

OPEN



# An integrated dataset of spatiotemporal and event data in elite soccer

Manuel Bassek<sup>1</sup>  , Robert Rein<sup>1</sup>, Hendrik Weber<sup>2</sup> & Daniel Memmert<sup>1</sup>

Data-driven match analysis in soccer is a growing discipline in both research and practice. However, public data is scarce, which raises the barrier for entering this field and decreases reproducibility of methods and results. To bridge this gap, this paper presents a dataset of official match information, event, and position data from seven matches of the German Bundesliga's first and second division. The match information contains meta data about the matches and their participants. The event data contain timestamps along with descriptions of discrete events, like passes, shots, or fouls. The position data contain the x/y-coordinates of every player and the ball. By integrating multiple data modalities – i.e., event logs with timestamps, and x-y coordinates of player and ball positions — the dataset offers a multidimensional view of match dynamics. This dataset supports the validation of existing analytical techniques and facilitates the development of new methodologies in sports analytics. With availability under CC-BY 4.0, it promotes transparency, reproducibility, and the idea of open science in match analysis research.

## Background & Summary

Performance analysis in team sports has become a data driven discipline in recent years<sup>1–3</sup>. For example, in professional soccer, game-related data are recorded using various methods and techniques for every single match across all professional leagues. These data have led to various advancements with respect to research insights and practices<sup>1,4</sup>. In general, three main data modalities are available: (i) Video data either from a broadcast streams or other dedicated camera systems that are used to capture, for example, the tactical behavior on the pitch (wide-angle scouting feed); (ii) event data using timestamps and descriptions of discrete match events like goals, passes and fouls; (iii) position data represented by x/y-pitch coordinates of the players and balls with a high spatial and temporal resolution. Position data have been used, for example, to characterize physiological demands of the players<sup>5,6</sup>, classify and evaluate tactical team behavior<sup>7–10</sup>, value individual actions<sup>11,12</sup>, or predict future states of the match<sup>13–15</sup>. However, these various data are usually not shared publicly due to their proprietary nature, their financial value and to maintain a competitive edge over competitors.

To date, several datasets of different modalities are publicly available. Video datasets like the *SoccerNet* dataset contains video data from multiple elite soccer matches (e.g., top-5 European leagues, FIFA World Cup). Research projects and challenges within the computer vision community as well as quality controlled manual annotations have enriched these datasets with additional information, like camera shot detection, action spotting, player identification and tracking, or pose estimation<sup>16–19</sup>. The event dataset *Wyscout soccer match event dataset* published by Papalardo *et al.*<sup>20</sup> contains one season in five national soccer competitions and two international tournaments with detailed descriptions about on-ball actions. Similarly, the *StatsBomb Open Data* repository<sup>21</sup> contains event data from various competitions and is constantly updated by the commercial event data provider Hudl StatsBomb (Agile Sports Technologies Inc., Lincoln, Nebraska, USA). Position data which, arguably, contain the most information are also the scarce ones in volume. Although extracted position data from the *SoccerNet* dataset is available<sup>19</sup>, reproducing and extending this data requires expertise in computer vision and substantial computing resources<sup>22</sup>. To the best of our knowledge, besides such data, only few public datasets of raw position data in soccer exist with one<sup>23</sup>, three<sup>24</sup> and nine<sup>25</sup> matches, respectively. The first dataset contains of one match tracked with an optical tracking system accompanied with physical (distance covered, distances in speed zones) and tactical (team centroid position) performance indicators<sup>23</sup>. The second dataset

<sup>1</sup>Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Cologne, Germany. <sup>2</sup>DFL, German Football League, Frankfurt, Germany. e-mail: [m.bassek@dshs-koeln.de](mailto:m.bassek@dshs-koeln.de)

used a combination of body sensors and video data to generate the position data<sup>24</sup>. The third dataset, has been generated by the commercial provider SkillCorner (Paris) by extracting position data from broadcast videos. Thus, player information may be limited by their visibility. Finally, Biermann *et al.*<sup>26</sup> published a dataset of video, position data, and manual annotations for 125 minutes of elite team handball.

