP+/T	%CP	S+/−	G+/-	% CP+	%CP-	% S+	% G+	% S-	% G-
LEV - 2016/2017 (12)	129.71	8.64%	23.70%	− 0.71	− 0.18	36.46%	63.54%	3.06%	0.38%	8.43%	1.51%
M05 - 2018/2019 (12)	121.79	7.49%	24.70%	− 0.62	− 0.26	30.30%	69.70%	3.42%	0.10%	7.85%	1.40%
BRE - 2013/2014 (12)	106.94	7.04%	23.93%	− 0.56	− 0.24	29.43%	70.57%	3.45%	0.23%	7.98%	1.63%
AUG - 2015/2016 (12)	107.03	6.95%	22.73%	− 0.38	-0.09	30.59%	69.41%	3.87%	0.00%	7.84%	0.52%
AUG - 2019/2020 (12)	104.40	6.82%	22.68%	− 0.52	-0.36	30.07%	69.93%	3.55%	0.68%	8.21%	3.14%
KOE - 2014/2015 (12)	119.38	6.73%	21.41%	− 0.21	-0.12	31.42%	68.58%	2.99%	0.46%	5.54%	1.34%
AUG - 2017/2018 (12)	121.62	6.41%	21.45%	− 0.65	0.00	29.88%	70.12%	2.71%	0.34%	7.40%	0.48%
FRA - 2013/2014 (13)	93.38	6.93%	25.98%	− 0.79	− 0.24	26.67%	73.33%	2.91%	0.00%	8.43%	1.32%
BRE - 2015/2016 (13)	125.35	6.69%	21.73%	− 0.32	− 0.18	30.78%	69.22%	3.02%	0.32%	6.08%	1.40%
H96 - 2017/2018 (13)	115.74	6.45%	22.21%	− 1.06	-0.35	29.06%	70.94%	2.52%	0.11%	9.35%	2.10%
AUG - 2016/2017 (13)	118.97	6.18%	21.63%	− 0.71	-0.09	28.57%	71.43%	3.20%	0.23%	8.32%	0.80%
H96 - 2014/2015 (13)	130.62	6.10%	19.88%	− 1.06	-0.21	30.69%	69.31%	1.81%	0.00%	8.50%	1.14%
SCF - 2018/2019 (13)	103.03	6.08%	22.01%	− 0.68	− 0.06	27.63%	72.37%	3.37%	0.39%	8.78%	0.90%
UBER - 2019/2020 (13)	114.96	5.50%	18.65%	− 0.48	− 0.24	29.48%	70.52%	2.99%	0.00%	7.41%	1.59%
M05 - 2017/2018 (14)	119.74	8.38%	23.38%	0.09	− 0.09	35.82%	64.18%	3.99%	0.42%	5.73%	1.15%
S04 - 2018/2019 (14)	116.68	7.71%	24.33%	− 0.88	− 0.32	31.71%	68.29%	2.80%	0.10%	8.65%	1.82%
HSV - 2016/2017 (14)	139.88	7.59%	20.90%	0.24	0.03	36.32%	63.68%	3.52%	0.50%	4.27%	0.63%
STU - 2014/2015 (14)	120.24	7.46%	22.06%	− 0.65	− 0.26	33.81%	66.19%	2.88%	0.22%	8.04%	1.84%
FRA - 2019/2020 (14)	123.00	7.35%	21.71%	− 1.12	− 0.12	33.85%	66.15%	3.28%	0.62%	11.32%	1.65%
SCF - 2013/2014 (14)	96.24	6.57%	22.31%	− 0.79	− 0.12	29.45%	70.55%	2.33%	0.14%	8.54%	0.97%

Table 11 continued

Team (Ranking)	T/M	CP+/T	%CP	S+/–	G+/-	% CP+	%CP-	% S+	% G+	% S-	% G-
DAR - 2015/2016 (14)	125.82	5.82%	16.88%	– 0.65	– 0.03	34.49%	65.51%	2.35%	0.28%	8.25%	0.63%
STU - 2013/2014 (15)	92.38	8.28%	25.66%	0.03	– 0.03	32.26%	67.74%	4.22%	0.50%	6.04%	0.92%
HOF - 2015/2016 (15)	121.85	8.21%	22.95%	– 0.74	– 0.24	35.75%	64.25%	2.73%	0.00%	8.35%	1.31%
M05 - 2016/2017 (15)	132.88	7.50%	20.98%	– 0.41	– 0.15	35.76%	64.24%	2.95%	0.42%	6.90%	1.48%
HBER - 2014/2015 (15)	127.85	7.34%	22.87%	– 0.12	– 0.12	32.09%	67.91%	2.52%	0.40%	4.30%	1.19%
M05 - 2019/2020 (15)	117.72	7.24%	23.48%	– 0.68	– 0.16	30.82%	69.18%	3.33%	0.43%	8.37%	1.46%
SCF - 2017/2018 (15)	114.32	6.51%	20.63%	– 0.68	– 0.18	31.55%	68.45%	2.87%	0.12%	8.38%	1.28%
AUG - 2018/2019 (15)	111.68	6.29%	21.83%	– 1.06	– 0.29	28.83%	71.17%	3.26%	0.00%	10.68%	1.69%
HSV - 2013/2014 (16)	95.32	7.44%	22.96%	– 0.59	– 0.44	32.39%	67.61%	4.17%	0.40%	10.14%	3.58%
FRA - 2015/2016 (16)	123.26	7.25%	23.50%	– 1.00	0.06	30.86%	69.14%	2.54%	0.81%	8.66%	0.88%
DUE - 2019/2020 (16)	107.52	7.11%	24.29%	– 1.04	– 0.24	29.25%	70.75%	2.30%	0.31%	8.87%	1.73%
WOB - 2016/2017 (16)	117.50	6.93%	22.38%	– 0.74	– 0.18	30.98%	69.02%	2.46%	0.00%	7.62%	0.97%
STU - 2018/2019 (16)	116.62	6.86%	22.65%	– 0.59	– 0.18	30.29%	69.71%	3.79%	0.11%	8.63%	1.12%
HSV - 2014/2015 (16)	119.79	6.58%	21.46%	– 0.50	– 0.21	30.66%	69.34%	3.20%	0.46%	7.43%	1.82%
WOB - 2017/2018 (16)	106.88	6.38%	24.38%	– 0.88	– 0.15	26.19%	73.81%	2.48%	0.00%	7.95%	0.76%
ING - 2016/2017 (17)	147.18	7.51%	20.76%	– 0.38	– 0.15	36.19%	63.81%	2.98%	0.29%	6.64%	1.21%
STU - 2015/2016 (17)	125.50	7.27%	22.73%	– 0.59	– 0.15	31.96%	68.04%	3.61%	0.52%	8.33%	1.52%
HSV - 2017/2018 (17)	129.35	7.21%	22.71%	– 0.79	– 0.24	31.73%	68.27%	3.30%	0.30%	8.80%	1.61%

Table 11 continued

Team (Ranking)	T/M	CP+/T	%CP	S+/−	G+/-	% CP+	%CP-	% S+	% G+	% S-	% G-
NUR - 2013/2014 (17)	99.47	6.86%	24.78%	− 1.09	− 0.12	27.68%	72.32%	3.58%	0.36%	11.06%	1.16%
BRE - 2019/2020 (17)	105.25	6.81%	24.03%	− 0.62	− 0.21	28.34%	71.66%	3.79%	0.33%	8.74%	1.61%
H96 - 2018/2019 (17)	109.35	6.70%	23.02%	− 1.59	− 0.32	29.09%	70.91%	2.22%	0.12%	12.03%	1.98%
SCF - 2014/2015 (17)	124.18	6.25%	19.78%	− 0.47	0.06	31.62%	68.38%	3.11%	0.48%	7.36%	0.35%
PAD - 2019/2020 (18)	106.12	7.50%	24.95%	− 0.04	− 0.20	30.06%	69.94%	3.32%	0.00%	4.97%	1.08%
PAD - 2014/2015 (18)	122.41	7.18%	22.01%	− 0.18	− 0.06	32.64%	67.36%	3.38%	0.33%	6.00%	0.81%
BRA - 2013/2014 (18)	104.42	6.99%	22.20%	− 0.42	− 0.09	31.50%	68.50%	3.01%	0.13%	7.06%	0.76%
DAR - 2016/2017 (18)	112.09	6.64%	21.18%	− 1.12	− 0.06	31.35%	68.65%	2.73%	0.37%	10.83%	0.90%
H96 - 2015/2016 (18)	119.06	6.35%	22.38%	− 1.03	− 0.24	28.37%	71.63%	2.54%	0.22%	8.94%	1.54%
KOE - 2017/2018 (18)	118.53	6.00%	23.50%	− 0.79	− 0.18	25.55%	74.45%	3.91%	0.53%	9.08%	1.56%
NUR - 2018/2019 (18)	104.85	5.89%	23.98%	− 1.03	− 0.26	24.56%	75.44%	1.99%	0.12%	8.06%	1.55%

References

Andrienko G et al (2017) Visual analysis of pressure in football. Data Mining Knowl Discov 31(6):1793–1839. https://doi.org/10.1007/s10618-017-0513-2

Andrienko G et al (2019) Constructing Spaces and Times for Tactical Analysis in Football. IEEE Trans Vis Comput Graph. https://doi.org/10.1109/tvcg.2019.2952129

Antipov EA, Pokryshevskaya EB (2020) Interpretable machine learning for demand modeling with high-dimensional data using Gradient Boosting Machines and Shapley values. J Revenue Pricing Manag 19(5):355–364. https://doi.org/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 (in print) Using Artificial Intelligence to Enhance Sport Performance, pp. 1-18. https://doi.org/10.3389/fspor.2021.624475.

Bergstra J et al (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, pp. 1–9

Bialkowski A et al (2014) “Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data”. In: Proceedings - IEEE International Conference on Data Mining, ICDM 2015-January, pp. 725–730. issn:15504786. https://doi.org/10.1109/ICDM.2014.133

Bialkowski A et al (2015) “Identifying team style in soccer using formations learned from spatiotemporal tracking data”. In: IEEE International Conference on Data Mining Workshops, ICDMW 2015.January, pp. 9-14. issn: 23759259. https://doi.org/10.1109/ICDMW.2014.167.29 https://gdpr-info.eu/, accessed 07/20/20.Data-Driven Detection of Counterpressing in Professional Football 15

Bojinov I, Bornn L (2016) “The Pressing Game: Optimal Defensive Disruption in Soccer”. In: MIT Sloan Sports Analytics Conference, pp. 1–8

Bourbousson Jérôme, Carole Sève Tim McGarry (2010) “Space-time coordination dynamics in basketball: Part 2. the interaction between the two teams”. In: Journal of Sports Sciences 28.3, pp. 349–358. issn: 02640414. https://doi.org/10.1080/02640410903503640

Brefeld U, Lasek J, Mair S (2019) Probabilistic movement models and zones of control. Mach Learn 108(1):127–147. https://doi.org/10.1007/s10994-018-5725-1

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-Augu, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Decroos T, Van Haaren J, Davis J (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. isbn: 9781450355520. https://doi.org/10.1145/3219819.3219832. url: https://people.cs.kuleuven.be/~jesse.davis/decroos-kdd18.pdf

Decroos T et al (2020) “VAEP: An Objective Approach to Valuing On-the-Ball Actions in Soccer (Extended Abstract)”. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), pp. 4696–4700. issn: 10450823. https://doi.org/10.24963/ijcai.2020/648.

Dewnacker Ian, Michael McCourt, Scott Clark (2016) “Bayesian Optimization for Machine Learning. A Practical Guidebook”. arXiv:1612.04858

Fairchild A, Pelechrinis K, Kokkodis M (2018) “Spatial analysis of shots in MLS: A model for expected goals and fractal dimensionality”. In: Journal of Sports Analytics 4.3, pp. 165–174. issn: 2215020X. https://doi.org/10.3233/jsa-170207.

Navarro Férnandez, J (2018) Analysis of Styles of Play in Soccer and Their Effectiveness. isbn: 9788413060576

Fernandez J, Bornn L (2018) “Wide Open Spaces : A statistical technique for measuring space creation in professional soccer”. In: MIT Sloan Sports Analytics Conference, pp. 1–19

Goes F, Kempe M, Koen L (2019). “Predicting match outcome in professional Dutch football using tactical performance metrics computed from position tracking data”. In: MathSport International Conference June, pp. 4-5

Goes F et al (2020) Interpretable machine learning for demand modeling with high-dimensional data using Gradient Boosting Machines and Shapley values. Euro J Sport Sci. https://doi.org/10.1080/17461391.2020.1747552

Grant AG et al (1999) Analysis of the goals scored in the 1998 World Cup. J Sports Sci 17(10):826–827

Herold M et al (2019) Machine learning in men's professional football: current applications and future directions for improving attacking play. Int J Sports Sci Coaching. https://doi.org/10.1177/1747954119879350

Hobbs J et al. (2018). "Quantifying the Value of Transitions in Soccer via Spatiotemporal Trajectory Clustering". In: MIT Sloan Sports Analytics Conferencece, pp. 1–11

Hughes M, Ian M Franks (2015) Essentials of performance analysis in sport. Vol. 53. 04, pp. 53-1831. https://doi.org/10.5860/choice.193440. url:https://books.google.de/books?hl=de&lr=&id=KorCDwAAQBAJ&oi=fnd&pg=PT14&dq=the+essentials+of+performance+analysis&ots=ZhJd6413Fq&sig=FMnUcL21bpzACnWTzNhxESEgUY4\#v=onepage&q=theessentialsofperformanceanalysis&f=false

Ibrahim L et al (2020) Explainable prediction of acute myocardial infarction using machine learning and shapley values. IEEE Access. https://doi.org/10.1109/access.2020.3040166

Kempe M et al (2014) Possession vs. direct play: evaluating tactical behavior in elite soccer. Int J Sports Sci 4(6A):35–41. https://doi.org/10.5923/s.sports.201401.05

Kim S (2004) Voronoi Analysis of a Soccer Game. Nonlinear Anal Model Control 9(3):233–240. https://doi.org/10.15388/na.2004.9.3.15154

Li TR et al (2019). Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. https://doi.org/10.3389/fphy.2019.00098

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

Link D, Hoernig M (2017) Individual ball possession in soccer. PLoS ONE 12(7):1–15. https://doi.org/10.1371/journal.pone.0179953

Liu K, Chen W, Lin H (2020) XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites. Mole Genet Genom 295(1):13–21. https://doi.org/10.1007/s00438-019-01600-9

Patrick L et al (2014) "Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data". In: MIT Sloan Sports Analytics Conference, pp. 1–9. url: http://www.sloansportsconference.com/?p=15790

Lundberg SM, SI Lee (2017) "Consistent feature attribution for tree ensembles". In: Proceedings of the 34th International Conference on Machine Learning, pp. 1–9

Yuan M et al (2020) What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values. J Theor Appl Electron Commerce Res 16(3):466–490. https://doi.org/10.3390/jtaer16030029

Pappalardo L et al (2019) A public data set of spatio-temporal match events in soccer competitions. Scientific Data 6(1):236. https://doi.org/10.1038/s41597-019-0247-7

Paul P et al (2018) "Mythbusting Set-Pieces in Soccer". In: MIT Sloan Sports Analytics Conference, pp. 1-12

Rathke A (2017) "An examination of expected goals and shot efficiency in soccer". In: Journal of Human Sport and Exercise 12.Proc2. issn: 1988-5202. https://doi.org/10.14198/jhse.2017.12.proc2.05. url: http://www.redalyc.org/articulo.oa?id=301052437005

Ratner Alexander J et al (2017) "Learning to compose domain-specific transformations for data augmentation". In: Advances in Neural Information Processing Systems Nips, pp. 3237–3247. issn: 10495258

Ratner A et al (2016) "Data programming: Creating large training sets, quickly". In: Advances in Neural Information Processing Systems Nips, pp. 3574–3582. issn: 10495258

Reep C, Benjamin B (1968) Skill and Chance in Association Football. J Royal Stat Soc Series A (General). https://doi.org/10.2307/2343726

Robert R, Memmert D (2016) "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science". In: SpringerPlus 5.1. issn: 21931801. https://doi.org/10.1186/s40064-016-3108-2.

Robberechts P (2019) "Valuing the Art of Pressing". In: StatsBomb Innovation in Football Conference 2019, p. 11. url: http://statsbomb.com/wp-content/uploads/2019/10/Pieter-Robberechts-Valuing-the-Art-of-Pressing.pdf

Roth Alvin E, Thomson W (1988) The Shapley Value: Essays in Honor of Lloyd S. Shapley. isbn: 052136177X. https://doi.org/10.2307/2554979.

Santos A. Benito et al (2018) "Data-driven visual performance analysis in soccer: An exploratory prototype". In: Frontiers in Psychology 9. issn: 16641078. https://doi.org/10.3389/fpsyg.2018.02416.

Shaw L, Glickman M (2019) "Dynamic analysis of team strategy in professional football". In: Barça sports analytics summit. Retrieved from https://static.capabiliaserver.com/frontend/clients/barca/wp_prod/wp-content/uploads/2020/01/56ce723e-barca-conferencepaper-laurie-shaw.pdf

Spearman W (2018) "Beyond Expected Goals". In: MIT Sloan Sports Analytics Conference, pp. 1–17

Steiner S et al (2019) Outplaying opponents-a differential perspective on passes using position data. German J Exerc Sport Res. https://doi.org/10.1007/s12662-019-00579-0

Travassos B et al (2013) Performance analysis in team sports: advances from an ecological dynamics approach. Int J Perform Anal Sport 13(1):83–95. https://doi.org/10.1080/24748668.2013.11868633

Vogelbein M, Nopp S, Hökelmann A (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. https://doi.org/10.1080/02640414.2013.879671.url: http://dx.doi.org/10.1080/02640414.2013.879671

Wang Y (2019) "A Xgboost Risk Model Via Feature Selection and Bayesian Hyper -Parameter Optimization". arXiv:1901.08433

Zhang W et al (2020)"Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization". Geoscience Frontiers 12(1): 469–477. https://doi.org/10.1016/j.gsf.2020.03.007

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).

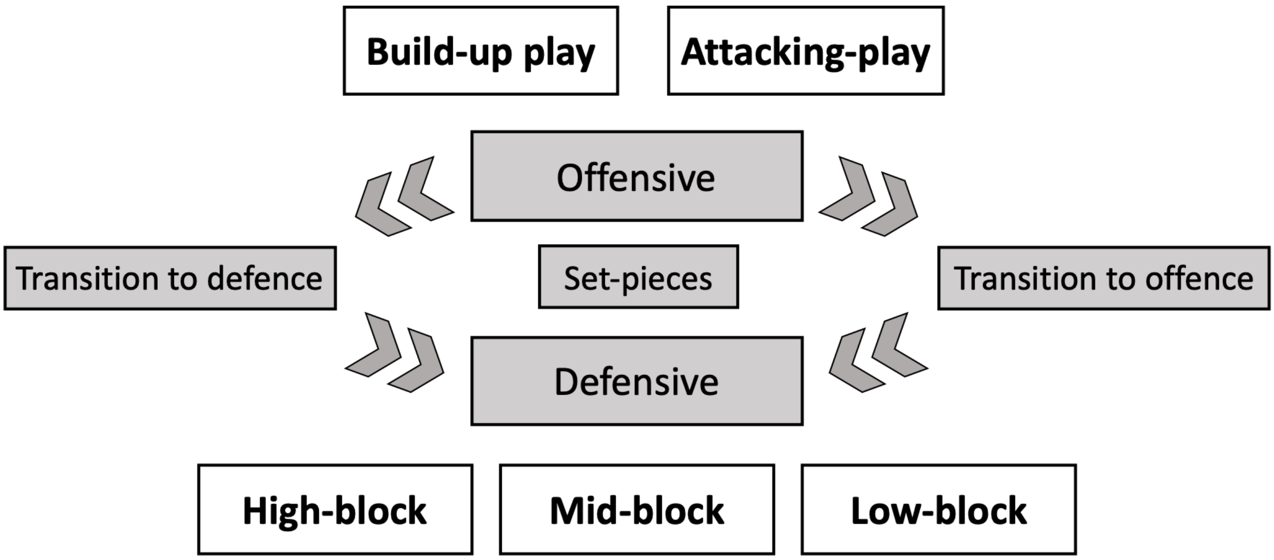

Fig. 1: Overview of tactical phases of play considered.

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.

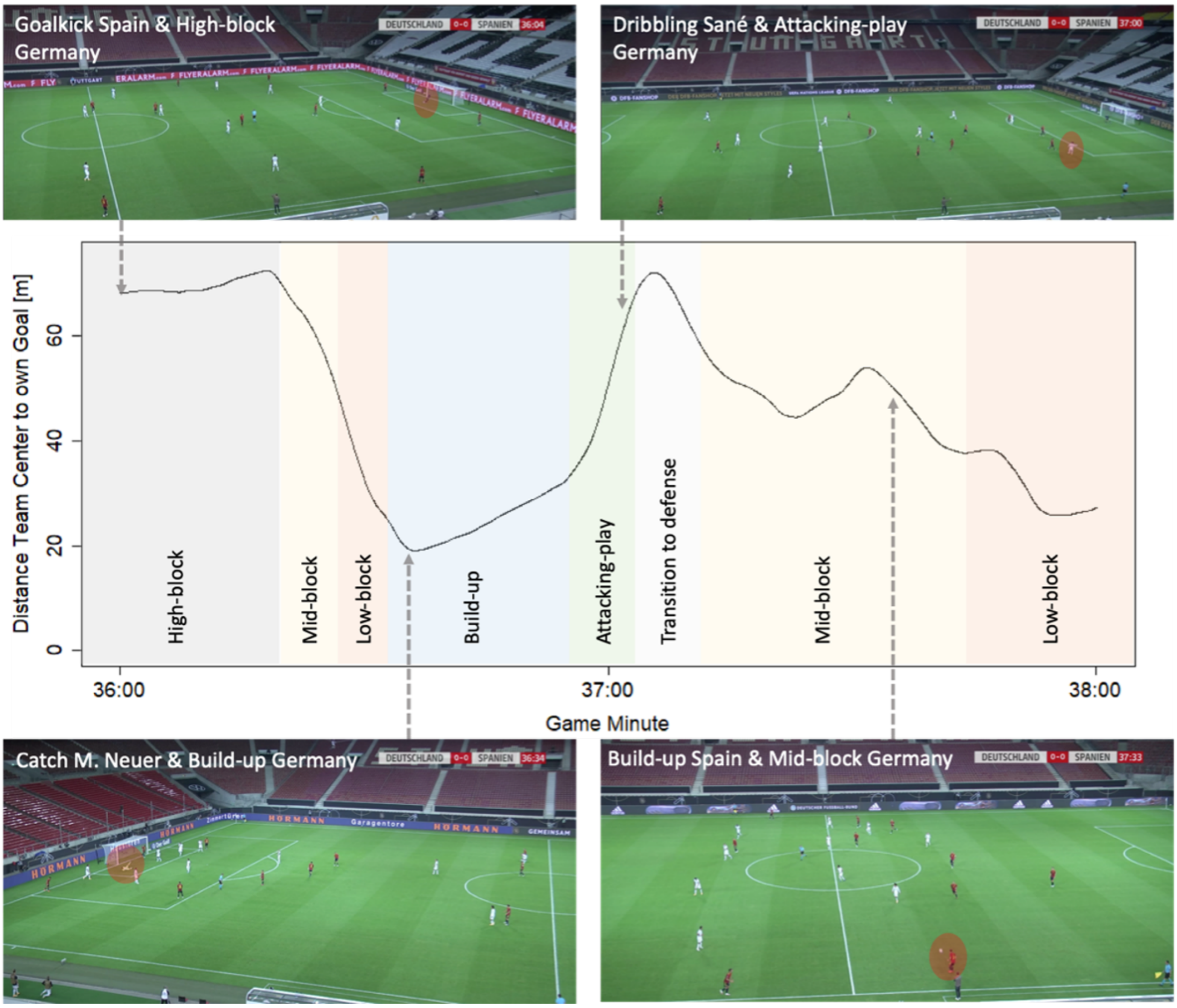

Fig. 2: Team behaviour per phase of play by the reference of Germany against Spain (3rd of September 2020, venue: Stuttgart, result: 1:1). The highlighted areas (red) in the video-footage mark the current ball action.