Although common practice in other scientific fields, in particular in computer science, researchers in sports sciences and match analysis rarely publish their results on publicly available benchmark datasets<sup>27</sup>. This may be in fact, due to the limited availability of public data sources. Additionally, the field of match analysis is currently not open for newcomers outside of clubs as they do not have access to data. Therefore, the dataset presented in this study aims at closing this gap by publishing official, synchronized event and position data from seven German Bundesliga matches. This dataset differs from the aforementioned datasets in three ways: (i) the dataset contains the official data provided by the German Football League (Deutsche Fußball Liga; DFL)<sup>28</sup> which is shared with the clubs and media instead of data generated by research projects or third party data provider; (ii) the data contains a continuous stream of all players' and the ball's position data, generated by a validated multi-camera tracking system; and (iii) contains both detailed event and tracking data from the same soccer matches.

Although the volume of the dataset may not allow for generalizable insights in the match analysis research domain, like several seasons of event data, and does not contain video footage of the matches, it serves as a complementary resource to be used for entry-level analyses as well as a high-quality benchmark dataset. It may therefore benefit the community by increasing the reproducibility of research from various disciplines and promoting open science practices in the field of match analysis.

## Methods

**Match information.** Match information data were collected by Sportec Solutions (Sportec Solutions AG, Unterföhringen, Germany). They are derived from various data sources (clubs, weather stations, ticket sales, on-site operators) and stored and distributed from the official DFL database<sup>28</sup>. Consent to publish this data is given mandatorily during the player registration process. All data was provided by the original data collector, i.e., DFL with the permission to publish them under CC-BY 4.0. General information about the competitors (player's names, jersey numbers, etc.) are entered by the respective clubs. Information about spectator numbers are based on the club's communication and validated with ticket sale numbers. Information about the weather conditions are collected by an on-site operator from local weather stations. An operator also specifies the team tactical formation from 41 possible formation templates. If none of the templates apply to the actual formation, he is instructed to select the most similar template.

**Event data.** A pool of 120 operators (trained human analysts) analyze over 616 matches every year, whereby five operators analyze each match. Operators undergo specific training for several days and are required to analyze several matches in an offline modus before their first live match. One on-site operator (speaker) and one operator in the operations center (writer) generate the initial event data stream aiming for minimum latency. Two additional operators follow the live video footage, add additional data, and check for event data quality. Each designated event is annotated in a custom software. Each event is created with a local UTC timestamp and a custom ID. Depending on the event, additional attributes, like participating players or location is added. Finally, these data are controlled for quality by a supervisor, who is a domain expert with previous one-week training and trial period.

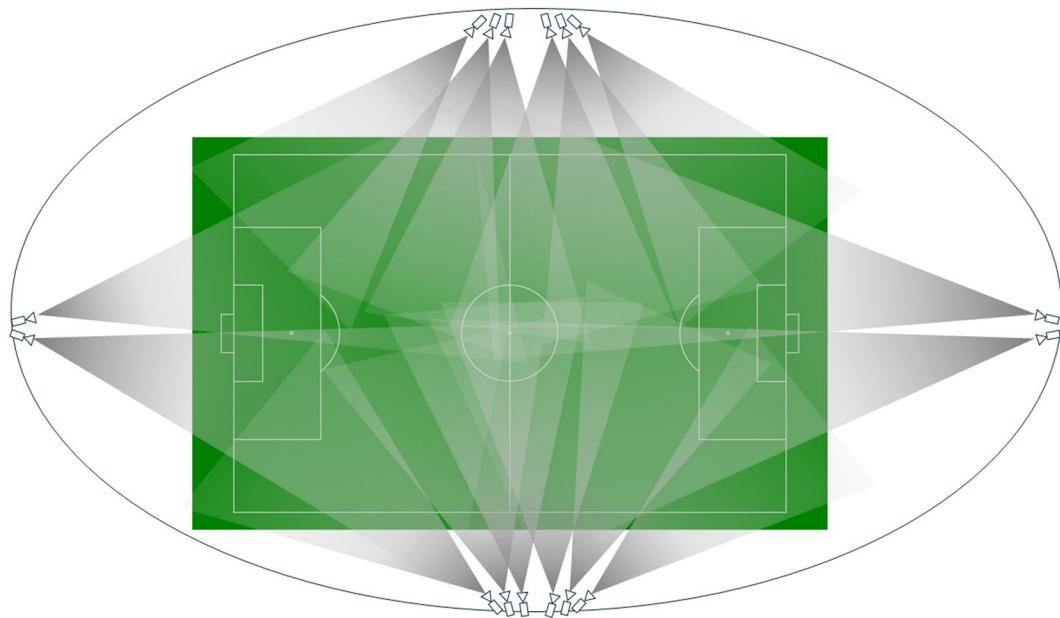
**Position data.** The position data was collected using the multi-camera tracking system TRACAB Gen5 by Chyron Hego (ChyronHego Corporation, Melville, New York, USA). The camera setup has been described in detail in a recent validation study<sup>29</sup> (see Technical Validation). Unlike Electronic Performance Tracking Systems, like Gobal Positioning Systems (GPS) or Local Positioning Systems (LPS), TRACAB does not rely on sensors worn by the athletes, but estimates position data from video data. Depending on the stadium, the system uses 16 to 20 cameras (1920 × 1080 pixels; 25 Hz). Six cameras each are positioned on both sides of the pitch and two cameras are positioned behind each goal. Figure 1 shows the general setup and camera coverage (for 16 cameras).