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).

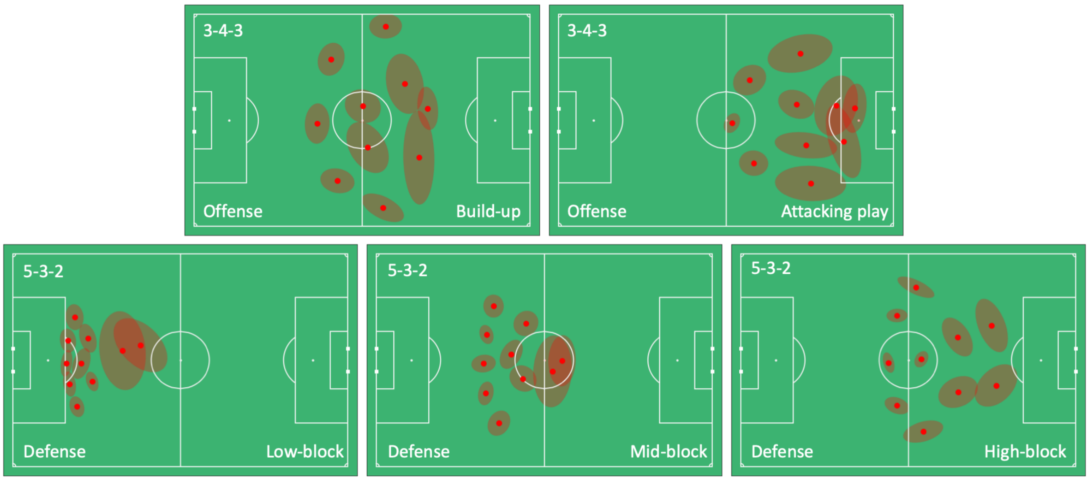

Fig. 3: Average player positions of a team per tactical phase of play during one match. The ellipses provide an estimate of how far the player would tend to move from their average position during each phase of play. The considered team plays from left to right. Player's positions are collected by optical 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):

$$W(\mu_1, \mu_2)^2 = \|m_1 - m_2\|^2 + \operatorname{trace}\left(C_1 + C_2 - 2\left(\sqrt{C_2}C_1\sqrt{C_2}\right)^{1/2}\right),$$

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).

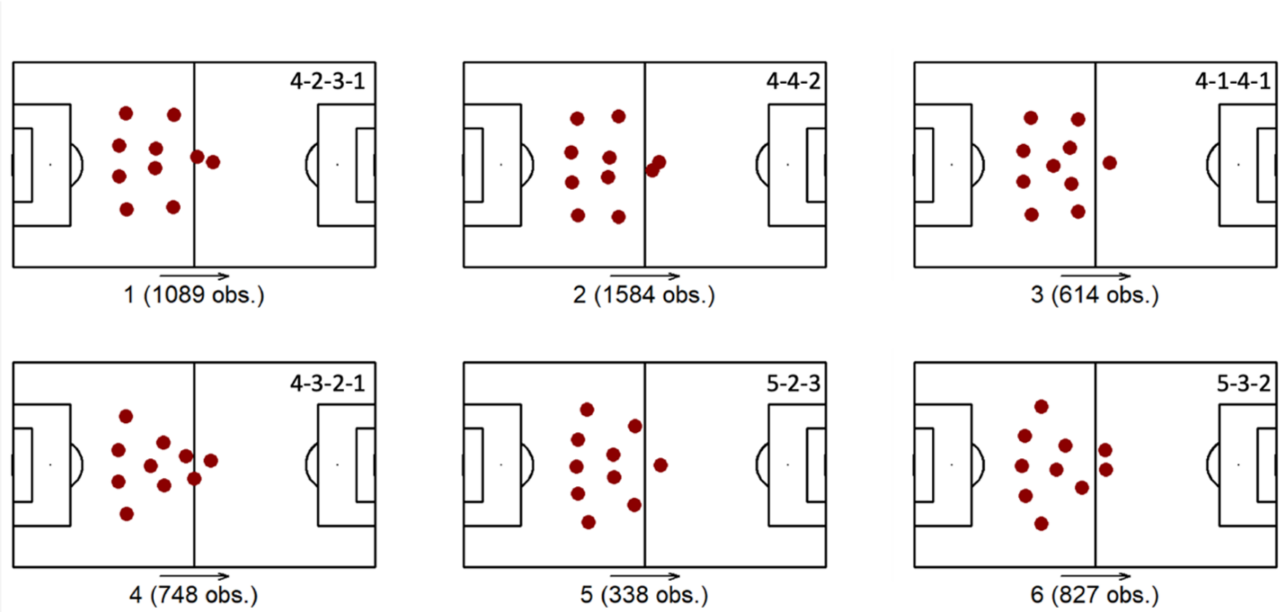

Fig. 4: Outcome of the clustering for mid-block including the number of observations (obs.) of our sample.

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.

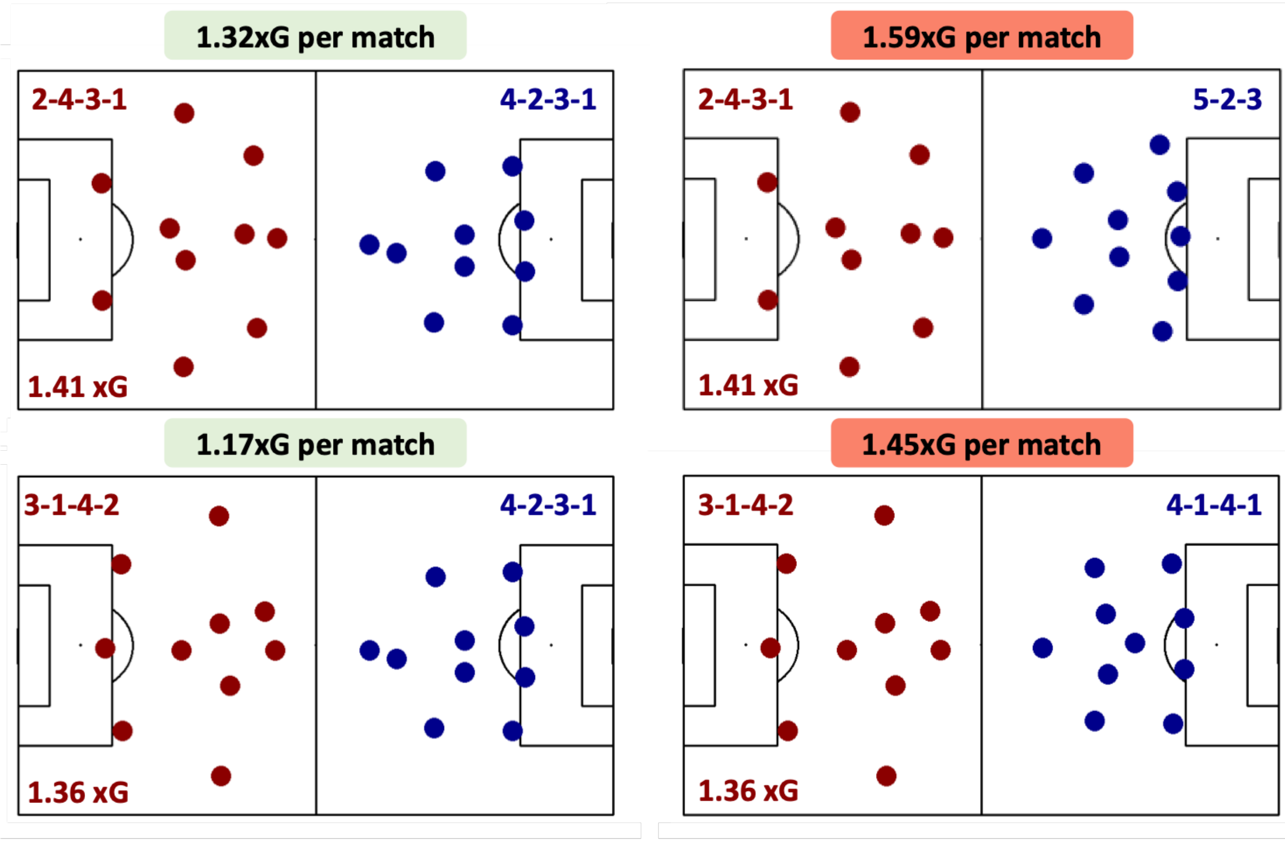

Fig. 5: Effectiveness of defensive formations (blue) against two (upper) and three (lower) player build-up (red).

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.

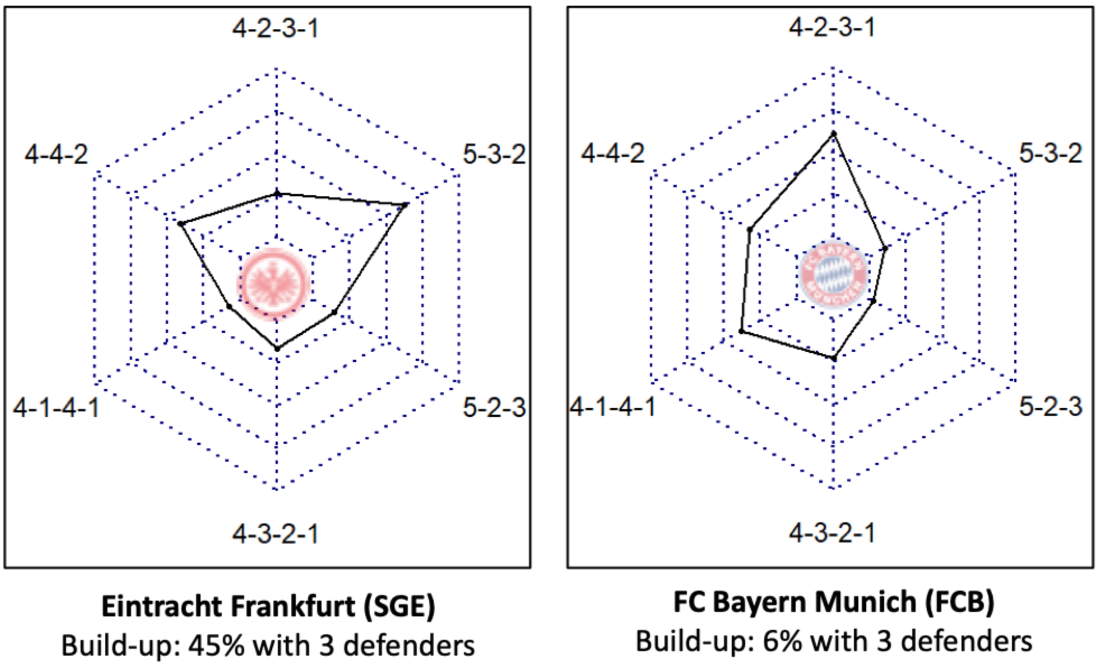

Fig. 6: Formations used by selected German Bundesliga clubs in the mid-block phase.

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.

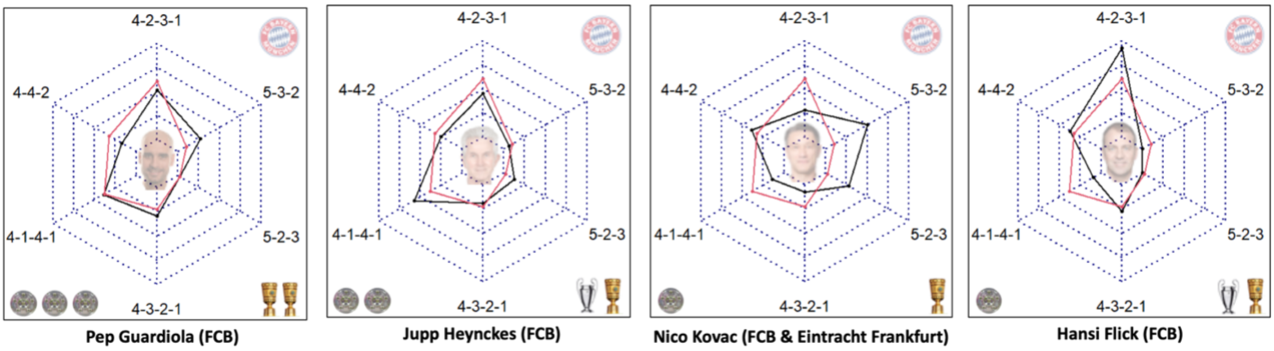

Fig. 7: FC Bayern Munich coaches by their formation (black) in comparison to the overall Bayern profile (red). The data from all coaches and FC Bayern are aggregated over the seasons 2013/2014 to 2019/2020. The trophies (Bundesliga championship, DFB-Cup and UEFA Champions-League) that each coach earned at his time at FC Bayern are displayed.

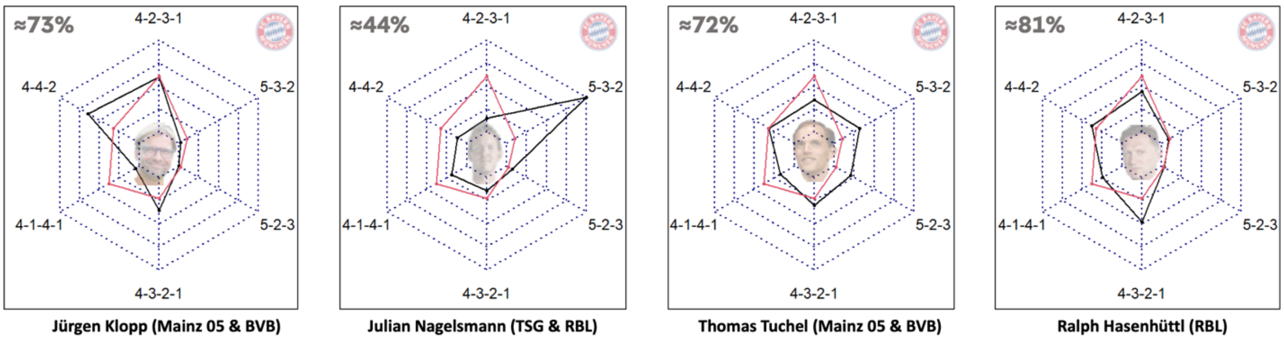

Fig. 8: Formation similarity. Who is the best fit for FC Bayern Munich? Top left the similarity of each coach compared to FC Bayern is displayed.

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

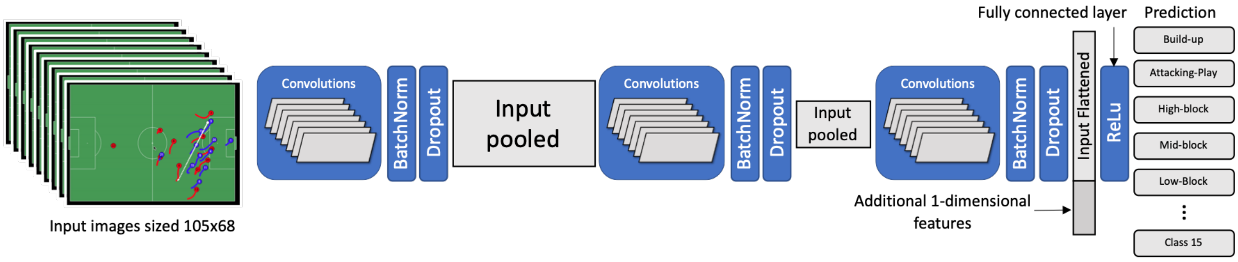

Fig. 9: Schematic architecture of the CNN predicting the phases of play.

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

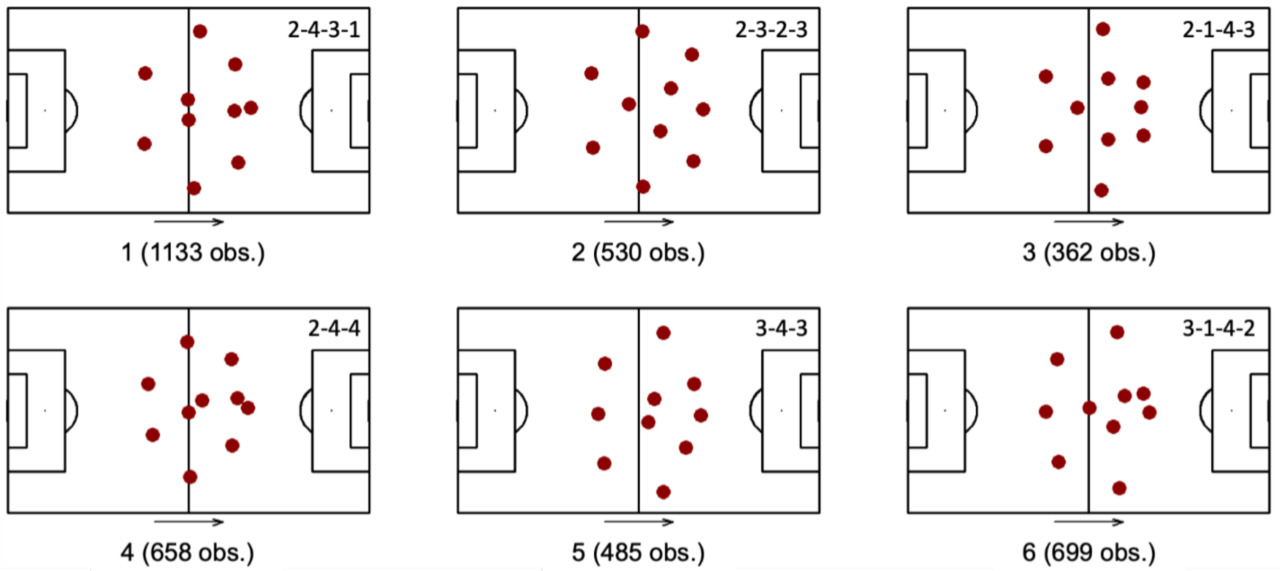

Fig. 10: Outcome of the clustering for build-up including the number of observations (obs.) of our sample.

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.

E Appendix—Study V: Individual Role Classification for Players Defending Corners in Football (Soccer)

Research Article

Gabriel Anzer, Pascal Bauer*, and Joshua Wyatt Smith

Individual role classification for players defending corners in football (soccer)

Categorisation of the defensive role for each player in a corner kick using positional data

https://doi.org/10.1515/sample-YYYY-XXXX
Received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY

Abstract: Choosing the right defensive corner-strategy is a crucial task for each coach in professional football (soccer). Although corners are repeatable and static situations, due to their low conversion rates, several studies in literature failed to find usable insights about the efficiency of various corner strategies. Our work aims to fill this gap. We hand-label the role of each defensive player from 213 corners in 33 matches, where we then employ an augmentation strategy to increase the number of data points. By combining a convolutional neural network with a long short-term memory neural network, we are able to detect the defensive strategy of each player based on positional data. We identify which of seven well-established roles a defensive player conducted (player-marking, zonal-marking, placed for counterattack, back-space, far-post, near-post, and far-post). The model achieves an overall weighted accuracy of 89.3%, and in the case of player-marking, we are able to accurately detect which offensive player the defender is marking 80.8% of the time. The performance of the model is evaluated against a rule-based baseline model, as well as by an inter-labeller accuracy. We show three concrete use-cases on how this approach can support a more informed and fact-based decision making process.

Keywords: Sports analytics, Football (Soccer), Tactical performance analysis, Applied machine learning, Positional and event data

Data science to support decision making in sport has become more predominant over the years. This is especially the case for baseball [1, 2], American football [3], and basketball [4–6]. The application of such methods in football (soccer) is only more recently gaining attention in literature [7–10]. Achieving a comparable impact on roster and in-play decisions is a more sophisticated challenge due to the complexity (i.e. low-scoring nature, invasive character) of the game. However, considering a repeating, static set-piece scenario (e.g. corner kicks) substantially reduces the complexity. Compared to baseball or American football, set-pieces only present a small fraction of a football match, nevertheless, their relevance is pointed out in literature [11, 12] and addressed carefully in sport science [13–25] and in practise through the installation of dedicated set-piece coaches by many teams.

The applied strategies for defending corner kicks in the top professional leagues are heterogeneous.[1] A pure player-marking approach (each offensive player is marked by one defender) is the most traditional

1 A video describing the most common roles can be found here: https://drive.google.com/file/d/1tqcxuV9-WXAvoDw4CfvJZsVnHQaxNG5G/view?usp=sharing.

Gabriel Anzer, Hertha BSC Berlin, Berlin, Germany
Institute of Sports Science, University of Tübingen, Tübingen, Germany,
***Corresponding author: Pascal Bauer,** DFB-Akademie, Deutscher Fußball-Bund e.V. (DFB), Frankfurt, Germany
Institute of Sports Science, University of Tübingen, Tübingen, Germany, pascal.bauer@dfb.de
Joshua Wyatt Smith, Department of Mathematics & Statistics, Concordia University, Montreal, Canada
Wyatt AI, Montreal, Canada, jwsmith@wyattai.com

tactic. In contrast, some teams prefer a pure zonal-marking approach (each defensive player is assigned to a specific area close to the goal). Power et al. [26] pointed out that a hybrid strategy, combining player- and zonal-marking is used most frequently, nowadays. Another crucial decision is how many players should be dedicated to defend the corner, and how many should be positioned for a potential counterattack. Traditionally, at least one player is placed close to the mid-line in case of a turnover. A controversial strategy decision is whether to position players at the posts [21]. Proponents of post-marking will argue that these players reduce the area of the goal that the goalkeeper has to cover by defending critical space next to the posts. Many modern coaches argue that having an additional player in a more pro-active role (e.g. player-/zonal-marking) prevents more shots, and therefore more goals in the long run. However, their benefit is not as obvious compared to post-marking players clearing situations that would otherwise lead to goals. Even if controversial practitioner discussions are ubiquitous since the existence of corners, statistical analysis conducted so far on the respective efficiency has only touched the surface.

Recent success in the area of computer vision has enabled off-the-shelf availability of highly accurate player and ball tracking data across all professional football leagues [27, 28]. Collection of players' coordinates now happens in an automated way, allowing for more advanced algorithms to gain new insights into the data. With respect to analyzing corners, Power et al. [26] were the first to make use of this data and looked at 12, 000 corners from the English Premier League (2016/2017 season). They trained a (supervised) neural network to detect player- versus zonal-marking on a team level using the players' positions and hand-collected labels. Their finding that 80% of teams defend corners using a hybrid approach means that analyzing which players are assigned to specific attacking players could be explored with more granularity. Thus, to determine the role of each defensive player, Shaw et al. [29][2] hand-labelled 500 corners and trained an extreme gradient boosting model based on 10 hand-crafted features. However, in the case of player-marking, still missing was the ability to identify which attacking player was being marked. In basketball, the problem of who is marking whom was solved for open play situations using hidden Markov models [31]. However, the problem of detecting individual roles during corners in football is more complex. This is due to the fact that there are roughly double the number of players (if goalkeepers are excluded), and that players can have different roles other than just player-marking.

The objective of this paper is to accurately classify the role of each defensive player. We detect whether a defender is *player-marking* (PM), *zonal-marking* (ZM), defending a *short-corner* (SD), marking the *near-post* (NP) or *far-post* (FP), defending the *back-space* (BS), or being positioned for a *counterattack* (CA). For those defenders that are classified as PM, we also detect which opposing player is marked. The automated and accurate detection of defensive roles provides many practical use-cases to match analysts and coaching staffs, e.g. for opponent analysis or player scouting.

The organisation of this paper is as follows: Section 1 describes the data used to train the algorithm and the definitions used for the above classes. Section 2 details the process in-which a baseline model was created using domain specific rules (Section 2.1), as well as the augmentation of the original dataset (Section 2.2) increasing the number of training samples. The neural-network algorithm is described in detail in Section 2.3 (Model) and Section 2.4 (Training). Section 2.5 describes the results of the classification process and compares a test dataset to hand-labelled data, consolidated between multiple domain experts. Section 3 discusses the diverse range of applications an algorithm such as this has, and Section 4 concludes with a discussion as well as what we foresee as possible future studies.

2 Note that a first version of the approach was also presented at the 7th International Workshop on Machine Learning and Data Mining for Sports Analytics [30].

1 Data and Definitions

1.1 Data

This study makes use of data from professional German teams provided by the Sportec Solutions AG.[3] In total, we make use of 33 matches containing 213 corner kicks (after excluding corners that were played short, or the ones containing data anomalies).

Positional data for each match is collected using optical tracking techniques with footage from cameras installed in each stadium. The Chyronhego[4] TRACAB system is used. All players and the ball positions relative to the pitch boundaries are provided, as well as metrics such as speed and acceleration sampled at 25 Hz. Several studies have evaluated the quality of positional data [27, 32, 33], especially the TRACAB system [28]. In order to locate corner kicks, we make use of so called *event data*, acquired by human operators tagging a predefined set of events manually. Since event data provides us a rough estimate of the time-frame the corner was played, we make use of positional data to get the exact frame following the method of Anzer et al. [34, 35]. The event data is also used to filter out "short-played" corner kicks, since we are only interested in corner crosses.

Tracking data of the players is taken from 0.5 s before the actual kick up until 0.8 s after the kick. This reduces noise in the players' movements and provides a more accurate representation of their trajectories during the critical time-window of our investigation. The selected time-window was determined by football experts observing video footage. Each corner is normalised to be taken from the bottom-right side of the pitch.

Figure 1 shows an example of the normalised tracking data for three separate corner kicks and the respective trajectories of the players and the ball. In each scenario, players of the defending team can be seen to deploy various strategies to defend their goal, either marking specific players, or specific zones. The definitions are discussed in the next section.

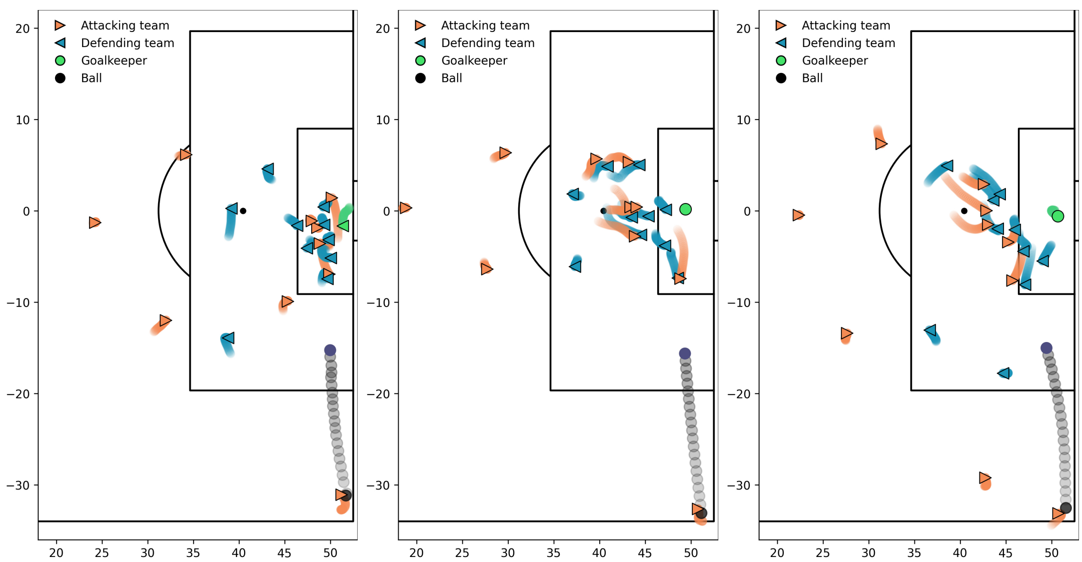

Fig. 1: Three separate corner kick scenarios showing the trajectories of all players and the ball 0.5 s before and 0.8 s after the actual kick. The attacking team plays from left to right and is shown in red. The ball path is displayed in black.

3 https://www.sportec-solutions.de/, accessed 07/22/2021.
4 https://chyronhego.com/, accessed 07/04/2021.

1.2 Definitions for defender roles

Before professional football matches the coaching staff typically provides each player with a defensive assignment covering almost all scenarios (substitutions, different scores, expulsions etc.). However, this information is only available to the respective team's coaching staff. Furthermore, players may adapt during corner kicks, or just not follow the original instructions. Therefore our focus is to detect the strategy for all defensive players during a corner kick, specifically until the ball passes the 16 m vertical boundary of the box.

It is important to note that strategies can be (and often are) changed dynamically depending on the progression of the ball, either from the original kick, or from interactions with players. Players also start approximating where the ball will land and consequently change their strategy. Furthermore, not all players position themselves for the assigned role at the same time. Some start marking their zone or opponent immediately after the ball went out of bounds for the corner, other players can take a while to find their position. However, everything begins with initial role assignment, which defines how the corner kick progresses. The possible assignments for the defending team are defined as:

- **Player-marking (PM)**: A player's objective during the corner is to prevent a specific opposing player from getting the ball. PM defenders position themselves close to their assigned attacker and follow their movement into the target area. In some cases, the defender may initially position themselves so that they intercept the intended run of the attacker. As the corner progresses, the perfect positioning is usually as close as possible to the attacker, on the side of their own goal.
- **Zonal-marking (ZM)**: A player's objective is to defend (and usually clear) the ball from a specific area. ZM players assume static positions close to the goal. They can attempt to intercept a ball by running towards the crossed ball in the direction of play.
- **Near-/Far-post (NP/FP)**: A player is post-marking if they are positioned directly next to the post with the primary aim to block a shot resulting from a corner. A post-marking player can either defend the NP or the FP, relative to which post is closer to the side of the corner kick. Player's are only considered NP/FP if they hold their position until the ball is cleared.
- **Short-defender (SD)**: A player is classified as a SD if they are positioned in such a way as to either prevent a short pass from being played, or to attack the respective players in case the corner is played short. Players that start in a ZM position and step out towards the corner only if the ball is played short are considered as ZM, since this is their primary intention.
- **Back-space (BS)**: A player is positioned close to the horizontal 16 m line (i.e. outside the usual target area of corners). It is their primary objective to get either the so called "second ball" after an initial header, or to defend longer distance shots from the back-space.
- **Counterattack (CA)**: A player typically aims to be a pass option in the case of turnover-situations and/or to receive long-distance clearances resulting from the corner. They are not defensive players in a corner kick.

Not all roles are mandatory during corners, and any combination of the above labels can occur. To ensure a consistent understanding (an "inter-labeller accordance"), we consolidated definitions for each strategy among the labellers.

2 Methods

2.1 Rule-based baseline model and rule-supported labelling

Domain experts (professional football match analysts and coaches; see Acknowledgements) use the above definitions to accurately label individual players' roles by observing the video footage. labelling a single corner means assigning 10 roles[5] for a time-window of 1.3 seconds, which can be technically demanding. Thus, we make use of rule-based functions to create an initial dataset of "weakly" labelled data. These rules range from purely geometric "cuts" (i.e. if a player is standing more than away 25 m from the goal-line they are a CA player), to geometric properties of the player in relation to the opposition and/or goal. Table 1 summarises the domain rules. At times rules may conflict with each other (i.e., a player that is "post-marking" according to rule 1 could also by labelled as "player-marking" qua rule 2). Thus, a player is assigned a final label via a majority vote based on the different rule predictions. It is therefore beneficial to have as many rules as possible for non-geometric classes. *Geometric rules* are prioritised. For example, if a player is more than 25 m away from the goal, they are a CA player, regardless of other classifications. Both majority vote and the prioritisation of geometric rules optimised the balanced accuracy on the hand-labelled data.

This initial dataset serves two valuable purposes: First, we use the rule-based predictions as a baseline model to evaluate the outcome of the CNN/LSTM approach. Second, it allows expert labellers to more quickly label data and by simply correcting labels they deem incorrect. A typical setup comparing the 2-D visualisation, video footage, and labels is shown in Figure 2. This setup drastically decreased labelling time, compared to one without any precomputed recommendations.

In order to avoid biases during our custom setup labelling process, for the purpose of this paper, two experts independently created the initial dataset. One expert labelled the data without the use of the framework, thus not seeing (or being "manipulated") by the rule-based labels, while the other one did. The labellers agree on 94.9% of all data-points. Each labeller had the chance to annotate players of an unclear assignment with "abstain". Excluding all scenes which were marked as abstain by at least one of the labellers yields an accordance of 98.1%. The labeller without rule-based support needed on average about 30% longer per corner.

Fig. 2: labelling configuration showing the radar visualisation, video footage (tactical camera), and output from the custom framework.

The inter-labeller reliability study showed that the definitions are reliable and the distinct strategies can be identified. Especially relevant for our use-cases is the detection of the player-marking strategy and the respective assignment. Both can be accurately detected by domain experts (PM-role: 94.3%; PM-assignment: 87.1%[6]; see Section 2.5). The discussions showed that ZM is often mixed up with other static roles like BS or short defender. In all considered cases, PM was only confounded with ZM. For some "static" strategies

5 Note that experts were not interested in strategies conducted by goalkeepers for this study.
6 Calculated as the accuracy of obtaining PM (0.943) times the accuracy of obtaining the correct assignment (0.924).

Tab. 1: Domain rules applied to raw positional data.

Nr.	Type	Rule
1	Post-marking†	**Distance to post:** A player is labelled as a NP or FP defender if they stay within a radius of 3 m from a post for more than half the considered time of the corner kick.
2	Short-defender†	**Distance to corner:** A player positioned at the corner of the pitch (less then 20 m away from side line and less then 20 m away from goal line).
3	Counterattack †	**Distance to goal:** A defensive player who is more than 25 m away from the goal in the x-coordinate direction.
4	Back-space †	**Backspace corridor:** A player who is within a corridor of 10 m above and below the centre line of the pitch (y-direction), as well as between 15 m and 25 m from the goal line (x-direction).
5	Zonal-marking	**Average speed:** The average speed of the defender is below 5 km/h.
6	Zonal-marking	**Distance covered:** The distance covered during the considered time is less than 4 m.
7	Player-marking	**Horizontal (Vertical) movement difference:** The difference between the summed $x/(y)$-coordinate values for the defender and the closest attacking player are ≤ 10 m.
8	Player-marking	**Close to teammate:** A defending player is within 3.5 m of their own teammate. The closest attacking player is designated as the marked player.
9	Player-marking	**Opposition tracking:** The closest opposing player to a defender (within a 4 m radius) is the same at the beginning and end of the time window considered.
10	Player-marking	**Bezier curve comparison:** The defender's bezier curve shares a very small difference (determined by inspection) with the attacking player's curve.
11	Player-marking	**Placement relative to goal:** The defender has a closer mean distance to goal centre and a smaller mean angle to the line between attacker and goal centre. A cutoff is determined via inspection.
12	Player/Zonal-marking	**Distance to opposition:** If a defender's mean distance to the closest opposing player is below a threshold, assume player-marking, otherwise assume zonal-marking. The threshold is determined via an exponential fit between two data points that satisfy the condition $y(x) = \begin{cases} 1.8 \text{ m}, & \text{if } x = \text{goal line} \\ 7 \text{ m}, & \text{at half-way line,} \end{cases}$ where the boundary conditions were determined by inspection.
14	Player/Zonal-marking	**Naive x-position:** If the defender starts within 6 m from the goal line assume zonal-marking. Otherwise, retrieve the closest opposition player and assume player-marking.

† = Geometric rules

the pure positioning of a player is enough to classify the role of the player (i.e. NP, FP, CA). However, more complex roles need contextual information from the corner. In the static cases (NP, FP and CA) the geometric rules already allow for such an accurate classification, that these are excluded in the following steps. An example strategy where simple geometric rules are not appropriate for an accurate detection is defending BS. Confusion often occurs between BS, SD, and ZM defenders and can not be inferred by solely looking at the positioning on the pitch. However, the respective role of a player is clear to experts when considering the context of each corner (e.g. trajectory of the ball, positioning of other players).

2.2 Data augmentation

We augment our data to increase the number of overall data points to be used in the training process, and to expose the NN to numerous new and unique scenarios that may occur. Based on how far a player is away from the goal line d, the x-coordinate of that player is perturbed such that $x \in [x - f(d), x + f(d)]$, where $f(d) = 0.1 \times 1.09^d$. Similarly, this is done for the y-coordinate. For example, a player who is covering the NP will be perturbed very little. However, other players will be exponentially perturbed the further away from the goal line they are, resulting in a different datapoint for the NN. Coordinates that fall outside of the field are set to be the field limits in the respective direction. The chosen coefficients are based on discussing their effects in different scenarios with experts.

Figure 3 shows an example of the original data compared to a perturbed dataset for the same corner. Differences can particularly be seen between the cluster of players at the approximate coordinates $(40, 5)$.

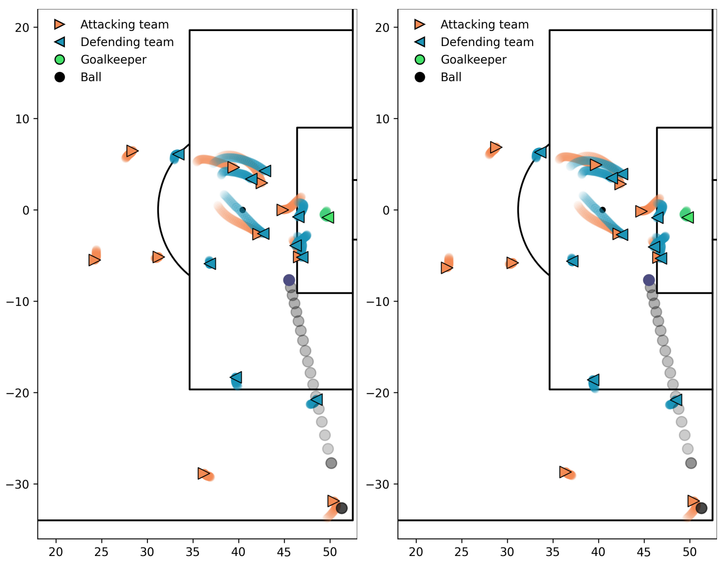

Fig. 3: Original corner (left figure) versus an example of the same augmented corner (right figure).

2.3 Model

A major extension in our approach compared to Power et al. [26] and Shaw et al. [29] is to classify the assignments of individual players in the case of PM. To enable such classification, two models with two separate inputs are trained simultaneously and joined in a final concatenation layer. The first NN, called the *role-classification network* (RCN), consists of stacked convolutional layers. The second network, called the *player-assignment network* (PAN), consists of stacked LSTM layers. A final dense layer combines the outputs of both NNs to provide a prediction such that $y \in \{\text{BS}, \text{SD}, \text{ZM}, \text{PM}\}$, while the PAN determines if the defender was assigned to a given attacker ($y \in [0, 1]$), if the RCN's output is PM. Consequently, a defender is assigned as a PM of an attacker if the RCN classifies them as PM. If that happens, the highest probable attacker from the output of the PAN is assigned as the player the defender is marking.

The input to the RCN takes the form of RGB images where a single defender's (x, y) coordinates are plotted against the whole attacking team in a 130×125 grid. The three color channels reflect each team and the ball. The PAN takes sequential frames of the (x, y) coordinates and speed of the same defending player and a single attacking player. Thus, we construct 10×10 data-points out of a single corner (before data augmentation is taken into consideration). Figure 4 shows the overall architecture, while Figure 5 shows examples of inputs to the RCN.

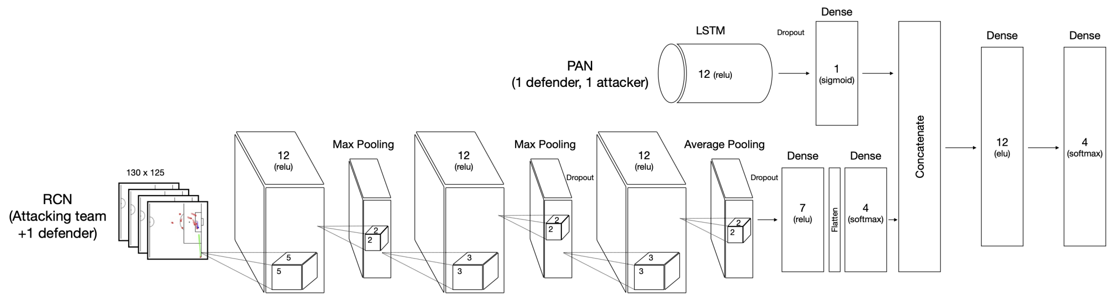

Fig. 4: Illustration of the overall model architecture. The PAN takes as input the features of a single attacking and a single defending player. The RCN takes RGB images of size 130×125 consisting of the full attacking team, a single defender, and the ball.

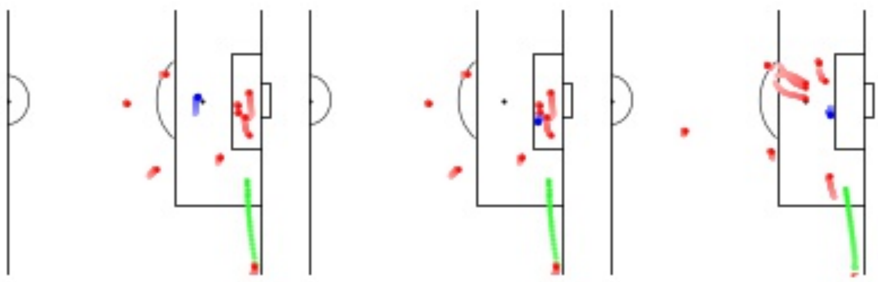

Fig. 5: RCN input examples for three different corner scenarios. Each team and a ball dominates the respective RGB channel, with light to dark trails indicating direction of travel, scaled by speed. The defender is shown in blue, the attacking team in red and the ball is displayed through the green channel.

2.4 Training

The model was trained on a NVIDIA T4 GPU. A grid search for optimum parameters was performed. Models were trained for 75 epochs, or the loss function showed negligible improvement over 15 epochs. Table 2 shows the final hyperparameters selected. A series of jobs were run, where Bayesian optimization was used to select successive regions, thus narrowing down the phase space. The final model was trained

Tab. 2: Final configuration for RCN and PAN parameters. Stride and pooling sizes are square.

	Hyperparameter	Description	Final Value
1	Learning rate	Step size per iteration	9×10^{-4}
2	Batch size	Number of training examples per iteration	800
3	RCN conv 1 (size, stride)	First convolutional layer	(12, 5)
4	Max pooling 1 stride	Downsample feature map using max values	2
5	RCN conv 2 (size, stride)	Second convolutional layer	(12, 3)
6	Max pooling 2 stride	Downsample feature map using max values	2
7	RCN conv 3 (size, stride)	Third convolutional layer	(12, 3)
8	Average pooling stride	Downsample feature map using average values	2
9	RCN dense	Fully connected layer	12
10	RCN dropout	Random fraction of nodes dropped in RCN	0.2
11	PAN dropout	Random fraction of nodes dropped in PAN	0.4
12	PAN LSTM size	Number of units in LSTM	12
13	RCN+PAN dense	Fully connected layer combining RCN and PAN	12

using an Adam optimiser [36], with a learning rate of 9×10^{-4} and batch size of 800.[7] An exponential learning rate decay was used after five epochs showed no improvement in the loss function. The model finished training when no improvement was seen in the loss function. The dataset was split into training (50%), validation (25%), and test (25%) sets, where the ratio of each class mentioned in Section 1.2 is maintained. Metrics obtained from the validation dataset are used to determine early stopping in the training process, as well as selection of new parameter space in the hyperparameter optimisation. During training, class weights for each of the defending categories are calculated on a batch-by-batch basis and applied to the categorical cross entropy loss function. According the augmentation described in section 2.2 we create in total ten augmented data-files for each corner. The separation between training and test datasets is maintained. To reduce the imbalance of the data for the PAN (a binary classifier for PM or "not-PM"), of the ten extra augmented data files per match, only data points that are considered PM are used for five of those augmentations.