All moving objects on the pitch are tracked using specialized computer vision algorithms for playfield detection, player and ball detection, and player labelling<sup>30</sup>. The players and the ball are then projected into a local 2D coordinate system via direct linear transformation using a calibrated homography matrix. Further post-processing algorithms are applied to filter outliers and smooth the raw data.

## Data Records

The present data set is released under a CC-BY 4.0 license with the authorization from the DFL and available at figshare<sup>31</sup>. The dataset can also be accessed via the software package *floodlight* for the Python programming language<sup>32</sup>. The dataset contains data from two matches of the German Bundesliga 2022/23 season, as well as five matches from the second division Bundesliga 2022/23 season (Table 1). The Bundesliga is a top-5 league according to the UEFA country coefficient<sup>33</sup>. For each match, three files are available: a match information file, a match events file, and a match position data file. All three files use an XML format as the file architecture. In total, the dataset contains information about 207 players from 10 teams, 11,137 events, and 1,002,644 frames of x/y-coordinates for players and the ball. All data categories are defined in the Catalogue of Definitions published by the DFL<sup>34</sup>. The video footage of the matches is not available in this dataset due to licensing restrictions.

**Match information.** The match information files use an XML container (Box 1) and specify general meta data about the competition, the environmental conditions and competitors.



**Fig. 1** The spatial distribution of the TRACAB Gen 5 cameras following Linke *et al.*<sup>29</sup>.

Match-ID	Div	Home	Away	Result	xG	Date	N <sub>Frames</sub>	N <sub>Events</sub>
J03WMX	1 <sup>st</sup>	1. FC Köln	FC Bayern München	1:2	1.33:1.18	2023/05/27	145,967	1,840
J03WN1	1 <sup>st</sup>	VfL Bochum 1848	Bayer 04 Leverkusen	3:0	1.97:0.78	2023/05/27	141,561	1,459
J03WPY	2 <sup>nd</sup>	Fortuna Düsseldorf	1. FC Nürnberg	0:1	1.01:1.78	2022/10/15	146,211	1,579
J03WOH	2 <sup>nd</sup>	Fortuna Düsseldorf	SSV Jahn Regensburg	4:0	2.66:0.52	2022/08/26	137,214	1,457
J03WQQ	2 <sup>nd</sup>	Fortuna Düsseldorf	FC St. Pauli	1:0	1.25:1.09	2022/11/05	142,345	1,679
J03WOY	2 <sup>nd</sup>	Fortuna Düsseldorf	F.C. Hansa Rostock	3:1	2.14:0.43	2022/09/10	142,536	1,578
J03WR9	2 <sup>nd</sup>	Fortuna Düsseldorf	1. FC Kaiserslautern	1:2	1.16:1.31	2022/11/11	146,810	1,545