2.5 Results

When evaluating the model, the order of how the NN makes predictions is important. First, the data point is predicted to be in one of the 4 classes, {BS, SD, ZM, PM}.[8] If the outcome is PM, only then is the PAN consulted and its prediction obtained. Since a wrong player-assignment of the PAN is not taken into consideration in the case of non-player-marking, an important metric for the PAN is in fact recall, while for the RCN it is a weighted or "balanced" accuracy. Figure 6 shows the confusion matrix using the test dataset for the RCN. A good classification can be seen among the different classes. Table 3 summarises these results per class. The overall balanced accuracy is 89.3%. A player is correctly classified as PM 90.3% of the time, additionally with the correct offensive player assignment occurring 80.8% of the time.[9] The accuracy for each class using the rule-based approach from Section 2.1 is included as a baseline model for comparison.

7 Due to how a single data point is constructed (Section 2.3) and fed into the NN, the total number of data points ends up being quite large with similar or even identical attributes (in the case of the inputs to the RCN). For this reason, large batch sizes and regularisation methods are essential.

8 For the purpose of this paper, the classes FP, NP and CA are removed since obtaining them geometrically is just as accurate and significantly more stable given the limited number of data points in these categories.

9 Calculated as the probability that PM is correctly predicted (0.895), multiplied by the recall of the PAN (0.903).

10 Calculated as the accuracy of obtaining PM (0.943) times the accuracy of obtaining the correct assignment (0.924).

11 Calculated as the accuracy of obtaining PM (0.726) times the accuracy of obtaining the correct assignment (0.855).

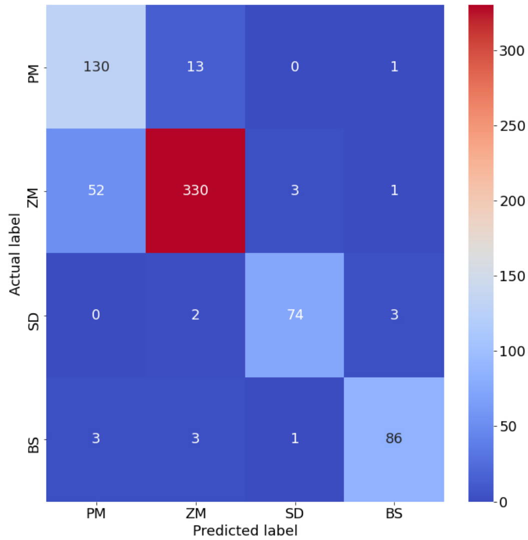

Fig. 6: Confusion matrix for the RCN on unseen test data. Numbers indicate the players in each class.

Tab. 3: Summary of the different classes and the corresponding number of data for each used in the training phase, as well as the overall classification achieved by our model on a test dataset. The inter labeller accordance is calculated on two matches.

Action	Data points	Aug-mented data	Inter-labeller accordance	Rule-based accuracy	Test data accuracy
Player-marking Class	743	6,634	94.3 %	72.6%	90.3%
↪ Correct Assignment			87.1%[10]	62.1%[11]	80.8%
Zonal-marking	793	4,843	89.6%	62.7%	86.9%
Near-post	15	105	93.3%	93.3%	-
Far-post	14	134	100%	100%	-
Counterattack	73	493	100%	100%	-
Short Defender	184	1,234	95.2%	85.7%	93.7%
Back-space	274	2,084	94.7%	70.3%	92.5%

3 Practical Application

Our algorithm is able to accurately determine individual defender roles in detail. This allows for a wide range of applications that can further be explored. We aim to outline a few of these use cases.

Use-Case 1: Automated match-report to monitor corner performance

Teams spend vast amounts of resources to prepare their strategy during corners for their upcoming matches. However, when analysing the performance of a team post-match, an objective quality assurance is often neglected or simply focuses on the few (possibly random) events, where goals were actually scored. Figure 7 shows an excerpt of a more granular match-report[12] that is used by the German national teams in post-match analysis. It provides an overview of the roles detected across all corners and can help to extract

12 The plot shows the post-match analysis of the U21 match Germany against Denmark in the round of the last 16 at the U21 European Championship 2021.

insights at a fraction of the time it would take match analysts. Player #10 (green) of the German team for example, was marked by #20 (red) of the Danish team during two corners. Interestingly, he scored a goal during the one corner where he was marked by #15 of the Danish team. From a defensive perspective, an insightful performance metric for ZM is how often they were able to reach the ball first ("first touch"). Figure 7 shows that number #13 and #15 (red), detected as zonal defenders (ZM), touched the ball first in three of the corners. Another insight that can easily be retrieved from this figure is that #19, #8 and #7 on the German team were not player-marked by an opponent, with #19 getting a first touch, and all three attempting a shot on goal following a corner kick.

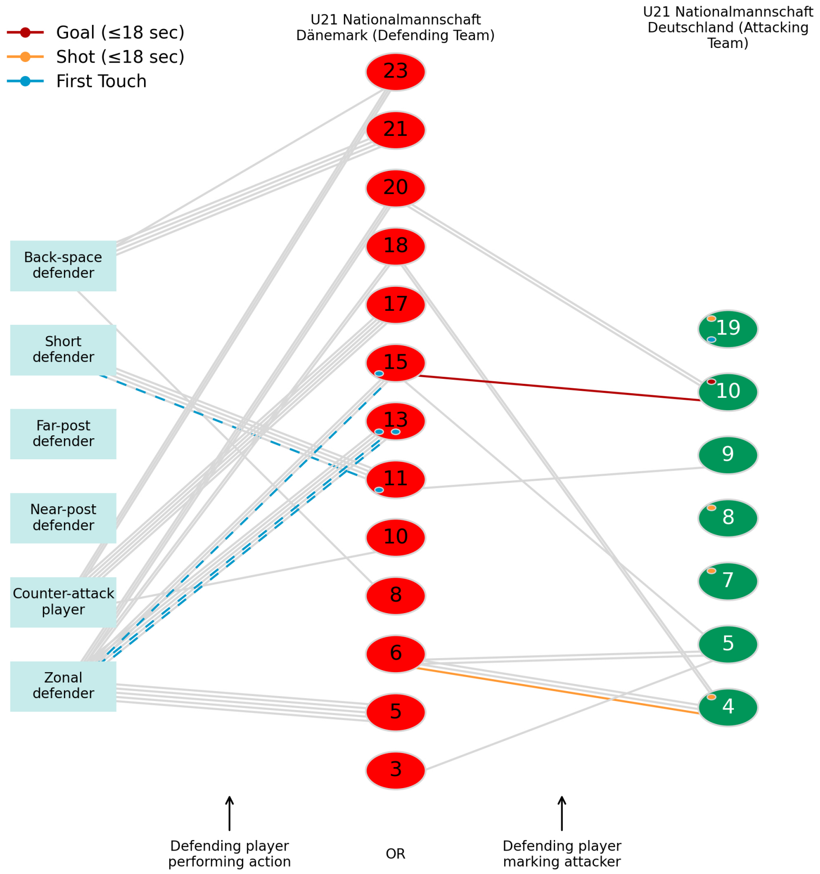

Fig. 7: Excerpt from the match-report of a European Championship match of Germany U21 against Denmark. The plot gives an overview of the player assignments of all corners where Denmark was defending, as well as their outcomes. For the defending team (jersey numbers in red) the role is indicated by the lines to the blue role-boxes. In case of player-marking, the player assignment is displayed by the connection to the green boxes (jersey numbers of the attacking team).

Use-Case 2: Analysis of individual players

Only about one in every 25 corners leads to a goal. It is not necessarily the best offensive players that take shots but rather depends on a number of factors such as the trajectory of the ball, how well a player is marked, and many other factors. Therefore, it is hard to objectively evaluate the performance of individual players. By automatically analysing several seasons of a player, an indication of individual quality for aerial duels can be given. By comparing how many shots and/or goals player-marked attackers create per defender, we can compare attackers and defenders to each other. Our approach allows us to spot players with the most first touches (per corner) when ZM, or the best PM-defenders preventing their assigned attackers to reach the ball or even create a shot/goal. This is a major extension compared to existing literature.

Individual analysis can also be used for opponent scouting. Match analysis departments analyse the upcoming opponent's corner strategies by observing such situations on a weekly basis. Our approach can support that process and gives more insights by looking at specific situations that occur sporadically over a season, but cannot be covered in detail by manual annotation due to time constraints (e.g. how does a team react after red-cards, by score, or in the last minutes of the game). Figure 8 shows an overview of individual player strategies of the German U21 national team across their matches in the European championship in 2021. This can be used efficiently for opponent analysis in order to get a first overview of the player's strategies.

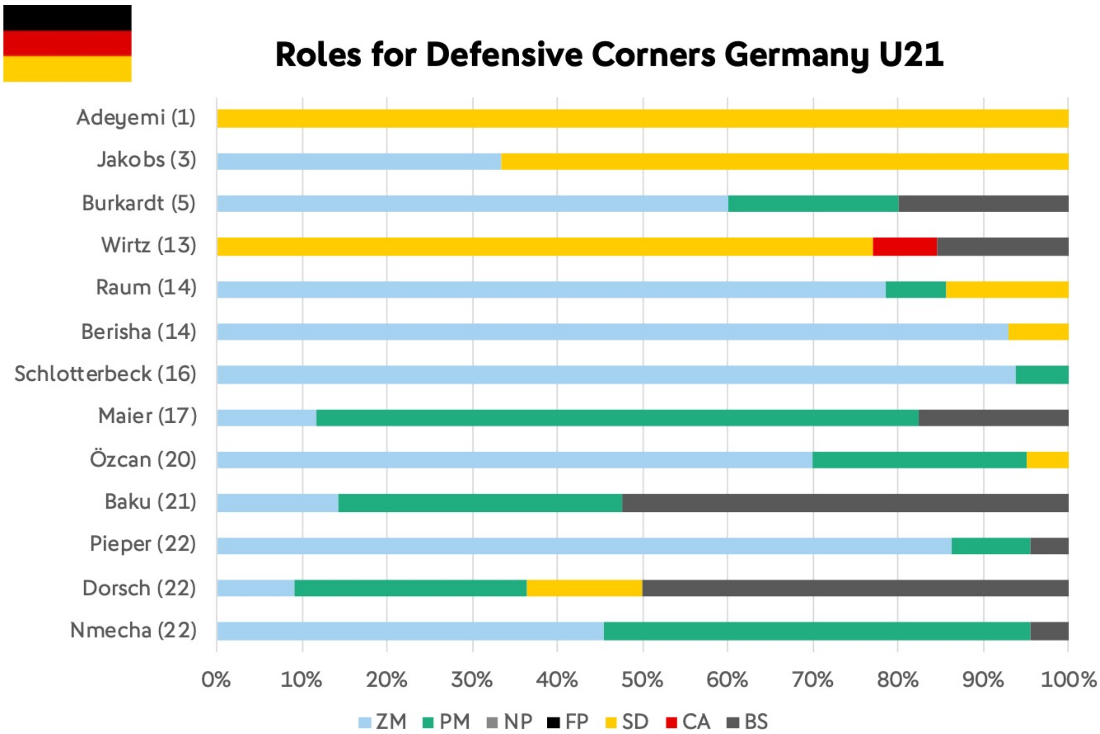

Fig. 8: The figure shows the defensive corner roles of the German U21 National Team across all matches during the 2021 European Championship. The values in parentheses indicates the number of corners a player participated in.

Use-Case 3: Long-term analysis of strategies and their efficiency

Due to the effort of manual detection, no detailed statistical investigations about whether some general strategies are more effective than others have been conducted so far. While Power et al. [26] presented a first indication about which strategy was applied more efficiently in the English Premier League season 2016/2017 on a team level, our approach enables us to understand the dominant category of hybrid marking (80% of the cases) in greater detail. Using our automated detection, we can compare the efficiency of all presented strategies for a team and/or on the individual player level on a long term basis. Shaw et al. [29] already pointed out this use case by raising the question *"Which attacking routines are most effective against a certain defensive set-up?"*, which could be answered using a sufficient sample size. Using our model, we can go one step further and suggest strategies on a player-level (i.e. which defender should be assigned to which attacker to minimize the goal scoring probability) instead of purely suggesting team strategies.

4 Discussion and Future Work

Currently, defensive strategies are chosen intuitively and heuristically in professional football. Coaches try to make informed decisions on whether to use PM, ZM or a hybrid form, and which opponents should be player-marked by whom. These decisions are typically prepared per match, where they try to optimise specific role-assignments based on strength and weaknesses of their own players as well as the opponent.

Manually annotating corner strategies, or their results, has a long history in sport sciences [13–25]. These hand-created annotations are limited in size, and therefore do not allow for long-term analysis. A Bundesliga team concedes on average only about 5.3 goals from corners per season (34 matches).[13] Hence, analysing only a single season or even only a single international tournament does not yield many significant insights into the efficiency of different corner strategies. Power et al. [37] performed a first step towards automating some manual annotations by detecting PM versus ZM on a team level. Hybrid marking is predominant nowadays, but team strategies vary between two and eight zonal-markers within that hybrid category. Power et al. [37] found that a hybrid strategy is used in 80.0% of the cases (English Premier League, 2016/2017 season), while our dataset (from matches after 2019) consists purely of corners defended in a hybrid formation. Shaw et al. [29] presented an approach detecting roles on a player level, however, individual player-marking assignments were not detected. For ZM, they quote 81.0% for both precision and recall on a training dataset. We improve upon in this score, achieving an accuracy of 86.9% on an unseen test dataset. Our work extends previous studies and solves a practical and relevant problem with sufficient accuracy. In the following, we lay out two major possibilities to further improve our current approach using recent developments in machine learning.

Using weak supervision to reduce the labelling effort

The combination of positional and event data provide a detailed reproduction of professional football matches, but most tactical strategies are too complex to detect them using a purely rule-based strategy. Nevertheless, rules can serve as a solid starting point. Elaborate supervised machine learning models (trained on human-made labels) are well established in football analytics literature—not only for corners [26, 29], but also for open-play strategies like counterattacks [38, 39], counterpressing [40] or patterns like overlapping runs [41]. Many rules aiming to detect tactical patterns require thresholds (see table 1), which can be subjective among experts. In the special case of corners, those thresholds (e.g. the distance to the goal until a player is considered a BS defender) cannot be set once for all corners but rather depend on several factors (e.g. whether the corner is an in-swinger or an out-swinger), or on the philosophy and instruction of the coach (i.e. some coaches want their BS defenders closer to the goal than others). Expert-labelling, used for supervised learning to overcome this issue, is very time-consuming and not ideal for rapidly evolving strategies. Individual strategies can even vary in nuances depending on different team-philosophies.

13 This average is calculated based on six seasons (2014/2015 until 2019/2020) of German Bundesliga and German 2. Bundesliga. Goals are counted whenever they occur within 18.0 seconds after the corner was executed.

Accordingly, approaches using semi-supervised methodologies are well suited for football specific problems. Ratner et al. [42] designed a weakly-supervised learning approach that significantly reduces the amount of required labels with minimal trade-off in accuracy for different application areas (e.g. natural language processing Hoffmann et al. [43], or saliency detection Zeng et al. [44]).

In our scenario, the rules formulated in section 2.1 to define our baseline model, could be used for such a weak-supervision approach. We tested this methodology by training our RCN/PAN on a randomly selected 75% of the hand-labelled data, with the remaining 25% of the labels coming from a majority vote of the rules defined in Table 1. We achieved similar results (balanced accuracy RCN: $89.6 \pm 1.6\%$; balanced accuracy PAN: $85.1 \pm 1.3\%$). This means that the labelling time can be further reduced without sacrificing accuracy.

Using graph neural networks to reduce the computing complexity

Another recent methodology that can be used to improve our work in future investigations are graph neural networks (GNN's) carefully discussed by Battaglia et al. [45]. Graphs model football as a multi-agent set-up, in which 22/23 agents (typically the ball is also modelled as an agent) interact with each other. A problem when modelling invasion sports is that no trivial ordering is given for the players. This issue with multi-agent classification problems on a team and a player level have been overcome in sports analytics literature using CNN's [46] and artificial orderings [47, 48]. An advantage of CNN approaches in football is that outliers in player trajectories are efficiently eliminated by the CNN. This is especially helpful in set-piece scenarios where many players interact with each other in a small space and tracking systems face many occlusions. A GNN that models players/agents as nodes and their interactions as edges, can overcome this problem very efficiently by creating permutation-invariant embeddings [49]. Combining a graph structure with recurrent units allows to model temporal sequences efficiently as well [50, 51].

Yeh et al. [52], Sun et al. [53], Kipf et al. [54], and Games [55] used graphs to predict players' trajectories in invasion sports (i.e. basketball and football), taking all player interactions into consideration. On static data, Stöckl et al. [56] used GNN's to predict future events in football via node-predictions. Dick et al. [57] were the first to present edge predictions on football data using a graph recurrent neural networks (GRNN) in order to perform classification tasks on player interactions. Given the recent development in the theory around graph neural networks, a combination of node-predictions (replacing the RCN) and edge-predictions (solving the task of the PAN), could drastically reduce computation power when compared to our current approach.

5 Conclusion

We detect defending corner strategies on a player level with a high accuracy using a combined NN approach. Data augmentation helped us to achieve a high generalisation, despite a low sample-size. Simple geometric rules reduced significant amounts of labelling time and can be used as labelling functions in a weak-supervision scenario. This allows us to add new data to our approach and improve the accuracy further with limited additional labelling effort. We improve on the accuracy compared to existing the literature, as well as include a novel extension, which is of high value for practitioners in professional football.

Acknowledgment: 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 men's National team), Jannis Scheibe (head match analyst of the German U21 men's national team), Leonard Höhn (head match analyst for the German women's national team) as well as Sebastian Geißler (former match analyst of Borussia Mönchengladbach).

Funding: Please insert information concerning research grant support here (institution and grant number). Please provide for each funder the funder's DOI according to https://doi.crossref.org/funderNames?mode=list.

Ethics and Reproducibility By informing all participating players, all tracking is compliant to the general data protection regulation (GDPR)[14]. 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 vendors and systems, peers working in the football industry can reproduce the results by using any kind of professional football data.

References

[1] Ramy Elitzur. "Data analytics effects in major league baseball." In: *Omega (United Kingdom)* 90 (Jan. 2020), p. 102001. ISSN: 03050483. DOI: 10.1016/j.omega.2018.11.010 (cit. on p. 1).

[2] Tom MacLennan. "Moneyball: The Art of Winning an Unfair Game." In: *The Journal of Popular Culture* 38.4 (May 2005), pp. 780–781. ISSN: 0022-3840. DOI: 10.1111/j.0022-3840.2005.140-11. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/j.0022-3840.2005.140_11.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.0022-3840.2005.140_11.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.0022-3840.2005.140_11.x (cit. on p. 1).

[3] Jeremy Hochstedler and Paul T Gagnon. "American Football Route Identification Using Supervised Machine Learning." In: *MIT Sloan Sports Analytics Conference, Boston (USA)* (2017), pp. 1–11 (cit. on p. 1).

[4] Changjia Tian et al. "Use of machine learning to automate the identification of basketball strategies using whole team player tracking data." In: *Applied Sciences (Switzerland)* 10.1 (2020), pp. 1–16. ISSN: 20763417. DOI: 10.3390/app10010024 (cit. on p. 1).

[5] Matthew van Bommel and Luke Bornn. "Adjusting for scorekeeper bias in NBA box scores." In: *Data Mining and Knowledge Discovery* 31.6 (2017), pp. 1622–1642. ISSN: 1573756X. DOI: 10.1007/s10618-017-0497-y (cit. on p. 1).

[6] Daniel Cervone et al. *A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes*. Vol. 111. 514. 2016, pp. 585–599. ISBN: 8750142011. DOI: 10.1080/01621459.2016.1141685 (cit. on p. 1).

[7] F. R. Goes et al. "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 (2020), pp. 1–16. ISSN: 15367290. DOI: 10.1080/17461391.2020.1747552. URL: https://doi.org/10.1080/17461391.2020.1747552 (cit. on p. 1).

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

[9] M. Herold et al. "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 (2019). ISSN: 2048397X. DOI: 10.1177/1747954119879350 (cit. on p. 1).

[10] Robert Rein and Daniel Memmert. "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science." In: *SpringerPlus* 5.1 (2016). ISSN: 21931801. DOI: 10.1186/s40064-016-3108-2 (cit. on p. 1).

[11] Diego Brito Souza et al. "A new paradigm to understand success in professional football: analysis of match statistics in LaLiga for 8 complete seasons." In: *International Journal of Performance Analysis in Sport* 19.4 (2019), pp. 543–555. ISSN: 14748185. DOI: 10.1080/24748668.2019.1632580. URL: https://doi.org/10.1080/24748668.2019.1632580 (cit. on p. 1).

[12] Gabriel Anzer, Pascal Bauer, and Ulf Brefeld. "The origins of goals in the German Bundesliga." In: *Journal of Sport Science* (2021). DOI: 10.1080/02640414.2021.1943981. URL: https://www.tandfonline.com/doi/full/10.1080/02640414.2021.1943981 (cit. on p. 1).

[13] Vasilis Armatas et al. "Analysis of the set-plays in the 18th World Cup in Germany." In: *Physical training* October.1 (2007) (cit. on pp. 1, 13).

[14] Robert H. Schmicker. "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* 12.2 (2013), pp. 70–79. ISSN: 16844769 (cit. on pp. 1, 13).

14 https://gdpr-info.eu/, accessed 07/20/21

[15] Craig Pulling, Matthew Robins, and Thomas Rixon. "Defending corner kicks: Analysis from the English premier league." In: *International Journal of Performance Analysis in Sport* 13.1 (2013), pp. 135–148. ISSN: 14748185. DOI: 10.1080/24748668.2013.11868637 (cit. on pp. 1, 13).

[16] Pilar Sainz de Baranda and David Lopez-Riquelme. "Analysis of corner kicks in relation to match status in the 2006 World Cup." In: *European Journal of Sport Science* 12.2 (2012), pp. 121–129. ISSN: 17461391. DOI: 10.1080/17461391.2010.551418 (cit. on pp. 1, 13).

[17] Toni Ardá Suárez et al. "Análisis de la eficacia de los saques de esquina en la copa del mundo de fútbol 2010. Un intento de identificación de variables explicativas." In: *Revista de Psicologia del Deporte* 23.1 (2014), pp. 165–172. ISSN: 1132239X (cit. on pp. 1, 13).

[18] Craig Pulling. "Long Corner Kicks in the English Premier League." In: *Kinesiology* 47.2 (2015), pp. 193–201. ISSN: 13311441 (cit. on pp. 1, 13).

[19] Claudio A. Casal et al. "Analysis of corner kick success in elite football." In: *International Journal of Performance Analysis in Sport* 15.2 (2015), pp. 430–451. ISSN: 14748185. DOI: 10.1080/24748668.2015.11868805 (cit. on pp. 1, 13).

[20] Ali Onur Cerrah, Barış Özer, and Ismail Bayram. "Quantitative Analysis of Goals Scored from Set Pieces: Turkey Super League Application." In: *Turkiye Klinikleri Journal of Sports Sciences* 8.2 (2016), pp. 37–45. ISSN: 1308-0938. DOI: 10.5336/sportsci.2016-50745 (cit. on pp. 1, 13).

[21] Craig Pulling and Jay Newton. "Defending corner kicks in the English Premier League: Near-post guard systems." In: *International Journal of Performance Analysis in Sport* 17.3 (2017), pp. 283–292. ISSN: 14748185. DOI: 10.1080/24748668.2017.1331577. URL: http://doi.org/10.1080/24748668.2017.1331577 (cit. on pp. 1, 2, 13).

[22] Daniel Fernández-Hermógenes, Oleguer Camerino, and Antonio García De Alcaraz. "Acciones ofensivas a balón parado en el fútbol." In: *Apunts. Educacion Fisica y Deportes* 129 (2017), pp. 78–94. ISSN: 20140983. DOI: 10.5672/apunts.2014-0983.es.(2017/3).129.06 (cit. on pp. 1, 13).

[23] C. A. Casal et al. "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* 17.68 (2017), pp. 715–728. ISSN: 1577-0354. DOI: 10.15366/rimcafd2017.68.009 (cit. on pp. 1, 13).

[24] Ben William Strafford et al. "Comparative analysis of the top six and bottom six teams' corner kick strategies in the 2015/2016 English Premier League." In: *International Journal of Performance Analysis in Sport* 19.6 (2019), pp. 904–918. ISSN: 14748185. DOI: 10.1080/24748668.2019.1677379. URL: https://doi.org/10.1080/24748668.2019.1677379 (cit. on pp. 1, 13).

[25] Raif Zileli and Mehmet Söyler. "Analysis of corner kicks in FIFA 2018 World Cup." In: *Journal of Human Sport and Exercise* 17.1 (2020). ISSN: 1988-5202. DOI: 10.14198/jhse.2022.171.15 (cit. on pp. 1, 13).

[26] Paul Power et al. "Mythbusting Set-Pieces in Soccer." In: *MIT Sloan Sports Analytics Conference, Boston* 102.2 (2018), pp. 1–12. ISSN: 15730565 (cit. on pp. 2, 7, 12, 13).

[27] Daniel Linke, Daniel Link, and Martin Lames. "Validation of electronic performance and tracking systems EPTS under field conditions." In: *PLoS ONE* 13.7 (2018), pp. 1–20. ISSN: 19326203. DOI: 10.1371/journal.pone.0199519 (cit. on pp. 2, 3).

[28] Daniel Linke, Daniel Link, and Martin Lames. "Football-specific validity of TRACAB's optical video tracking systems." In: *PLoS ONE* 15.3 (2020), pp. 1–17. ISSN: 19326203. DOI: 10.1371/journal.pone.0230179 (cit. on pp. 2, 3).

[29] Laurie Shaw and Sudarshan Gopaladesikan. "Routine inspection: A playbook for corner kicks." In: *MIT Sloan Sports Analytics Conference, Boston (USA)* (2021) (cit. on pp. 2, 7, 13).

[30] Jan Van Haaren et al. *Machine Learning and Data Mining for Sports Analytics*. September. 2013, p. 2013. ISBN: 9783030649111. DOI: 10.1007/978-3-030-64912-8 (cit. on p. 2).

[31] Alexander Franks et al. "Characterizing the spatial structure of defensive skill in professional basketball." In: *Annals of Applied Statistics* 9.1 (2015), pp. 94–121. ISSN: 19417330. DOI: 10.1214/14-AOAS799 (cit. on p. 2).

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

[33] A. Redwood-Brown, W. Cranton, and C. Sunderland. "Validation of a real-time video analysis system for soccer." In: *International Journal of Sports Medicine* 33.8 (2012), pp. 635–640. ISSN: 01724622. DOI: 10.1055/s-0032-1306326 (cit. on p. 3).

[34] Gabriel Anzer and Pascal Bauer. "Expected Passes—Determining the Difficulty of a Pass in Football (Soccer) Using Spatio-Temporal Data." In: *Data Mining and Knowledge Discovery, Springer US* (2022). ISSN: 1573-756X. DOI: 10.1007/s10618-021-00810-3 (cit. on p. 3).

[35] Gabriel Anzer and Pascal Bauer. "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 (2021), pp. 1–18. DOI: 10.3389/fspor.2021.624475. URL: https://www.frontiersin.org/articles/10.3389/fspor.2021.624475/full (cit. on p. 3).

[36] Diederik Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: *International Conference on Learning Representations* (Dec. 2014) (cit. on p. 9).

[37] Paul Power et al. ""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 (2017), pp. 1605–1613. DOI: 10.1145/3097983.3098051 (cit. on p. 13).

[38] Dennis Fassmeyer et al. "Toward Automatically Labeling Situations in Soccer." In: *Frontiers in Psychology (Special Research Topic on Collective Behaviour in Team Sports; accepted for publication)* (2021) (cit. on p. 13).

[39] Jennifer Hobbs et al. "Quantifying the Value of Transitions in Soccer via Spatiotemporal Trajectory Clustering." In: *MIT Sloan Sports Analytics Conference, Boston (USA)* (2018), pp. 1–11 (cit. on p. 13).

[40] Pascal Bauer and Gabriel Anzer. "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 (2021), 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. 13).

[41] Gabriel Anzer et al. "Detection of tactical patterns using semi-supervised graph neural networks." In: *MIT Sloan Sports Analytics Conference, Boston, USA (Accepted for Research paper track 2022)* 16 (2022), pp. 1–3 (cit. on p. 13).

[42] Alexander Ratner et al. "Snorkel: Rapid training data creation with weak supervision." In: *Proceedings of the VLDB Endowment* 11.3 (2017), pp. 269–282. ISSN: 21508097. DOI: 10.14778/3157794.3157797 (cit. on p. 14).

[43] Raphael Hoffmann et al. "Knowledge-based weak supervision for information extraction of overlapping relations." In: *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 1.May 2014 (2011), pp. 541–550 (cit. on p. 14).

[44] Yu Zeng et al. "Multi-source weak supervision for saliency detection." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* (2019), pp. 6074–6083. ISSN: 23318422 (cit. on p. 14).

[45] Peter W. Battaglia et al. *Relational inductive biases, deep learning, and graph networks.* 2018 (cit. on p. 14).

[46] Uwe Dick and Ulf Brefeld. "Learning to Rate Player Positioning in Soccer." In: *Big Data* 7.1 (2019), pp. 71–82. ISSN: 2167647X. DOI: 10.1089/big.2018.0054 (cit. on p. 14).

[47] Hoang M. Le et al. "Coordinated multi-agent imitation learning." In: *34th International Conference on Machine Learning, ICML 2017.* Vol. 4. 2017, pp. 3140–3152. ISBN: 9781510855144 (cit. on p. 14).

[48] Keisuke Fujii. "Data-Driven Analysis for Understanding Team Sports Behaviors." In: *Journal of Robotics and Mechatronics* 33.3 (2021), pp. 505–514. ISSN: 0915-3942. DOI: 10.20965/jrm.2021.p0505 (cit. on p. 14).

[49] Luana Ruiz, Fernando Gama, and Alejandro Ribeiro. "Gated Graph Recurrent Neural Networks." In: *IEEE Transactions on Signal Processing* 68 (2020), pp. 6303–6318. ISSN: 19410476. DOI: 10.1109/TSP.2020.3033962 (cit. on p. 14).

[50] Ehsan Hajiramezanali et al. "Variational Graph Recurrent Neural Networks." In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.* NeurIPS (2019) (cit. on p. 14).

[51] Yujia Li et al. "Gated graph sequence neural networks." In: *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings* 1 (2016), pp. 1–20 (cit. on p. 14).

[52] Raymond A. Yeh et al. "Diverse generation for multi-agent sports games." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June (2019), pp. 4605–4614. ISSN: 10636919. DOI: 10.1109/CVPR.2019.00474 (cit. on p. 14).

[53] Chen Sun et al. "Stochastic prediction of multi-agent interactions from partial observations." In: *Seventh International Conference on Learning Representations (ICLR), New Orleans (USA)* (2019), pp. 1–15. ISSN: 23318422 (cit. on p. 14).

[54] Thomas Kipf et al. "Neural relational inference for Interacting systems." In: *35th International Conference on Machine Learning, ICML 2018* 6 (2018), pp. 4209–4225. ISSN: 1938-7228 (cit. on p. 14).

[55] Multi-agent Sports Games. "A Graph Attention Based Approach for Trajectory Prediction in Multi-agent Sports Games." In: *Preprint (arXiv)* (2019) (cit. on p. 14).

[56] Michael Stöckl et al. "Making Offensive Play Predictable - Using a Graph Convolutional Network to Understand Defensive Performance in Soccer." In: *MIT Sloan Sports Analytics Conference, Boston (USA)* (2021) (cit. on p. 14).

[57] Uwe Dick, Maryam Tavakol, and Ulf Brefeld. "Rating Player Actions in Soccer." In: *Frontiers in Sports and Active Learning (Special Issue: Using Artificial Intelligence to Enhance Sport Performance)* 3 (2021), p. 174. DOI: 10.3389/fspor.2021.682986. URL: https://www.frontiersin.org/article/10.3389/fspor.2021.682986 (cit. on p. 14).

F Appendix—Study VI: Toward Automatically Labeling Situations in Soccer

frontiers in Sports and Active Living

ORIGINAL RESEARCH
published: 03 November 2021
doi: 10.3389/fspor.2021.725431

Toward Automatically Labeling Situations in Soccer

Dennis Fassmeyer[1], Gabriel Anzer[2,3†], Pascal Bauer[2,4†] and Ulf Brefeld[1]*

[1] Machine Learning Group, Leuphana University of Lüneburg, Lüneburg, Germany, [2] Department of Sport Psychology and Research Methods, Institute of Sports Science, University of Tübingen, Tübingen, Germany, [3] Sportec Solutions AG, Subsidiary of the Deutsche Fußball Liga (DFL), Munich, Germany, [4] DFB-Akademie, Deutscher Fußball-Bund e.V. (DFB), Frankfurt, Germany

We study the automatic annotation of situations in soccer games. At first sight, this translates nicely into a standard supervised learning problem. However, in a fully supervised setting, predictive accuracies are supposed to correlate positively with the amount of labeled situations: more labeled training data simply promise better performance. Unfortunately, non-trivially annotated situations in soccer games are scarce, expensive and almost always require human experts; a fully supervised approach appears infeasible. Hence, we split the problem into two parts and learn (i) a meaningful feature representation using variational autoencoders on unlabeled data at large scales and (ii) a large-margin classifier acting in this feature space but utilize only a few (manually) annotated examples of the situation of interest. We propose four different architectures of the variational autoencoder and empirically study the detection of corner kicks, crosses and counterattacks. We observe high predictive accuracies above 90% AUC irrespectively of the task.

Keywords: sports analytics, soccer, tracking data, variational autoencoders, labeling situations

OPEN ACCESS

Edited by:
Rui Marcelino,
University Institute of Maia, Portugal

Reviewed by:
Paizis Christos,
Université de Bourgogne, France
Hendrik Meth,
Stuttgart Media University, Germany

***Correspondence:**
Ulf Brefeld
brefeld@leuphana.de

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

Specialty section:
This article was submitted to Elite Sports and Performance Enhancement, a section of the journal Frontiers in Sports and Active Living

Received: *15 June 2021*
Accepted: *06 October 2021*
Published: *03 November 2021*

Citation:
Fassmeyer D, Anzer G, Bauer P and Brefeld U (2021) Toward Automatically Labeling Situations in Soccer. Front. Sports Act. Living 3:725431. doi: 10.3389/fspor.2021.725431

INTRODUCTION

The acquisition of tracking/positional and event data has become ubiquitous in professional football. The benefits of the resulting digital reproduction of a match, widely available in professional leagues, are twofold: Firstly, coaches, analysts and other decision makers in clubs may use data as an objective and quantitative alternative to traditional analyzes of performance, and, secondly, the collected data enables media to tell automated stories, to provide data-driven insights in what is happening on the pitch.

For example, match-analysis departments have historically spend vast amounts of time analyzing their upcoming opponent before each match by manually evaluating video footage. This work intensive approach is nowadays being supported or even partially replaced by automatic insight generation based on available data. While some information is easily accessible from the collected data, e.g., extracting the preferred formation of a team (Shaw and Glickman, 2019), other (rather tactical) pieces of information cannot be automatically computed yet, either because they are too complex (e.g., how teams behave during counterattacks), depend on the actual game philosophy of a team, require large amounts of tactical knowledge, or are considered a niche with only few interested followers. Detecting such events and patterns automatically offers a huge potential for performance analysis and may revolutionize current pre- and post-match performance analyses in professional football.

When speaking about data in soccer, we differentiate between positional/tracking and event data. Positional data, describing player and ball positions at any point in time of a match, are collected automatically via computer vision algorithms and dedicated tracking cameras. Event data, on the other hand, provides basic annotations of game events (mainly on ball actions like passes, shots, tackles, etc.) and is still acquired manually by human operators. The manual collection of such events is unsurprisingly labor and cost intensive and involves up to five operators per game. The goal of this article is to bridge the gap from the status quo toward fully-automatic annotations of soccer games.

There are several recent studies aiming to detect basic events directly out of video footage (Ekin et al., 2003; Wickramaratna et al., 2005; Kolekar and Palaniappan, 2009) or positional data (Zheng and Kudenko, 2010; Motoi et al., 2012; Richly et al., 2016; Stein et al., 2019) and others focus on the identification of sophisticated tactical patterns (Hobbs et al., 2018; Andrienko et al., 2019; Shaw and Sudarshan, 2020; Anzer et al., 2021; Bauer and Anzer, 2021). The proposed approaches provide useful solutions for their respective tasks. However, they are also restricted to either a particular data source or type of events or pattern that is to be detected; none of the above approaches offer an all-encompassing framework to deal with general detection problems.

A challenge for designing a general detector of game situations is the available data structure. While vast amounts of positional data of players and ball exist, collecting the associated labels of interest is an expensive endeavor and requires manual annotation by human experts. For example, counterattack detection first involves defining strict criteria and definitions of counterattacks before engaging in extensive search processes to annotate the matching game snippets. Consequently, it is vital to reliably extract the game situations with little external supervision. In that sense, classical supervised learning methods fail to be a viable candidate since the algorithms typically require large amounts of annotated data to achieve a good generalization error (Erhan et al., 2010). However, a strategy to mitigate the necessity of a large number of labels is to incorporate abundantly available *unlabeled data* into the training process. While there are many conceivable ways to operate within such a semi-supervised framework, we focus particularly on the variational autoencoder (VAE) (Kingma and Welling, 2013; Rezende et al., 2014) family of methods.

Variational autoencoders learn implicit low-dimensional feature representations for input data by jointly training a probabilistic encoder and decoder network. The idea is that the original observations can be reconstructed (approximately) from this lower-dimensional feature space. In fact, our semi-supervised strategy relies on inferring these semantically salient representations for annotated situations, hence reducing the need to solve a large supervised learning problem in feature space. Our instance of semi-supervised learning achieves a substantial increase in generalization ability in cases where only a few observed labels are available (Kingma et al., 2014). An essential contribution of this paper is to lift the underlying principles to spatiotemporal structures to capture the temporal and spatial dependencies of positional data. Existing body of research on extending VAEs to sequential data mainly focuses on the generative aspects of the models rather than on their potential benefits in the context of semi-supervised learning (Chung et al., 2015; Goyal et al., 2017).

In this paper, we propose novel VAE-based feature extraction methods. Starting from the vanilla VAE, we begin with proposing a rather straight forward generalization that can be applied to positional data. A second contribution incorporates existing auxiliary labels in the training process. The idea of the auxiliary labels is to foster discriminative causes of variation in the inferred latent feature representation. The main contribution however is the development of sequential counterparts of the two VAEs to match the spatiotemporal problem domain. After one of the VAEs has been trained using unlabeled or auxiliary labeled data, only a few of the feature representations, for which labels of interest exist, are fed into a support vector machine to train the final classifier. We empirically evaluate the effectiveness of our approach on three different detection tasks, involving the detection of cornerkicks, crosses (labels obtained from event data), and counterattacks (labels manually annotated by experts). We observe detection rates above 90% AUC for all tasks and discuss several findings on methodological issues derived from further experimentation.

The remainder is structured as follows. Section Problem Setting introduces the formal problem setting. The static and sequential models are presented in sections Static Models, Sequential Models, respectively. We report on our empirical findings in section Empirical Evaluation and provide a discussion in section Discussion. Section Related Work reviews related work and section Conclusion concludes.

PROBLEM SETTING

Positional data from professional soccer is introduced as follows. Let $\mathcal{A}$ be the set of agents (i.e., players and ball) and $\mathcal{T}$ be the set of timesteps. For each element of the cartesian product $\mathcal{A} \times \mathcal{T}$, whereabouts of all agents on the pitch in form of two-dimensional coordinates $(g, h) \in \mathbb{R}^2$ are observed. It will be convenient to further divide the set of agents into three disjoint subsets, $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$, corresponding to the players on teams 1, team 2, and the ball[1], respectively.

Individual spatiotemporal movements of the agents allow to augment the positional data with additional pieces of information such as the (approximated) velocity of players ($\frac{dg}{dt}, \frac{dh}{dt}$). More precisely, linearized motion for agent $a \in \mathcal{A}$ is computed via

$$(\Delta g_t^{(a)}, \Delta h_t^{(a)}) = (g_{t'}^{(a)} - g_t^{(a)}, h_{t'}^{(a)} - h_t^{(a)})$$

with $t' > t$ and $(\Delta g_t^{(a)}, \Delta h_t^{(a)}) = (0, 0)$ for the case of $t' \notin \mathcal{T}$, i.e., using a small time window between two consecutive frames. Further defining $\mathcal{Y}$ as an *auxiliary label space* that consists of inexpensive labels (e.g., provided by event data), we are given a subset of event annotations $\mathcal{T}_{\mathcal{Y}} \subset \mathcal{T}$ s.t. $|\mathcal{T}_{\mathcal{Y}}| \ll |\mathcal{T}|$, referred

[1]We have $\mathcal{A}_i \subset \mathcal{A}$ s.t. $\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 = \mathcal{A}$ and $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 = \emptyset$.

to as $y_S := \{y_t : t \in \mathcal{T}_\mathcal{Y}\}$. We further denote $\mathcal{Y}_b$ as the (binary) *target space* described by an action value of interest and a "no action" value with $\mathcal{T}_{\mathcal{Y}_b} \subset \mathcal{T}$ ($|\mathcal{T}_{\mathcal{Y}_b}| \ll |\mathcal{T}|$) defining the set $y_B := \{y_t : t \in \mathcal{T}_{\mathcal{Y}_b}\}$[2]. We denote the composite of all pixel coordinates and velocity values of agents a at a certain timestep as $\boldsymbol{x}_t \coloneqq \{(g_t^{(a)}, h_t^{(a)}, \Delta g_t^{(a)}, \Delta h_t^{(a)})\}_{a \in \mathcal{A}}$ and formulate our objective as quantifying the probability over $\mathcal{Y}_b$ given the state representation $\boldsymbol{x}_t$ for all $t \in \mathcal{T}$.

An emerging issue is to find a pertinent representation of the described data for model training. A plain random concatenation of the agents' coordinates and velocities at time t is clearly inappropriate in the sense that divergent instantiations of agent orderings also translate into divergent representations for the exact same state. Accordingly, the function that tranforms instances of $\{(g_t^{(a)}, h_t^{(a)}, \Delta g_t^{(a)}, \Delta h_t^{(a)})\}_{a \in \mathcal{A}}$ into an input representation of a neural network needs to be invariant under permutation of the agents. Since the locations of the agents are given as pixel coordinates, we choose to convert these coordinates into an image-based representation, resulting in a consistent representational structure across different game settings.

The mechanism for capturing position and motion information in a 3-dimensional image representation $\boldsymbol{x}_t$ is based on the approach presented in Dick and Brefeld (2019). Here, the pitch size (105 × 68) defines the axes in the horizontal and vertical directions, with each channel of the tensor encoding a different subset of the available information. The first 3 channels capture positional information of $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ (in that very order) by assigning constant 1 s to the coordinates defined by $(g_t^{(a)}, h_t^{(a)})\ \forall a \in \mathcal{A}$ and the corresponding channel. Since agent positions live in real-world coordinates, a transfer into image pixels requires a translation $(g_t^{(a)}, h_t^{(a)}) + t$ with $t = (\frac{105}{2}, \frac{68}{2})$, effectively shifting the origin from the center of the image to the top left corner. The remaining channels track motion information, with velocity values acting as value assignments for the indices instead of constant 1 s. The speed values in g direction (Δg_t) is covered for $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$ in channels 4, 6 and 8; the information in h direction (Δh_t) is handled by channels 5, 7 and 9. All other values in the resulting input representation $\boldsymbol{x}_t \in \mathbb{R}^{105\times 68\times 9}$ are 0.

In summary, the final dataset representing a soccer game is a collection of tensor representations for each timestep $\mathcal{D} = \{\boldsymbol{x}_1, .., \boldsymbol{x}_{|\mathcal{T}|}\}$ with additional label sets y_S (auxiliary labels) and y_B (target labels). The goal is to use the available evidence and auxiliary labels to construct detectors that work effectively to identify situations of interest defined in $\mathcal{Y}_b$. To this end, we adopt a two-stage optimization procedure, which relies on the derivation of semantically meaningful feature representations. This instance of semi-supervised learning is advantageous in the present context because a large part of the model training is already accomplished independently of the specific game situation of interest. Consequently, the general detection design can be described based on the following stages:

[2] A description of the exact form and type of the label information used in this work is given in section Experimental Setup.