**Table 1.** Meta data for all seven available matches. N<sub>Frames/Events</sub>: Number of frames and events in the respective match.

```

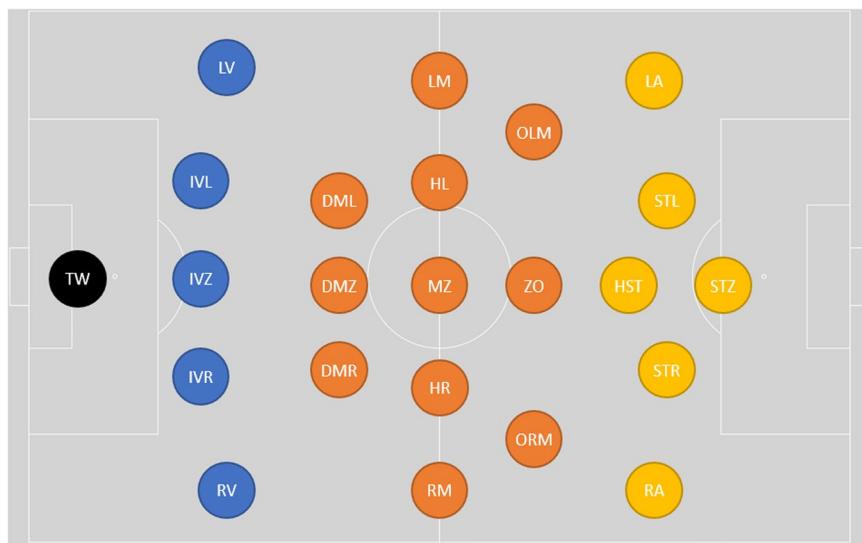
<MatchInformation>
  <General TypeOfSport="Fußball" CompetitionName="Bundesliga" CompetitionId="DFL-COM-000001" Host="Die Liga - Fußballverband e.V.
  (Ligaverband)" Type="Ligabetrieb" MatchDay="34" Season="2022/2023" SeasonId="DFL-SEA-0001K6" PlannedKickoffTime="2023-05-27T13:30:00
  .000+00:00" KickoffTime="2023-05-27T13:30:12.220+00:00" MatchId="DFL-MAT-J03WMX" DlProviderId="182265" MatchTitle="1. FC Köln:FC
  Bayern München" HomeTeamName="1. FC Köln" HomeTeamId="DFL-CLU-000008" GuestTeamName="FC Bayern München" GuestTeamId="DFL-CLU-000006"
  Result="1:2" />
  <Environment Country="Deutschland" StadiumId="DFL-STA-000008" StadiumName="RheinEnergieSTADION" StadiumAddress="Aachener Straße 999, 50933
  Köln" NeutralVenue="false" Roof="open" Floodlights="on" Temperature="30" AirHumidity="21" AirPressure="1023" PitchErosion="low"
  NumberOfSpectators="50000" Capacity="50000" Precipitation="none" SoldOut="true" PitchX="105.00" PitchY="68.00" />
  <Teams>
    <team TeamId="DFL-CLU-000006" TeamName="FC Bayern München" Role="guest" PlayerShirtType="Ausweichtrikot" PlayerShirtMainColor
    ="#25282A" PlayerShirtSecondaryColor="#232222" PlayerShirtNumberColor="#D22630" LineUp="4-2-3-1">
      <Players>
        <Player PersonId="DFL-OBJ-002DR" ShirtNumber="27" FirstName="Yann" LastName="Sommer" Shortname="Y. Sommer" Starting="true"
        PlayingPosition="TW" TeamLeader="false" />
        <Player PersonId="DFL-OBJ-000LT" ShirtNumber="26" FirstName="Sven" LastName="Ulreich" Shortname="S. Ulreich" Starting="false"
        TeamLeader="false" />
      </Players>
      <TrainerStaff>
        <Trainer PersonId="DFL-OBJ-000R0" Role="headcoach" FirstName="Thomas" LastName="Tuchel" Shortname="T. Tuchel" />
        <Trainer PersonId="DFL-OBJ-000QY" Role="assistantHeadcoach" FirstName="Arno" LastName="Michels" Shortname="A. Michels" />
      </TrainerStaff>
    </OfficialStaff></OfficialStaff>
  </Team>
</MatchInformation>

```

**Box 1** Exemplary excerpt from the match information file of match J03WMX. In order to make the layout appropriate, all but two players and staff have been removed.

*General.* General information includes the type of sport and competition, league, season, matchday, kick-off time (ISO 8601 format), and home and away team names and IDs.

*Environmental.* Environmental conditions include the country, stadium name, ID, address, capacity, pitch dimensions (in m), weather conditions (temperature in °C, humidity in %, atmospheric pressure in hPa), and number of spectators.



**Fig. 2** Classification of player roles in playing direction from left to right, adapted from<sup>52</sup>.

```

<Event MatchId="DFL-MAT-J03WMX" X-Source-Position="52.50" EventTime="2023-05-27T15:30:12.230+02:00" Y-Source-Position="34.00" EventId="18226500000006" X-Position="52.50" Y-Position="34.00">
<Kickoff TeamLeft="DFL-CLU-00000G" TeamRight="DFL-CLU-000008" GameSection="firstHalf">
  <Play SemiField="false" Player="DFL-OBJ-0027GG" Team="DFL-CLU-00000G" FromOpenPlay="false" PenaltyBox="false" FlatCross="false" Height="flat" Distance="medium" PlayOrigin="ownHalf" PlayAngle="179.19" Recipient="DFL-OBJ-0027KL" BallPossessionPhase="6" Evaluation="successfullyCompleted">
    <Pass FreeKickLayup="false"/>
  </Play>
</Kickoff>undefined</Event>undefined<Event MatchId="DFL-MAT-J03WMX" X-Source-Position="38.27" EventTime="2023-05-27T15:30:15.059+02:00" Y-Source-Position="33.80" EventId="18226500000007" X-Position="38.27" Y-Position="33.80">
<Play SemiField="false" Player="DFL-OBJ-0027KL" Team="DFL-CLU-00000G" FromOpenPlay="true" PenaltyBox="false" FlatCross="false" Height="high" Distance="long" PlayOrigin="ownHalf" PlayAngle="320.43" BallPossessionPhase="7" Evaluation="unsuccessful">
  <Pass FreeKickLayup="false" Direction="diagonalBall"/>
</Play>undefined</Event>undefined<Event MatchId="DFL-MAT-J03WMX" X-Source-Position="79.65" EventTime="2023-05-27T15:30:33.781+02:00" Y-Source-Position="68.00" EventId="18226500000008" X-Position="79.65" Y-Position="68.00">
<ThrowIn Team="DFL-CLU-000008" Side="right" DecisionTimestamp="2023-05-27T15:30:16.916+02:00">
  <Play SemiField="false" Player="DFL-OBJ-0002B0" Team="DFL-CLU-000008" FromOpenPlay="false" PenaltyBox="false" FlatCross="false" Height="high" Distance="medium" PlayOrigin="ownHalf" PlayAngle="348.75" Recipient="DFL-OBJ-0027KL" BallPossessionPhase="8" Evaluation="unsuccessful">
    <Pass FreeKickLayup="false"/>
  </Play>
</ThrowIn>undefined</Event>undefined

```

**Box 2** The meta data and first five events from the match events file of match J03WMX.

**Teams.** Team information files contain the competing teams' IDs, names, them playing home or away, jersey colors (in hexadecimal) and tactical formation (e.g., "4-2-3-1"). Additionally, each player is listed with his ID, name, shirt number, playing position in German abbreviation (e.g., "LV" for Linksverteidiger, i.e., left back; TW for Torwart, i.e., goal keeper). Figure 2 shows the position abbreviations of players on the pitch. The starting players and team captain are listed as "true" or "false". The coaching staff is listed with their IDs, names, and role (e.g., "trainer"; "assistantTrainer"). Similarly, the official staff is listed (e.g., "doctor", "teamManager").

**Referees.** The referees are listed with their IDs, names and role (e.g., "referee"; "firstAssistant"; "fourthOfficial").

**OtherGameInformation.** This contains the gross and net playing time for the first and second half in seconds.

**Event data.** The event data files (Box 2) contain information about discrete events categorized into player, team and referee actions. Events where one or more players participate are considered as player actions. These events include on-ball actions, tackles, fouls, off-side, and other player actions (i.e., actions that cannot be classified into specific player actions). All set pieces (e.g., kick-off, throw-in, corner, free kick) as well as counter attacks are considered team actions. Events that involve a referee decision are categorized as referee actions (e.g., start and final whistle, substitutions, sanctions).

The events are structured hierarchically (Table 2). Events are derived from general parent classes (e.g., player actions) and further specialized (e.g., on-ball action, pass). In that process, each subclass of events (e.g., blocked shot, saved shot, successful shot) inherits their parent (shot, on-ball action, player action) class's characteristics. Each event also gets contextualized with attributes (e.g., x/y-coordinates in m, goal expectancy). All events contain the attribute timestamp which specifies the time instant when the event occurred (in ISO 8601 format).

Figure 3 shows the distribution of all occurring event types (i.e., differentiated to every subclass) in the dataset. As visible, the 'Play' which specifies a player's action is the most often occurring event. A 'Play' is an attempt from a player to switch ball control to a teammate. However, the same event class 'Play' can be derived from different parent classes. For instance, a free-kick is considered a team action but can be executed as a play. Also,