1. The training of a VAE-based feature extraction module to transform the high-dimensional tensor data $\boldsymbol{x}_t$ into a low-dimensional embedding space.
2. The training of a classifier using the derived embeddings and the available label information.

Irrespective of the first step's choice, we use a support-vector machine (SVM) (Cortes and Vapnik, 1995) for the second step. The technical contributions of this paper address the first stage and introduce novel feature extraction methods in sections Static Models and Sequential Models. See **Figure 1** for an illustration of the information flow.

STATIC MODELS

In this section, we present static models that operate only on a single timestamp to predict a labeling of the encoded situation. The term static stems from an equivalence class of model architectures whose resulting optimization targets are derived based on the assumption that each tensor frame is iid., i.e., the computation factors across the individual timesteps of a game. Note, however, that the data points themselves contain sequential information due to the inclusion of motion vectors for each agent. We discard the time subscripts for the tensor representations $\boldsymbol{x}$ since we operate within a static domain.

Preliminaries

The idea of a variational autoencoder (VAE) (Kingma and Welling, 2013; Rezende et al., 2014) is to learn a deep generative model $p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p_\theta(\boldsymbol{x}|\boldsymbol{z})$ by maximizing the marginal log-likelihood of the training data $\mathcal{D}$. Due to intractabilities that arise from the integration over the latent variables $\boldsymbol{z}$, the marginal likelihood is substituted by some variational lower bound to infer the model parameters. This requires introducing a variational approximation $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, which is used to approximate the intractable true posterior. The resulting (negative) evidence lower bound (ELBO) denotes the VAE training criterion and enables concurrent optimization of θ and ϕ,

$$\log p_\theta(\boldsymbol{x}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathcal{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})] \equiv -\mathcal{L}_{\text{VAE}}(\theta, \phi; \boldsymbol{x}). \quad (1)$$

The first term of (1) quantifies the reconstruction error and the second term measures the distance between variational approximation and the pre-defined prior in terms of the KL divergence. The learned variational distributions $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ capture semantically meaningful low-dimensional feature representations of the higher-dimensional observations $\boldsymbol{x}$. This encoded information facilitates finding a generalizable discriminator, especially when labels are scarce. The merits of such a semi-supervised instance are e.g., explored in the M1 model in Kingma et al. (2014), where samples from the approximate posterior distribution over the latent variables $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ are used as input data for a downstream classifier (e.g., an SVM) to learn a decision boundary in latent space.

SoccerVAE

We begin with a rather straight forward application of VAEs to the problem at-hand. The SoccerVAE uses the same optimization

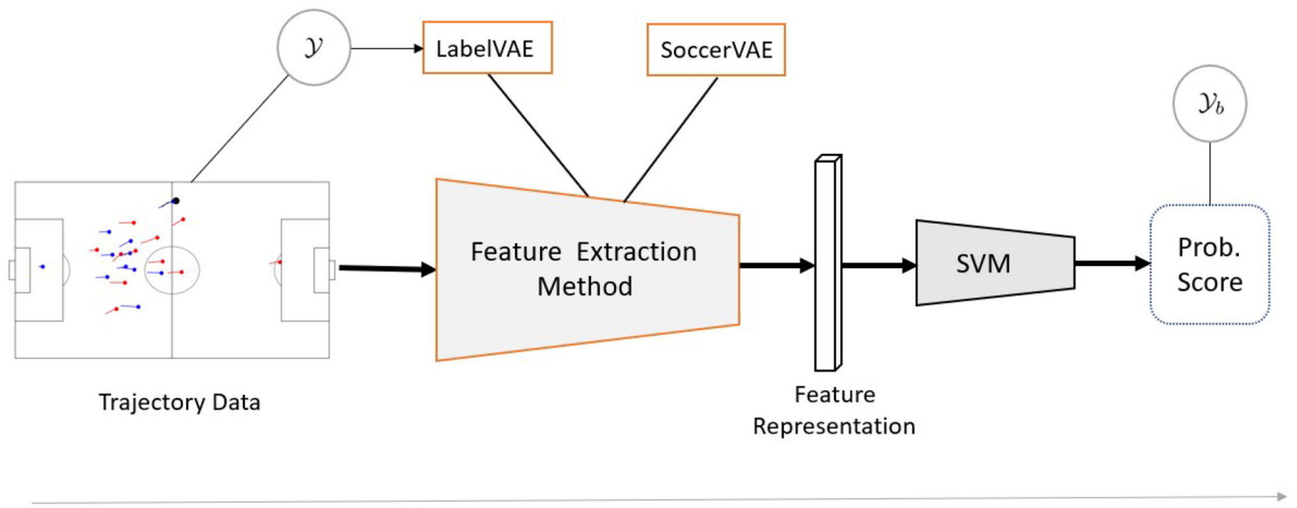

FIGURE 1 | Detection worflow. The building blocks outlined in orange are presented in sections Static Models, Sequential Models and denote the technical contributions of this paper. The proposed LabelVAE methods for feature extraction operate on $\mathcal{Y}$, while the SoccerVAE methods are trained fully unsupervised.

target as the vanilla VAE (cf. Equation 1) so that only the input and resulting choices on distribution type and architecture design need to be considered[3]. Regarding the former, the generating distribution of the generative model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$ is modeled as a multivariate distribution of independent Bernoulli parametrized by a decoder neural net with parameters θ:

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \text{Bernoulli}(\boldsymbol{x}|\boldsymbol{\mu}(\boldsymbol{z};\theta)) = \prod_{j=1}^{D} \text{Bernoulli}(x_j|\mu_j(\boldsymbol{z};\theta)),$$

where D is the dimensionality of $\boldsymbol{x}$ and $\boldsymbol{\mu}(\mu_1, \ldots, \mu_D)^\top$ aggregates the individual $\mu_j \in [0, 1]$ parameters for each pixel. This consitutes a reasonable design choice as we constrain the observed values to lie in the interval $[0, 1]$.

Our generative and inference network definitions can be seen as instantiations of the class of CNN proposed by Radford et al. (2015). Specifically, the network $\boldsymbol{\mu}(\boldsymbol{z};\theta)$, which incrementally converts a sampled vector $\boldsymbol{z}$ to the observation space $\boldsymbol{x} \in \mathbb{R}^{105\times68\times9}$, is implemented using fractional-strided convolutions with ReLU activations (Nair and Hinton, 2010) and a sigmoid activation for the output layer, as well as batch normalization layers to reparametrize the intermediate layer activations (Ioffe and Szegedy, 2015; Bjorck et al., 2018). Each of the convolutional layers has kernels of the same size, with the number of kernels per layer decreasing proportionally to the depth of the network. All four proposed models deal with continuous priors given in form of standard multivariate Gaussians. The inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is a diagonal Gaussian parametrized by an encoder neural net with parameters ϕ,

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\mu(\boldsymbol{x};\phi), \text{diag}(\sigma^2(\boldsymbol{x};\phi))).$$

[3]Unless explicitly stated, these choices are reused in the derivation of the other models.

The role of the encoder is to transform a static game situations into fixed-size vector representations. We use strided convolutions with the leaky rectified activation (Maas et al., 2013; Xu et al., 2015) and batch normalization to process the input tensors. A fully-connected layer is dedicated to mapping the final representation onto the parameter space of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, i.e., to the mean and standard deviation vector of a diagonal Gaussian, which are used in conjunction with $\mathcal{N}(\boldsymbol{\epsilon}|0, I)$ to generate the latent vector $\boldsymbol{z}$.

LabelVAE

The goal is to infer continuous latent embeddings that capture beneficial properties to detect a predefined (generally speaking: rarely occurring) game situation of interest. Hence, the quality of our approach is not primarily measured by reconstruction errors but in terms of the ability to discriminate between different types of situations in the subsequent supervised learning task. The second static model thus aims at directly optimizing a classification network. The model uses a VAE over the input variables that serves an effective regularizer. However, our envisaged optimization strategy is based on the extraction of general feature representations via pre-trained parameters to enable flexible adaption to the task at-hand.

The generative model reflects that causal factors of the observed $\boldsymbol{x}$ can be broadly categorized into label-specific and label-unspecific factors,

$$p_\theta(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{z})p(\boldsymbol{z})p(\boldsymbol{a}), \qquad (2)$$

where we assume that $\boldsymbol{a}$ encapsulates all relevant label-specific information and $\boldsymbol{z}$ the remaining label-unspecific characteristics. The dependency structure of the inference model embodies the consideration that the data-specific latent information $\boldsymbol{z}$ may vary with respect to the class-specific information of $\boldsymbol{a}$, that is,

$$q_\phi(\boldsymbol{a}, \boldsymbol{z}|\boldsymbol{x}) = q_\phi(\boldsymbol{a}|\boldsymbol{x})q_\phi(\boldsymbol{z}|\boldsymbol{a}, \boldsymbol{x}). \qquad (3)$$

The above approximate posterior is amenable to approximating the true posterior over the latent variables to provide a tractable lower bound on the log-likelihood $\log p_\theta(\boldsymbol{x})$. The resulting (negative) ELBO is the optimization target of an unsupervised data point

$$\log p_\theta(\boldsymbol{x}) = \log \int\int p_\theta(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{z}) dz d\boldsymbol{a} \tag{4}$$

$$= \mathbb{E}_{q_\phi(\boldsymbol{a}|\boldsymbol{x})}\Bigg[\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{a})}[-\log p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{a})] - \mathcal{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{a}) \parallel p(\boldsymbol{z})]\Bigg] - \mathcal{KL}[q_\phi(\boldsymbol{a}|\boldsymbol{x}) \parallel p(\boldsymbol{a})] \tag{5}$$

$$\equiv -\mathcal{L}_u(\theta, \phi; \boldsymbol{x})$$

To encourage the model to capture the most relevant variational factors in the representations obtained via inference, we embed the available supervised learning signals concurrently with the unsupervised learning signals by means of an auxiliary classifier. Thus, the learning process is given by jointly maximizing the probability of each frame $\log p_\theta(\boldsymbol{x})$ and minimizing the auxiliary loss given the latent space realizations $\boldsymbol{a}$,

$$\mathcal{L}_s(\theta, \phi, \xi; \boldsymbol{x}, y) = \mathcal{L}_u(\theta, \phi; \boldsymbol{x}) - \alpha \mathbb{E}_{q_\phi(\boldsymbol{a}|\boldsymbol{x})}[\log q_\xi(y|\boldsymbol{a})], \tag{6}$$

where ξ are the parameters of the classifier, α is a hyperparameter encoding the trade-off between generative and discrimative learning and $q_\xi(y|\boldsymbol{a}) = \text{Cat}(y|\boldsymbol{\pi}(\boldsymbol{a}; \xi))$. Equation (6) is essentially a regularized classification objective. More precisely, the second term quantifies the performance of a deep classification network with injected noise from the sampling operation $\boldsymbol{a} \sim q_\phi(\boldsymbol{a}|\boldsymbol{x})$ and the variational loss $\mathcal{L}_u$ can be viewed as a form of regularization imposed on the learned representations of the supervised prediction model.

The full training criterion is then given by collecting $\mathcal{L}_s$ and $\mathcal{L}_u$ for the supervised and unsupervised data points of the evidence $\mathcal{D}$:

$$\mathcal{L}_{LabelVAE}(\theta, \phi, \xi; \mathcal{D}_u, \mathcal{D}_s) = \sum_{(\boldsymbol{x},y)\sim\mathcal{D}_s} \mathcal{L}_s(\theta, \phi, \xi; \boldsymbol{x}, y) + \gamma \sum_{\boldsymbol{x}\sim\mathcal{D}_u} \mathcal{L}_u(\theta, \phi; \boldsymbol{x}), \tag{7}$$

where $\mathcal{D}_s := \{(\boldsymbol{x}_t, y_t), \forall t \in \mathcal{T}y\}$ and $\mathcal{D}_u := \mathcal{D} \setminus \mathcal{D}_s$, and trade-off γ balances the contribution of the unsupervised term to the overall objective. This can be advantageous in situations where the labeled data is very sparse ($N_l \ll N_u$) and therefore aim to externally impinge on the relative weight that is otherwise implicitly given by the data set (Siddharth et al., 2017). We define the feature vector for SVM training by concatenating the derived variables $\boldsymbol{a}$ and $\boldsymbol{z}$ into a single vector: $[\boldsymbol{a}, \boldsymbol{z}]$.

SEQUENTIAL MODELS

A clear limitation of the static models of the previous section is that their input is solely a single snapshot of the game. Although direction of movement and velocities may add context to the otherwise isolated situation, the idea of processing short sequences around these situations may add important information. Hence, in this section, we present sequential variants of the previously introduced models.

We denote a slice of consecutive frames from the game $\mathcal{D}$ as $\boldsymbol{x}_{\leq T}$, where T denotes the length of the game segment. Importantly, this implies that the time specifications of the frames $\boldsymbol{x}_t$ refer more narrowly to the timestep in a segment within the soccer game $\boldsymbol{x}_{\leq T} = \boldsymbol{x}_1, ..., \boldsymbol{x}_T$ and no longer to the timestep in the overall game (as we describe it in section Problem Setting).

SeqSoccerVAE

A viable avenue for inferring sequence-level features is to reconstruct the input sequence using a single global latent variable $\boldsymbol{z}$. While most approaches from the literature have been developed for modeling data distributions, we revisit this approach primarily to aggregate game sequences/multi-agent trajectories into informative vectors. Here we simply adapt the static VAE objective (1) to a sequential definition by assigning a temporal dimension to the data points:

$$\mathcal{L}_{\text{SeqSoccerVAE}}(\theta, \phi; \boldsymbol{x}_{\leq T}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}_{\leq T})}[\log p_\theta(\boldsymbol{x}_{\leq T}|\boldsymbol{z}) - \mathcal{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{x}_{\leq T}) \parallel p(\boldsymbol{z})]. \tag{8}$$

To model the components constituting Equation (8), we generalize the parameter functions for a given point to architectures suitable for sequential data. Accordingly, the parameters of the approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x}_{\leq T})$ are obtained from the last hidden state of an encoder RNN (parameterized by ϕ) working on the input sequence, and the generating distribution $p_\theta(\boldsymbol{x}_{\leq T}|\boldsymbol{z})$ is modeled by a decoder RNN (parameterized by θ) conditioned on the sampled hidden code alongside the previous data point, yielding the generating distribution $p_\theta(\boldsymbol{x}_{\leq T}|\boldsymbol{z}) = \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_t|\boldsymbol{z}, \boldsymbol{x}_{<t})$. Thus, we force the model to encode all information about the data into the latent variable since it is the only source of information available for data reconstruction. The overall workflow of the SeqSoccerVAE is illustrated in **Figure 2**.

SeqLabelVAE

The static LabelVAE in section LabelVAE seeks to leverage discriminative information already existing in the data by injecting them into the latent space via a classification network to facilitate the detection of game situations. In this section, we propose a sequential generalization of the LabelVAE that builds upon the dependencies in inference and generative parts of its peer. Accordingly, the SeqLabelVAE utilizes a label-specific partition of the latent space into $\boldsymbol{a}_t$ and $\boldsymbol{z}_t$, describing two distinct pieces of information about the data. We address the temporal dependency for successive observations by generating conditional independence for the random variables (the data and the latent variables) given the hidden states of two separate RNN networks,

$$\boldsymbol{h}_t^{enc} = f_\phi(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}^{enc})$$
$$\boldsymbol{h}_t^{dec} = g_\theta(\boldsymbol{a}_t, \boldsymbol{z}_t, \boldsymbol{h}_{t-1}^{dec}),$$

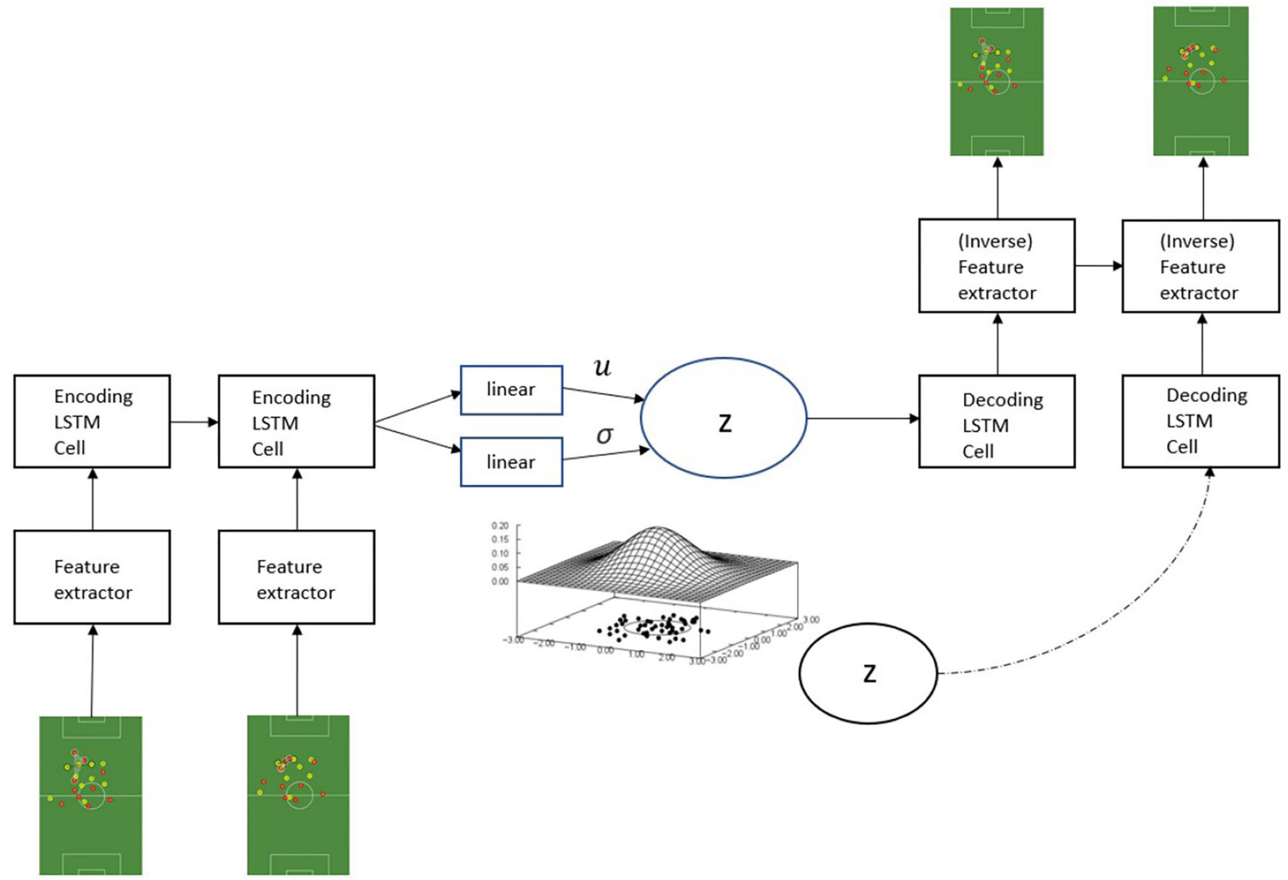

FIGURE 2 | The SeqSoccerVAE. The model takes a sequence of frames as input and then extracts corresponding feature representations per timestep, which are passed to an LSTM network that outputs a global summary of the sequence. The sampled latent representation is used to reconstruct the sequence.

where $\boldsymbol{h}_t^{enc}$ denotes the recurrent state for the inference model and $\boldsymbol{h}_t^{dec}$ denotes the recurrent state for the generative model.

The latent variables of the generative model at time t encode the observation $\boldsymbol{x}_t$ indirectly via the state representation $\boldsymbol{h}_t^{dec}$, yielding the conditional distribution $p_\theta(\boldsymbol{x}_t|\boldsymbol{z}_{\leq t}, \boldsymbol{a}_{\leq t})$. As in the previous models, we restrict ourselves to standard multivariate Gaussian priors for both latent variables per timestep. Using unconditional prior distributions may reduce the approximability of observation sequences, but our focus is on obtaining informative feature representations rather than on generating sequences. For the inference model, we condition the LabelVAE dependency structure of the posterior approximation on the RNN state $\boldsymbol{h}_t^{enc}$, resulting in the factorization

$$q_\phi(\boldsymbol{z}_{\leq T}, \boldsymbol{a}_{\leq T}|\boldsymbol{x}_{\leq T}) = \prod_{t=1}^{T} q_\phi(\boldsymbol{z}_t|\boldsymbol{a}_t, \boldsymbol{x}_{\leq t}) q_\phi(\boldsymbol{a}_t|\boldsymbol{x}_{\leq t}).$$

The derivations in the remainder of this section is analogous to the derivation of the static LabelVAE objective. Specifically, we optimize an unsupervised training instance by maximizing the ELBO

$$\mathcal{J}_u(\theta, \phi; \boldsymbol{x}_{\leq T}) = \mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq T}, \boldsymbol{a}_{\leq T}|\boldsymbol{x}_{\leq T})}\Bigg[\sum_{t=1}^{T} - \log p_\theta(\boldsymbol{x}_t|\boldsymbol{z}_{\leq t}, \boldsymbol{a}_{\leq t}) + \mathcal{KL}[q_\phi(\boldsymbol{z}_t|\boldsymbol{x}_{\leq t}, \boldsymbol{a}_t) \parallel p(\boldsymbol{z}_t)] + \mathcal{KL}[q_\phi(\boldsymbol{a}_t|\boldsymbol{x}_{\leq t}) \parallel p(\boldsymbol{a}_t)]\Bigg].$$

Also, we enforce the latent variables to encode discriminative information by introducing an auxiliary classifier for the supervised training loss

$$\mathcal{J}_s(\theta, \phi; \boldsymbol{x}_{\leq T}, y) = \mathcal{J}_u(\theta, \phi; \boldsymbol{x}_{\leq T}) - \alpha \mathbb{E}_{q_\phi(\boldsymbol{a}_{\leq T}|\boldsymbol{x}_{\leq T})}\Big[\sum_{t=1}^{T} \log q_\xi(y_t|\boldsymbol{a}_{\leq t})\Big],$$

where $\log q_\xi(y_t|\boldsymbol{a}_{\leq t})$ is the per timestep classification loss and α is the hyperparameter that controls the trade-off between classification and generation. Note that the label $y \in \mathcal{Y}$ denotes the event annotation for the game situation $\boldsymbol{x}_{\leq T}$, such that each frame is assigned an identical label: $y_1 = ... = y_T = y$.