Event type	Parent-classes	Sub-classes	Attributes
Play	ThrowIn, FreeKick, Kickoff, CornerKick, GoalKick	Pass, Cross	SemiField, Player, Team, FromOpenPlay, PenaltyBox, FlatCross, Height, Distance, PlayOrigin, PlayAngle, Recipient, BallPossessionPhase, Evaluation
OtherBallAction			Player, Team, DefensiveClearance, BallPossessionPhase
TacklingGame			WinnerTeam, Winner, WinnerRole, PossessionChange, GoalKeeperInvolved, Loser, LoserRole, WinnerResult, LoserTeam, Type
Delete	<i>This event indicates an event deleted in post processing and does not contain information.</i>		
ShotAtGoal	Freekick, Penalty	ShotWide, SavedShot, BlockedShot, ShotWoodWork, SuccessfulShot, OtherShot	Team, ExtendedTypeOfShot, Pressure, GoalDistanceGoalkeeper, ShotOrigin, AssistTypeShotAtGoal, AssistShotAtGoal, AngleToGoal, PlayerSpeed, TakerBallControl, CounterAttack, ChanceEvaluation, SetupOrigin, TypeOfShot, TakerSetup, AfterFreeKick, DistanceToGoal, InsideBox, BuildUp, AssistAction, xG, BallPossessionPhase, ShotCondition
Foul			TeamFouler, Fouler, Fouled, TeamFouled, FoulType

**Table 2.** Most frequent events, their parent- and sub-classes, and attributes.

if executed in a special way, the play can get another subclass, e.g., pass → cross. An overview of attributes for the most important events, their possible parent and subclasses, and their attributes is listed in Table 2.

The detailed list and definitions of all events and attributes is available in the Catalogue of Definitions Official Match Data<sup>34</sup>.

The goal expectancy (xG) is attributed to each shot at goal. Goals are in soccer are relatively rare and often influenced by randomness<sup>35</sup>. They may therefore not reflect the actual offensive performance of a team. The xG value estimates the probability of each shot to get converted into a goal<sup>11</sup>. Given a large number of observed shots as training data, shots from similar situations can be cumulated into bins and the xG value calculated as the bin's conversion rate. The sum of xG of all shots, can be interpreted as a more accurate approximation of the offensive performance. The xG-model used in this dataset adjusts the xG value based on ten features: (i) the shot location, (ii) the speed of the player taking the shot, (iii) number of defenders in the line of the shot, (iv) goalkeepers position, (v) a “pressure”-metric<sup>36</sup> on the player taking the shot, (vi) the body part, (vii) the amount of ball control prior to the shot, (viii) the amount of ball control when taking ball possession, (ix) whether the shot followed a free kick, and (x) whether the shot was a free kick<sup>11</sup>. The model has been trained on data from 105,627 shots taken in the German Bundesliga. An evaluation of feature importance using Shapely values shows that the features distance to goal, distance of the goalkeeper to goal and the angle to goal have the highest impact on the model output.

**Position data.** The position data files (Box 3) specify the raw positions of each player and the ball together with further meta data. The meta data contain the respective match ID, pitch size and the start time of the data collection, i.e., kickoff of first half. For each player and the ball, the positions are stored in a list of frames for each game section (first/second half) and player. Each frame has the attributes frame number ( $N$ ), local time stamp ( $T$  in ISO 8601), x- and y-coordinates ( $X/Y$  in m), distance covered since the preceding frame ( $D$  in cm), speed ( $S$  in km/h), acceleration ( $A$  in m/s<sup>2</sup>), and minute of play ( $M$ ).

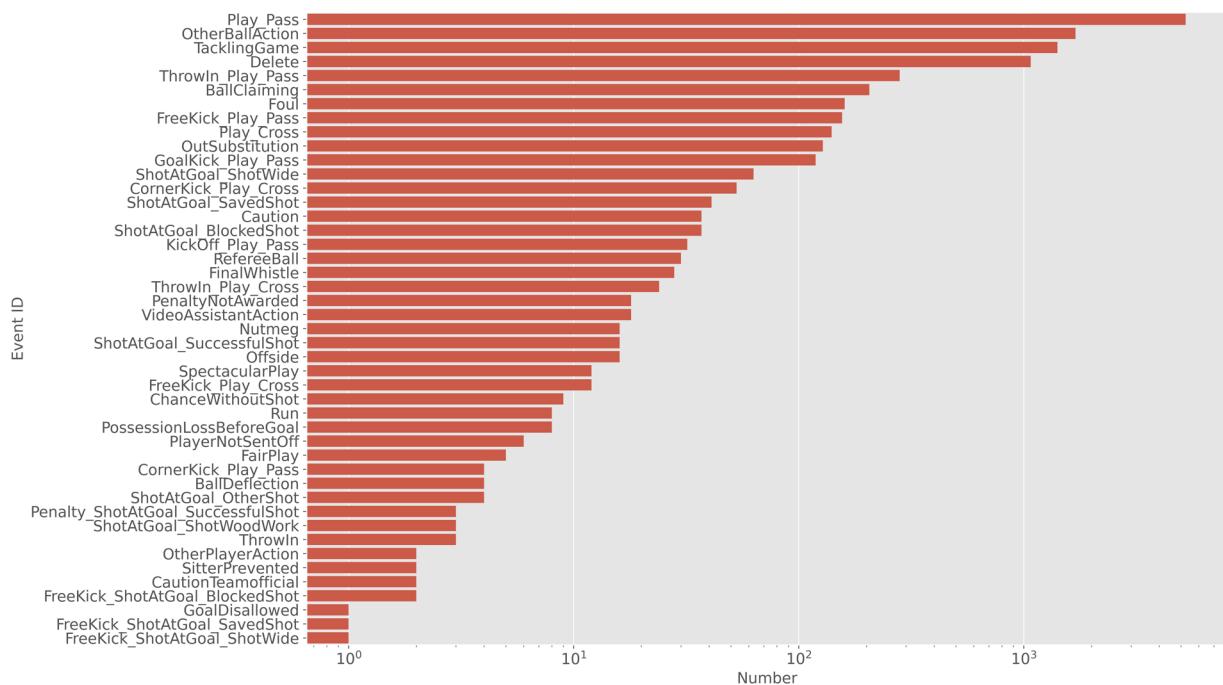
For the ball, in addition to the position data further information about the ball height ( $Z$ ), the ball possession status (BallPossession; 1 = home team in possession, 2 = away team in possession) and the match state (BallStatus; 0 = ball inactive, 1 = ball active) are provided. The ball possession is defined by the player in control of the ball. The match state is inactive when the match is interrupted by the referee, for example after a foul or during a substitution

## Technical Validation

**Match information.** The match information is extracted from the official DFL database. All data are validated throughout the process of data collection, e.g. with data from local weather stations, ticket sales records, and quality control by independent raters. To assure completeness of this dataset, all information was manually validated against other public data sources (e.g., Transfermarkt.de). The validity of the players positions, which result in the selected tactical formation is depicted in Fig. 4. The figure shows the kernel-density of players with different positions from the home team in the first half of match J03WOH (Table 1). Their general space occupation matches their position in the team formation.

**Event data.** The possible errors of event tagging can be divided into semantic<sup>26</sup> and temporal<sup>37</sup>. Semantic errors refer to false information attributed to the event. For example, a pass is labelled as successful, although in reality it failed. In this dataset, all information is checked for plausibility live and after the match to guarantee high semantic quality of the labels.

Temporal errors refer to inaccurate timestamps. As the event data and the position data are generated through separate processes inaccuracy may occur. This often leads to further problems downstream when analyzing the data, for example when analysts want to synchronize the event data with position data<sup>11</sup>. Some automatic approaches have been proposed in the literature<sup>11,37</sup>. However, these have not been applied to this dataset. Still, the quality of the timestamps is controlled after the game is finished. To illustrate the temporal accuracy,



**Fig. 3** Distribution of event occurrences. The x-axis is scaled on the logarithm to base 10.

```

<Positions EventTime="2023-05-27T13:30:12.600+00:00">
  <MetaData MatchId="DFL-MAT-J03WMX" Type="pitch-size">
    <PitchSize X="105.00" Y="68.00" />
  </MetaData>
  <FrameSet GameSection="firstHalf" MatchId="DFL-MAT-J03WMX" TeamId="BALL" PersonId="DFL-0B1-0000XT">
    <Frame N="10000" T="2023-05-27T13:30:12.600+00:00" X="-0.18" Y="-0.17" Z="0.00" D="0.00" S="1.98" A="0.00" M="1" BallPossession="2" BallStatus="1"/>
    <Frame N="10001" T="2023-05-27T13:30:12.640+00:00" X="-0.40" Y="-0.19" Z="0.00" D="57.78" S="2.02" A="0.00" M="1" BallPossession="2" BallStatus="1"/>
    <Frame N="10002" T="2023-05-27T13:30:12.680+00:00" X="-0.97" Y="-0.21" Z="0.01" D="56.64" S="2.02" A="0.00" M="1" BallPossession="2" BallStatus="1"/>
    <Frame N="10003" T="2023-05-27T13:30:12.720+00:00" X="-1.54" Y="-0.24" Z="0.01" D="57.78" S="2.02" A="0.00" M="1" BallPossession="2" BallStatus="1"/>
    <Frame N="10004" T="2023-05-27T13:30:12.760+00:00" X="-2.12" Y="-0.26" Z="0.02" D="57.75" S="2.02" A="0.00" M="1" BallPossession="2" BallStatus="1"/>
    <Frame N="10005" T="2023-05-27T13:30:12.800+00:00" X="-2.69" Y="-0.28" Z="0.02" D="56.67" S="51.88" A="0.00" M="1" BallPossession="2" BallStatus="1"/>
    <Frame N="10006" T="2023-05-27T13:30:12.840+00:00" X="-3.26" Y="-0.31" Z="0.03" D="57.78" S="51.30" A="0.00" M="1" BallPossession="2" BallStatus="1"/>
    <Frame N="10007" T="2023-05-27T13:30:12.880+00:00" X="-3.83" Y="-0.33" Z="0.03" D="56.64" S="51.77" A="0.00" M="1" BallPossession="2" BallStatus="1"/>
  </FrameSet>