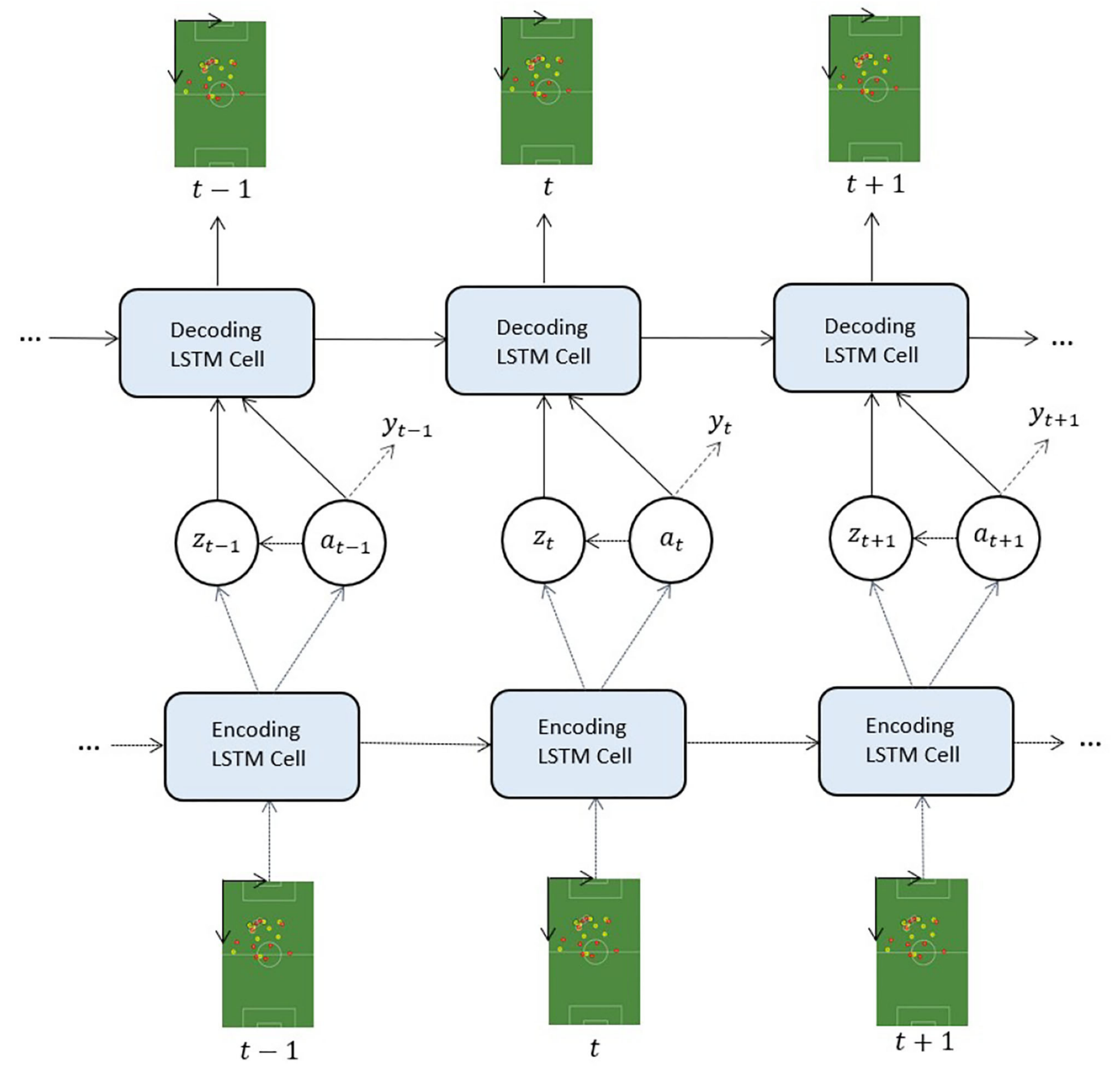

FIGURE 3 | The SeqLabelVAE. Dashed lines indicate modules performing inference, solid lines denote components performing generation, and dotted lines point out the auxiliary classifier.

We define the feature vector for classifier training on $\mathcal{Y}_b$ by concatenating the derived variables $\boldsymbol{a}_{\leq T}$ and $\boldsymbol{z}_{\leq T}$ into a single vector: $[\boldsymbol{a}_1^{\mathrm{T}}, ..., \boldsymbol{a}_T^{\mathrm{T}}, \boldsymbol{z}_1^{\mathrm{T}}, ..., \boldsymbol{z}_T^{\mathrm{T}}]$. The SeqLabelVAE architecture is sketched in **Figure 3**.

EMPIRICAL EVALUATION

Data

We operate on two matches of the German national team. The tracking data consist of (g, h) positions of all players and ball, sampled at 25 frames per second. Following Dick et al. (2018), the tensor representations of the games are computed as follows. Firstly, the origin centered representation of the player position is transformed into pixel values of the tensor representation. This is done by adding half of the size of the pitch along the horizontal and vertical direction to the position of the agents. To approximate the velocities of the players and the ball at each timestep, we compute differences in positions over the last five frames (corresponding to a time lag of 0.2 s), yielding movement vectors of the form $(\Delta g_t, \Delta h_t) = (g_{t+5} - g_t, h_{t+5} - h_t)$. Since we assume the outputs to be Bernoulli distributed, we map the resulting speed values onto the range $[0, 1]$. To obtain the final input representation described in section Problem Setting, we incorporate the coordinates and velocity values into a 0-tensor of the size of the target shape $(105, 68, 9)$. The updated tensor forms the input for a single timestamp. Every game consists of about $140,000$ such frame representations.

Experimental Setup

As described in section Problem Setting, we define our setup with two different label spaces: the auxiliary label space $\mathcal{Y}$ that includes all available (inexpensive) labels and the binary (expensive) label space $\mathcal{Y}_b$ that indicates occurrences of the game situation of interest. The auxiliary label space $\mathcal{Y}$ defines the label information y_S and originates in our study from the event data of the respective games (roughly 4000 observations per game). Note that only LabelVAE and SeqLabelVAE make use of these inexpensive labels in the training process to capture discriminative variations in the respective feature spaces. For simplicity, we focus on 5 auxiliary lables, $\mathcal{Y} = \{shot, cross^4, ground, pass, other\}$. If more than one auxiliary label is active in a snapshot, we select the minority label for the observations in question.

By contrast, the label space $\mathcal{Y}_b$ defines the label information y_B used for SVM training and depends on the task at-hand. Our exemplary use cases target game actions of increasing difficulties by predicting variables encoded already in the available auxiliary labels or annotated manually by human experts. Accordingly, when employing fully unsupervised feature extraction methods (i.e., SoccerVAE and SeqSoccerVAE), targets y_B are the only label information required. We elaborate on the exact construction of the set y_B when discussing the predictive results in the following section.

We use one game for training and model selection and the other game for testing. In the training process, parameters of the static and sequential VAEs are optimized as well as parameters of the support vector machine which serves as the final classifier. After training, the best parameters are fixed and used for processing the test game. For every frame in the test game, probabilities of the quantities of interest are computed as follows: A static (section Static Models) or sequential (section Sequential Models) approach computes the embedding of the situation which is then used as input to the support vector machine which computes the prediction of interest and a softmax turns this prediction into a probability.

To assess the detection performance, we mainly use two different performance metrics: the area under the ROC curve (AUC) and the F1 score. We calculate the relevant components that constitute the F1 score (true positives (TPs), false positives (FPs), and false negatives (FNs)) as follows. To identify an action, we apply a threshold to the derived probability estimates for each frame of the test game. The independently detected frames are then converted into coherent game situations (or positive prediction instances), defined as a set of detected consecutive frames where the time gap between 2 successive frames is less than 10 s. The average length of the detected sequences depends on the concrete application, but it is in the range of a few seconds in most cases. We obtain TP values (FP values) if any (no) element within the extracted sequences is assigned the label of interest. Further, we define FN values as true action frames that remain undetected, i.e., do not occur within the positively predicted regions. We compute F1-scores for 50 distinct threshold values in the range between 0.6 and 0.98 and only report the maximum F1-score in the subsequent section[5].

We compare our approaches to a fully supervised deep convolutional network that directly processes the tensor frames. The architecture of the baseline is identical to the feature extraction modules of our inference models, i.e., it consists of convolutional and batch normalization layers with LeakyReLu activation functions. The output dimensionality equals 1, and we use the standard binary cross-entropy loss for training. That is, the baseline directly computes the prediction of the desired label without a need for an additional SVM but lacks the reconstruction part of the proposed networks. We train the model with RMSprop (Tieleman and Hinton, 2012) and a batch size of 4. All methods are implemented with Tensoflow 2.0 (Abadi et al., 2016)[6].

To ensure clarity regarding the used baseline architecture, we replaced "feature extraction modules of our inference models" with "encoder network of the SoccerVAE." We report the comparison with this supervised baseline in **Table 1**.

Predictive Accuracies

We showcase the expressivity of our approaches on three tasks with gradually increasing difficulty, the first one being the automatic **detection of cornerkicks**. The task should be the easiest one as the spatial distribution of agents is very indicative and event data provides ground-truth labels. The second task is the **detection of crosses**. Again, ground-truth is provided by event data, however, the spatial distribution of the agents is not as obvious as for cornerkicks. For both tasks, we train the models on one game and use another one for testing and evaluation.

The third task is the **detection of counterattacks** and clearly more involving than the former two. The task is more difficult than the previous two as many different temporal aspects need to be learned by the model, including gaining and maintaining ball possession, etc. Labels for this task are provided by human experts. Since the effort of labling is tedious, we train the models only on the first half of a game and evaluate on the second.

We begin with the detection of cornerkicks. For this straight forward task, the variational autoencoders are trained on a single game. The subsequent SVM is trained on 16 labeled examples per class (cornerkick vs. no cornerkick), where the negative examples are randomly drawn from the training game. The test game contains 26 cornerkick situations. The baseline uses the same training and testing set as the downstream SVM. **Table 1** (top rows) summarizes the results for the different metrics on the test/validation game. All semi-supervised approaches outperform the fully-supervised baseline with SeqLabelVAE being the best predictor in this task. Comparing the static

[4]The auxiliary label cross also includes corners and freekicks.

[5]Therefore, unlike the reported AUC values, the F1 scores are validation values as we engage in threshold optimization.

[6]The source code is available at https://github.com/fassmeyer/labeling-situations.

TABLE 1 | Results for the detection of cornerkicks, crosses and counterattacks.

Task	Model	AUC	TP-Rate	Precision	F1	Length
Cornerkick	Baseline	0.909	0.904	0.478	0.620	13.624
	SoccerVAE	0.944	0.940	0.578	0.716	14.445
	LabelVAE	0.967	0.877	0.670	0.760	8.451
	SeqSoccerVAE	0.975	0.886	0.792	0.824	11.054
	SeqLabelVAE	**0.986**	0.920	0.785	**0.850**	14.560
Cross	Baseline	0.827	0.765	0.507	0.606	20.070
	SoccerVAE	0.920	0.933	0.575	0.707	24.229
	LabelVAE	0.924	0.927	0.577	0.711	24.812
	SeqSoccerVAE	0.931	0.983	0.578	0.728	19.138
	SeqLabelVAE	**0.940**	0.812	0.683	**0.739**	16.750
Counterattack	SeqSoccerVAE	0.835	0.855	0.533	0.651	7.586
	SeqLabelVAE	**0.912**	0.745	0.726	**0.730**	3.712

The highest values are indicated in bold face. The average length of the detected segments is given in seconds. All numbers are averages on the test game.

models shows decent improvements of the LabelVAE over the SoccerVAE. Furthermore, the average length of the detected sequences is significantly lower for the LabelVAE. Since the average length is a good indicator concerning the width of the predicted amplitudes, the value can be interpreted as a confidence measure of the predictions. Though LabelVAE performs worse than the sequential models, the static models provide solid results in this task, presumably because the agents' distribution on the playing field is easily distinguishable from other game situations. When comparing the sequential models, we find that the SeqLabelVAE performs better than SeqSoccerVAE. This improvement however comes at the cost of detection lengths.

Next, we study the detection of crosses using the same extracted features as for cornerkick detection. The classifier is trained on 33 examples per class (cross vs. no cross), and the test game consists of 38 cross situations. **Table 1** (center rows) summarizes the results for the different metrics for the test/validation game. The trends are largely consistent with those of the corner detection task but at a lower overall level. The drop in performance stems from the variance in spatial distributions of agents that render the detection of crosses naturally more difficult than cornerkicks.

For the detection of counterattacks, static methods cannot sensibly be applied as the sequential nature and complexity of the situation (change of ball possession, maintaining ball possession thereafter, etc.) cannot be captured by focusing on only a single point in time. Consequentially, we only evaluate the sequential models using the first half of a manually annotated game with 27 counterattack situations for training and use the second half containing 33 situations for testing the classifier. The inherent complexity of counterattacks render the task much more challenging compared to the detection of cornerkicks or crosses. **Table 1** (bottom rows) shows the results. As in the previous cases, the SeqLabelVAE emerges as the model of choice. Albeit detection performances are below previous ones, the findings show the potential of the models in challenging domains with manual labels. The detection rate of counterattacks is still above 91% AUC.

Analyzing LabelVAE

To shed light on the effect of the auxiliary labels used in LabelVAE and SeqLabelVAE, we visualize the latent space of the former using t-SNE (Van der Maaten and Hinton, 2008) in **Figure 4**. Recall that the generative model of LabelVAE makes use of two latent variables $\boldsymbol{a}$ and $\boldsymbol{z}$. The former encodes label-specific information while the latter captures all label-unspecific traits. Thus, both latent variables are supposed to capture different properties which actually holds true for the trained models as can be seen in the figure. Every point in the figure corresponds to a game situation and its color indicates the attached auxiliary label. The difference of the two latent variables is clearly visible and accentuated by a clear separation into action clusters (right part of figure) for $\boldsymbol{a}$ and the absence of any class structure (left part) for $\boldsymbol{z}$. Since both variables are used to reconstruct the tensor frames, but merely variable $\boldsymbol{a}$ concurrently needs to accurately discriminate between the different actions, it stands to reason that $\boldsymbol{z}$ captures position-specific information useful for frame reconstruction.

Recall, that the empirical results for the LabelVAE in **Table 1** are based on concatenating the two latent variables $\boldsymbol{a}$ and $\boldsymbol{z}$ into a single feature vector yielding an AUC of 96.7% for cornerkicks. Passing on only a single variable to the SVM decreases the performance to 94.0% for $\boldsymbol{z}$ and 90.1% for $\boldsymbol{a}$, respectively. Hence, the two variables complement one another and focus on different aspects of the problem.

Qualitative Assessment

To shed light on the nature of the proposed methodology, we compare the structure of correctly and incorrectly predicted examples for the detection of counterattacks on the example of SeqSoccerVAE. We begin with a correctly identified counter attack in **Figure 5**. The upper part of the figure shows the detection probabilities computed from the output of the SVM. The black indicator on top of the figure at timestamp 68.330 indicates the true label by the experts. The SeqSoccerVAE classifies the indicated segment above the threshold (dashed line) between timestamps 68.265 and 68.355 as a successful

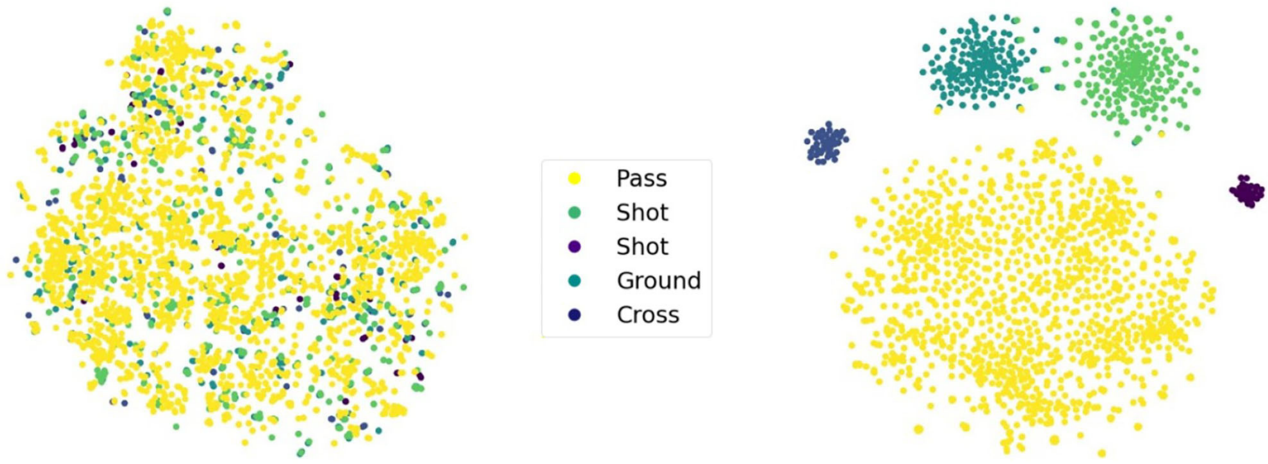

FIGURE 4 | t-SNE visualizations of variable $\boldsymbol{z}$ (left) and variable $\boldsymbol{a}$ (right) of the LabelVAE. Each point in the plot is a tensor frame and the color represents a particular game action.

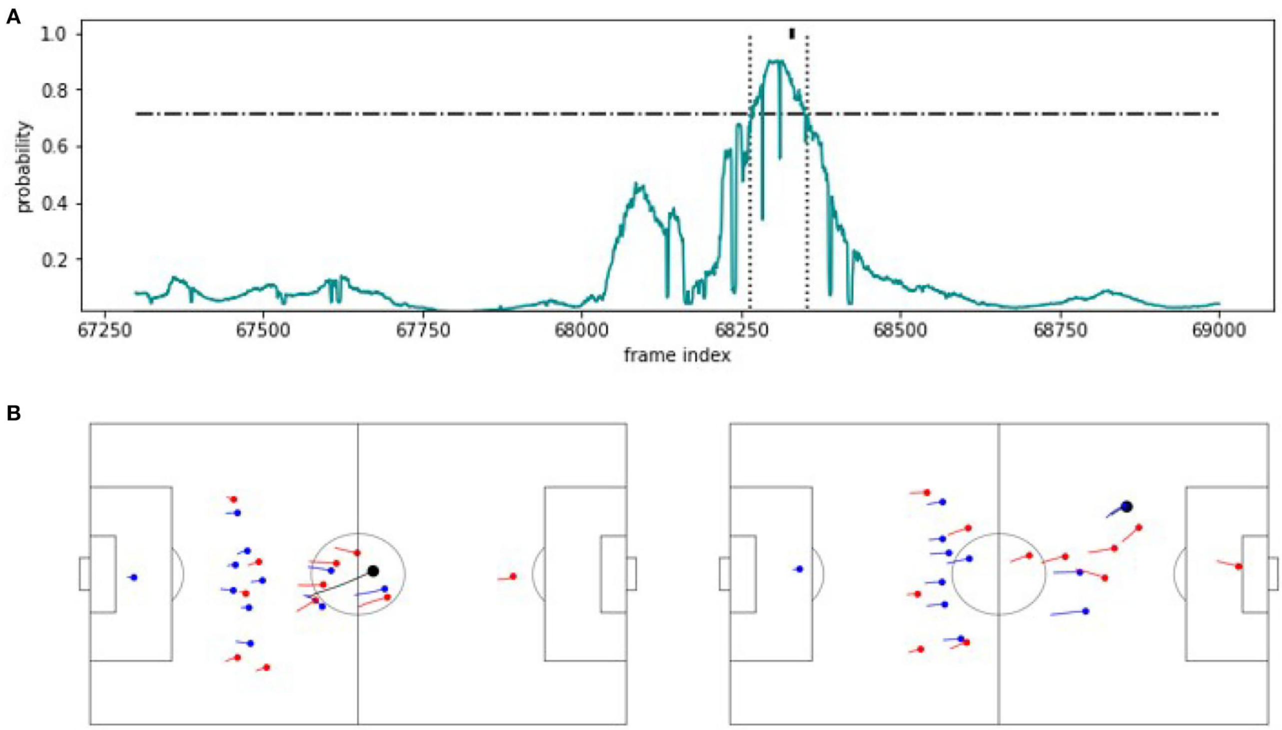

FIGURE 5 | TP example. **(A)** Predicted probability values for an approximately 70 s excerpt from the validation game. The horizontal line indicates the threshold applied on the probabilities to recognize a counterattack. The vertical lines mark the beginning and the end of the extracted sequence. The mark at the probability value 1 indicates the counterattack annotation. **(B)** The corresponding snapshot visualizations for the beginning (left) and the end (right) of the detected scene.

counterattack. The two figures below display the snapshots at the beginning and end of the detected scene and clearly show the successful counterattack that over both halves of halfes of the pitch.

By contrast, **Figure 6** shows a false negative. The detection probabilities shown in the upper part of the figure stay constantly below the threshold and consequentially, the turnover is missed by the classifier. Interestingly, the expert annotation is at a position, where the probability for a counterattack has decreased entirely and stays around zero. We credit this poor performance to the rather crowded origin of the situation and the many defending players behind the

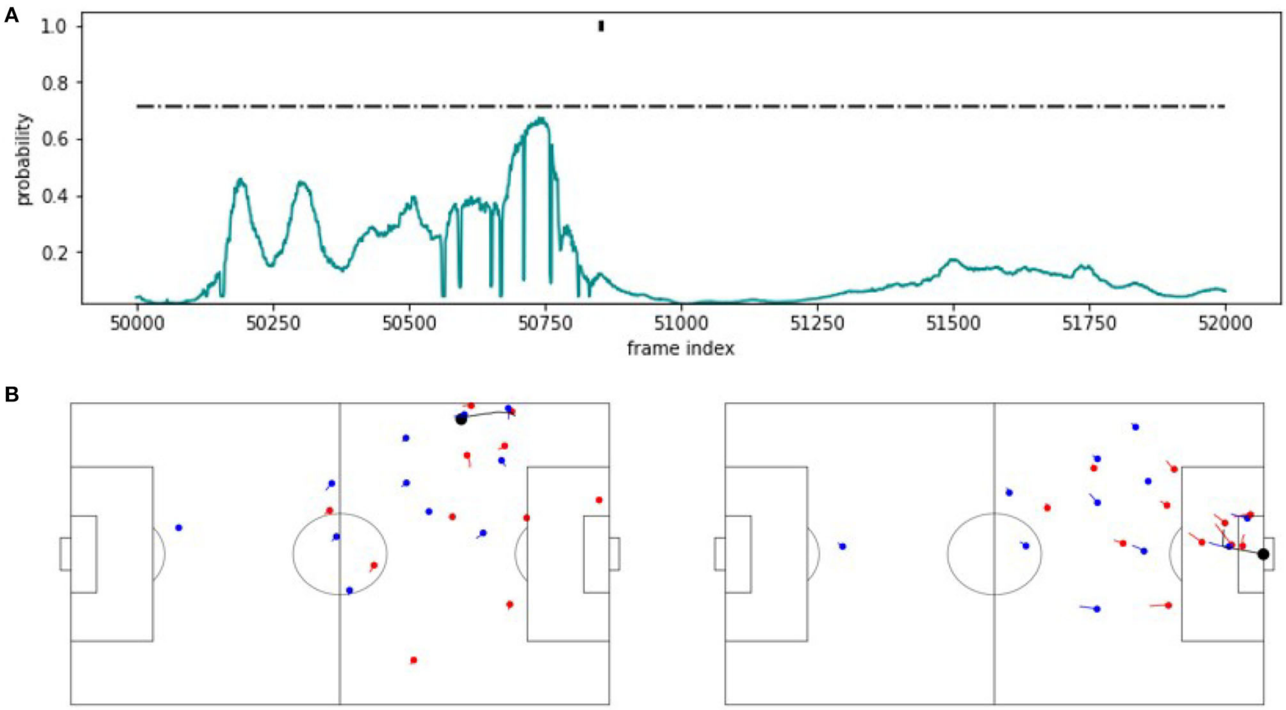

FIGURE 6 | FN example. **(A)** Predicted probability values for an approximately 80 s excerpt from the validation game. The horizontal line indicates the threshold applied on the probabilities to recognize a counterattack. The mark at the probability value 1 indicates the corner annotation. **(B)** Snapshot visualizations of a potential start and end point around the counterattack annotation.

ball. The situation is clearly different from the one shown in **Figure 5**.

Last but not least, **Figure 7** shows a false positive. As can be seen, the situation resembles the one agent in **Figure 5** but here, the turnover fails and correspondingly, there is no expert annotation. This result expresses both, the strength and the limitation of the SeqSoccerVAE, and possibly the use of VAEs in general for such tasks. By using an autoencoder, we implicitly assume that similar situations in feature space will have a similar outcome in the real world. On one hand, this assumption allows to use many unlabeled situations to extract meaningful features and render the entire classification approach with only a handful of (expert) labels feasible. On the other hand, once the feature representation is fixed, the subsequent SVM is unlikely to differentiate neighboring situations although their labels suggest separation. However, the overall performance impressively shows that the latter case does not occur very often, resulting in an excellent total detection rate.

Importance of Labeled Data

The idea of the paper grounds on splitting the original problem of labeling situations in soccer into two: an unlabeled[7] grouping of similar situation by a variational autencoder (VAE) and feeding the learned feature representation into a support vector machine (SVM) to compute the final prediction. This approach promises a much better structured feature space that allows the SVM to learn an accurate hyperplane with only a few labeled instances. This, in turn, renders the approach useful for practitioners as they only need to provide manual labels for a handful of situations.

To investigate the models' applicability in a practical context, we quantify the (human) labeling effort to achieve accurate performance for the detection of cornerkicks and counterattacks, respectively. **Figure 8** shows the results. The y-axis shows AUCs and the x-axis depicts the number of positive training examples which are (manually) labeled. In addition, the same amount of negative examples are introduced, however, these are randomly drawn from the training games and do not need manual attention. To reduce the effect of the randomness in the training sets, we report on averages over five runs; error bars indicate standard error. The left part of the figure shows the results for the SoccerVAE and the detection of cornerkicks. A training set with only six instances, three (manually) labeled positive and three randomly drawn negative ones, is sufficient to obtain optimal performance. Adding more instances to the training set does not lead to further improvements.

For the detection of counterattacks with the SeqLabelVAE (right part of figure), the performance stabilizes for about seven manually labeled data points. Increasing the size of the training set further reduces the variance that is introduced by selecting only a few positive and negative examples and renders the classifier more robust. However, the key message is

[7]Recall that we use auxiliary labels in (Seq)LabelVAE to enforce sensible groupings.

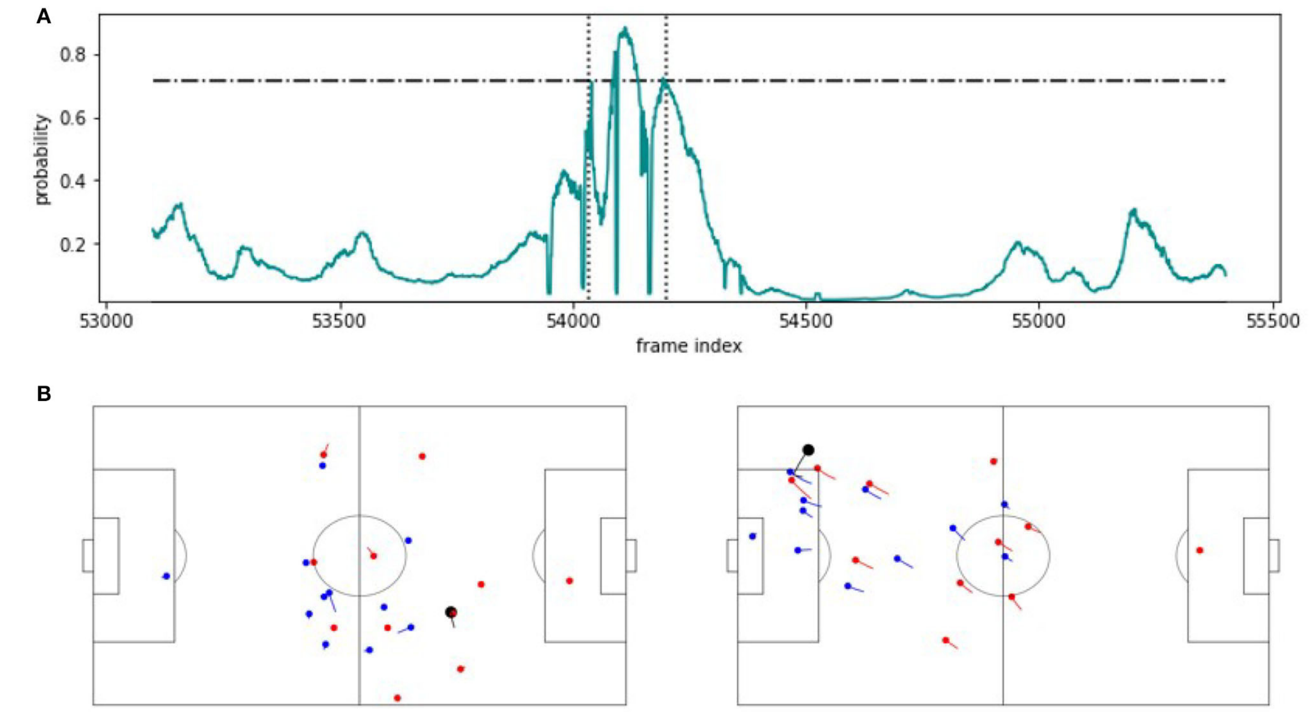

FIGURE 7 | FP example. **(A)** Predicted probability values for an approximately 100 s excerpt from the validation game. The horizontal line indicates the threshold applied on the probabilities to recognize a counterattack. The vertical lines mark the beginning and the end of the extracted sequence. **(B)** The corresponding snapshot visualizations for the beginning (left) and the end (right) of the detected scene.

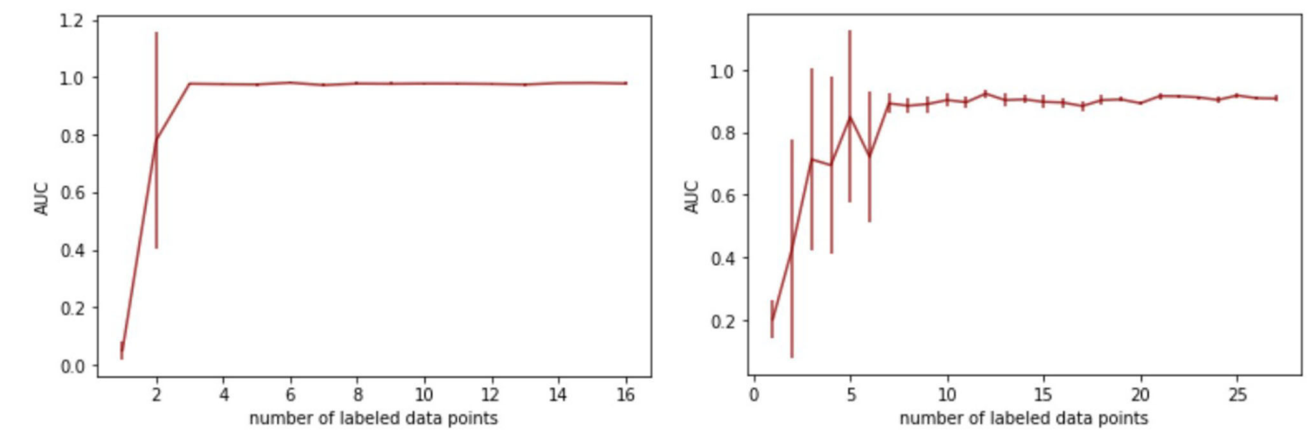

FIGURE 8 | Resulting AUC scores for different training set sizes of the SVM. Left: SoccerVAE for cornerkicks. Right: SeqLabelVAE for counterattacks.

that only seven manual annotations suffice to accurately detect counterattacks with a detection rate (AUC) of over 90%.

DISCUSSION

Our approach allows us to detect basic events (cornerkicks and passes) as well as more complicated patterns (i.e., counterattacks) without requiring massive sets of annotated data and without falling back to rule-based approaches. The detection of a more complicated pattern, namely counterattracks, is addressed in Hobbs et al. (2018) using an unsupervised clustering. By making use of a few expert-labels, we combine a data-driven approach with expert guidance. The autoencoder-based approach introduced by of Karun Singh[8] is improved in two ways: First, we use a variational autoencoder and second, we extend the

[8]Opta Analytics Pro Forum, 2019 London https://www.youtube.com/watch?v=H1iho17lnoI.

approach to use time series instead of static snippets of positional data. Bauer and Anzer (2021) compare a rule based model, to a machine learning based one to identify the tactical pattern of counterpressing automatically across 20, 000 labels from 97 matches. For their trained model they extract 137 hand-crafted features. The advantage of our approach is that it not only requires far fewer labeled observations, but also works with very simple basic features. It is easily reproducible for any other pattern, and, can be adjusted quickly even if definitions of patterns slightly change when the game-philosophy shifts (e.g., because of a coaching change).

Besides the potential to reduce the costs of manual event data collection, our approach enables several team performance affecting applications: The automatic detection of relevant patterns saves coaches and match-analysis departments not only time, but furthermore increases consistency and offers scalability. This can consequently be used to perform long-term analysis across multiple seasons or even leagues. Furthermore, besides match-analysis this methodology could also be integrated in the player scouting process, by identifying certain beneficial individual action patterns and finding players that exhibit these frequently.

While our work describes the technical framework to achieve these results, for it to be usable in a club environment, one would need to integrate it in an application that fits in seamlessly into daily routines of match-analysis or scouting departments.

RELATED WORK

This paper explores issues related to VAE-based semi-supervised learning, with the main contribution in this field introduced by Kingma et al. (2014). Our SoccerVAE and LabelVAE are clearly inspired by their proposals M1 and M2. Specifically, the authors integrate label information into the assumption of the data generation process, thereby obviating the necessity for the otherwise required supervised learning task on extracted label-feature pairs. Recent work by Joy et al. (2020) argues that explicitly modeling the connection between labels and their corresponding latent variables improves the classification accuracy compared to the M2 approach and allows to learn meaningful representations of data effectively. Maaløe et al. (2016) also improve M2 classification performance by introducing an auxiliary variable that leaves the original model unchanged but increases the flexibility of the variational posterior. This can result in convergence to a parameter configuration that is closer to a local optimum of the actual data likelihood (due to potentially better fits to the complex posterior) while maintaining the computational efficiency of fully factorized models. Siddharth et al. (2017) choose a more generalized formulation of semi-supervised learning with VAE compared to the models in the work by Kingma et al. (2014). Their framework allows choosing complex models, such as when a random variable determines the number of latent variables itself.

In addition to static semi-supervised tasks, this work methodologically touches a branch of research that describes methods involving autoencoders to model sequential data. Bayer and Osendorfer (2014) incorporate stochasticity into vanilla RNNs by making the independently sampled latent variables an additional input at each timestep. Chung et al. (2015) apply a similar model termed VRNN to speech data, sharing parameters between the RNNs for the generative model and the inference network. In Goyal et al. (2017), the latent variable participates to the prediction of the next timestep, and the variational posterior is informed about the whole future in the sequence modeled by an RNN processing the sequence backwards. While the previously mentioned methods sample a separate latent variable at each timestep, Bowman et al. (2015) propose an RNN-based VAE to derive global latent representations for sentences. The approach to modeling human-drawn images discussed in Ha and Eck (2017) shares many architectural similarities to Bowman et al. (2015), but uses an additional backward RNN encoder. Teng et al. (2020) introduce a semi-supervised training objective for modeling sequential data where the model specification draws inspiration from Kingma et al. (2014) and Chung et al. (2015).

CONCLUSIONS

We studied automatic annotation of non-trivial situations in soccer. We proposed to separate the problem into an unsupervised autoencoder to learn a meaningful feature representation and a supervised large-margin classification. The advantage of this separation lied in the use of abundant unlabeled data that allowed for learning a nicely structured feature space so that only a few labeled examples were needed in the classifier to learn the target concept of interest.

We proposed two variants of autoencoders, a straight forward application of existing results (SoccerVAE) and a more sophisticated variant that used auxiliary labels and allowed for even more discriminative feature spaces (LabelVAE). In addition to these two static variants, we devised their sequential peers to account for the spatiotemporal nature of soccer. Empirically, we studied the performance of the four approaches on three different detection tasks, involving cornerkicks, crosses, and counterattacks. The SeqLabelVAE turned out the best competitor and outperformed all others with detection rates of 91% AUC or higher in all problems for only a few labeled examples.

While our methods emerged as valuable tools for detection tasks in soccer, there are some shortcomings that could be addressed in future work. A possible starting point is to compare the implicit regularization of our semi-supervised approach against supervised sequential models with alternate regularization methods (Semeniuta et al., 2016). From the perspective of achieving the lowest possible generalization error, there are several avenues for potential variations. Future work might include alternate probabilistic assumptions (Goyal et al., 2017; Joy et al., 2020) such as conditioning the variational distribution on the full input sequence (Goyal et al., 2017), novel regularization techniques for VAE (Tolstikhin et al., 2017; Ma et al., 2019; Deasy et al., 2020), other approaches to semi-supervised learning (Kingma et al., 2014; Dai and Le, 2015) such as transfer learning (Fabius and Van Amersfoort, 2014; Srivastava et al., 2015), or to achieving consistent agent representations such as graph-networks (Sun et al., 2019; Yeh et al., 2019) and tree-based role alignments (Lucey et al., 2013; Sha et al., 2017; Felsen et al., 2018).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We would like to thank the German Football Association (DFB) for providing the data for this study.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv [Preprint]. arXiv:1603.04467.*

Andrienko, G., Andrienko, N., Anzer, G., Bauer, P., Budziak, G., Fuchs, G., et al. (2019). Constructing spaces and times for tactical analysis in football. *IEEE Trans. Vis. Comput. Graph.* 27, 2280–2297. doi: 10.1109/TVCG.2019.2952129

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

Bauer, P., and 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 Min. Knowl. Disc.* 35, 2009–2049. doi: 10.1007/s10618-021-00763-7

Bayer, J., and Osendorfer, C. (2014). Learning stochastic recurrent networks. *arXiv [Preprint]. arXiv:1411.7610.*

Bjorck, J., Gomes, C., Selman, B., and Weinberger, K. Q. (2018). Understanding batch normalization. *arXiv [Preprint]. arXiv:1806.02375.*

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015). Generating sentences from a continuous space. *arXiv [Preprint]. arXiv:1511.06349.* doi: 10.18653/v1/K16-1002

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A., and Bengio, Y. (2015). A recurrent latent variable model for sequential data. *arXiv [Preprint]. arXiv:1506.02216.*

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Dai, A. M., and Le, Q. V. (2015). Semi-supervised sequence learning. *arXiv [Preprint]. arXiv:1511.01432.*

Deasy, J., Simidjievski, N., and Liò, P. (2020). Constraining variational inference with geometric jensen-shannon divergence. *arXiv [Preprint]. arXiv:2006.10599.*

Dick, U., and 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., and Brefeld, U. (2018). Rating continuous actions in spatial multi-agent problems.

Ekin, A. A., Tekalp, M., and Mehrotra, R. (2003). "Automatic soccer video analysis and summarization," in *IEEE Transactions on Image Processing 12.7*, 796–807.

Erhan, D., Courville, A., Bengio, Y., and Vincent, P. (2010). "Why does unsupervised pre-training help deep learning?" in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings* (Sardinia), 201–208.

Fabius, O., and Van Amersfoort, J. R. (2014). Variational recurrent auto-encoders. *arXiv[Preprint]. arXiv:1412.6581.*

Felsen, P., Lucey, P., and Ganguly, S. (2018). "Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 732–747.

Goyal, A., Sordoni, A., Côté, M.-A., Ke, N. R., and Bengio, Y. (2017). Z-forcing: Training stochastic recurrent networks. *arXiv [Preprint]. arXiv:1711.05411.*

Ha, D., and Eck, D. (2017). A neural representation of sketch drawings. *arXiv [Preprint]. arXiv:1704.03477.*

Hobbs, J., Power, P., Sha, L., Ruiz, H., and Lucey, P. (2018). "Quantifying the value of transitions in soccer via spatiotemporal trajectory clustering," in *MIT Sloan Sports Analytics Conference*, 1–11.

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning* (Lille: PMLR), 448–456.

Joy, T., Schmon, S. M., Torr, P. H., Siddharth, N., and Rainforth, T. (2020). Rethinking semi-supervised learning in vaes. *arXiv [Preprint]. arXiv:2006.10102.*

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. (2014). Semi-supervised learning with deep generative models. *arXiv [Preprint]. arXiv:1406.5298.*

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv [Preprint]. arXiv:1312.6114*

Kolekar, M. H., and Palaniappan, K. (2009). Semantic concept mining based on hierarchical event detection for soccer video indexing. *J. Multimedia* 4, 298–312. doi: 10.4304/jmm.4.5.298-312

Lucey, P., Bialkowski, A., Carr, P., Morgan, S., Matthews, I., and Sheikh, Y. (2013). "Representing and discovering adversarial team behaviors using player roles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Portland, OR: IEEE), 2706–2713.

Ma, X., Zhou, C., and Hovy, E. (2019). Mae: Mutual posterior-divergence regularization for variational autoencoders. *arXiv [Preprint]. arXiv:1901.01498.*

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. (2016). "Auxiliary deep generative models," in *International Conference on Machine Learning* (New York, NY: PMLR),1445–1453.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning (ICML)*. 30.

Motoi, S., Misu, T., Nakada, Y., Yazaki, T., Kobayashi, G., Matsumoto, T., et al. (2012). Bayesian event detection for sport games with hidden Markov model. *Pattern Anal. Appl.* 15, 59–72. doi: 10.1007/s10044-011-

Nair, V., and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Icml.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv [Preprint]. arXiv:1511.06434.*

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning* (Bejing: PMLR), 1278–1286.

Richly, K., Bothe, M., Rohloff, T., and Schwarz, C. (2016). "Recognizing compound events in spatio-temporal football data," in *IoTBD 2016-Proceedings of the International Conference on Internet of Things and Big Data March 2018* (Funchal), 27–35.

Semeniuta, S., Severyn, A., and Barth, E. (2016). Recurrent dropout without memory loss. *arXiv [Preprint]. arXiv:1603.05118.*

Sha, L., Lucey, P., Zheng, S., Kim, T., Yue, Y., and Sridharan, S. (2017). Fine-grained retrieval of sports plays using tree-based alignment of trajectories. *arXiv [Preprint].arXiv:1710.02255.*

Shaw, L., and Glickman, M. (2019). "Dynamic analysis of team strategy in professional football," in *Barca Sports Analytics Summit*, 1–13.

Shaw, L., and Sudarshan, G. (2020). "Routine inspection: A playbook for corner kicks," in *International Workshop on Machine Learning and Data Mining for Sports Analytics* (Cham: Springer).

Siddharth, N., Paige, B., Van de Meent, J.-W., Desmaison, A., Goodman, N. D., Kohli, P., et al. (2017). Learning disentangled representations with semi-supervised deep generative models. *arXiv [Preprint]. arXiv:1706.00400.*

Srivastava, N., Mansimov, E., and Salakhudinov, R. (2015). "Unsupervised learning of video representations using lstms," in *International Conference on Machine Learning* (Lille: PMLR), 843–852.

Stein, M., Seebacher, D., Karge, T., Polk, T., Grossniklaus, M., and Keimet, D. A. (2019). From movement to events: improving soccer match annotations. *Lecture Notes Comput. Sci.* 11295, 130–142. doi: 10.1007/978-3-030-05710-5247_11

Sun, C., Karlsson, P., Wu, J., Tenenbaum, J. B., and Murphy, K. (2019). Stochastic prediction of multi-agent interactions from partial observations. *arXiv [Preprint]. arXiv:1902.09641.*

Teng, M., Le, T. A., Scibior, A., and Wood, F. (2020). Semi-supervised sequential generative models. *arXiv [Preprint]. arXiv:2007.00155.*

Tieleman, T., and Hinton, G. (2012). "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," in *COURSERA: Neural Networks for Machine Learning*, *Vol. 4*, 26–31.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2017). Wasserstein auto-encoders. *arXiv [Preprint]. arXiv:1711.01558.*

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605.

Wickramaratna, K., Chen, M., Chen, S.-C., and Shyu, M.-L. (2005). "Neural network based framework for goal event detection in soccer videos," in *Proceedings-Seventh IEEE International Symposium on Multimedia* (Irvine, CA: IEEE), 21–28.

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv [Preprint]. arXiv:1505.00853.*

Yeh, R. A., Schwing, A. G., Huang, J., and Murphy, K. (2019). "Diverse generation for multi-agent sports games," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 4610–4619.

Zheng, M., and Kudenko, D. (2010). Automated event recognition for football commentary generation. *Int. J. Gaming Comput. Mediated Simulat.* 2, 67–84. doi: 10.4018/jgcms.2010100105

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.