```

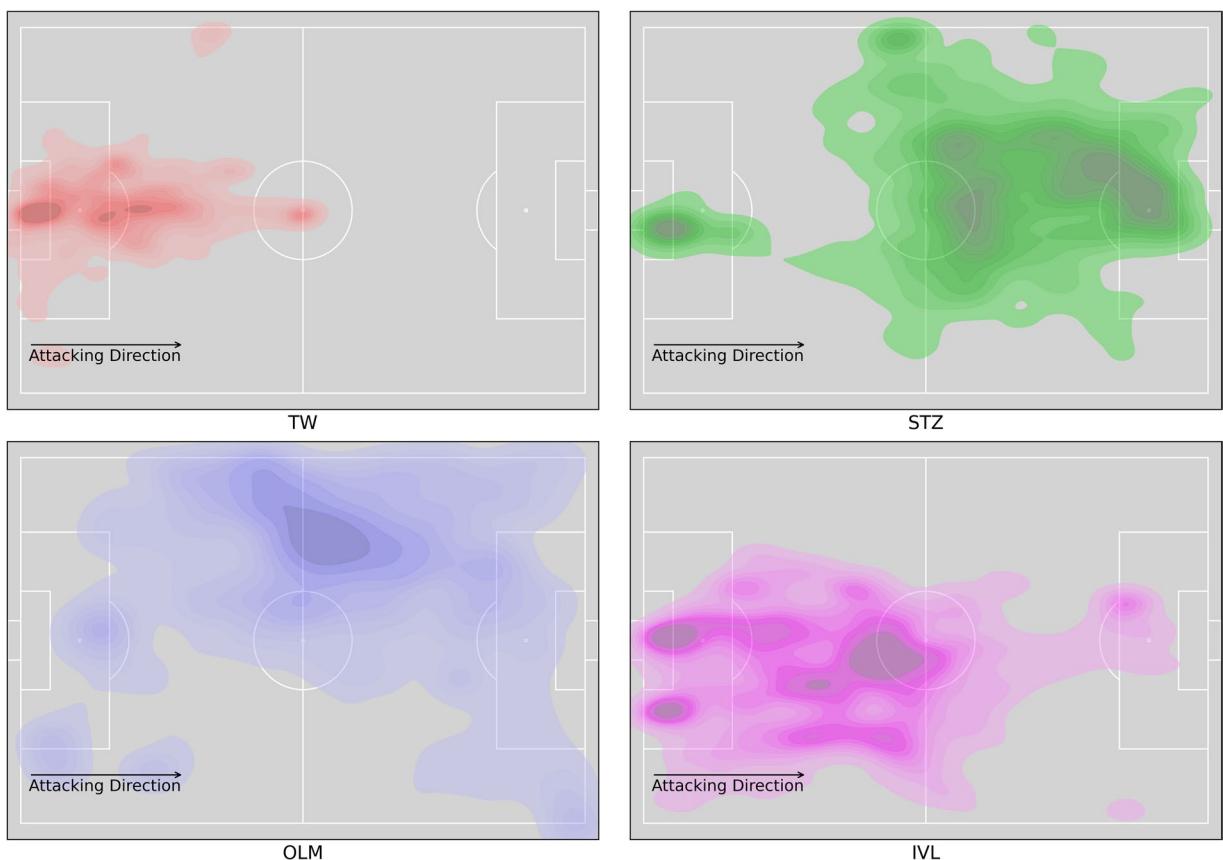
**Box 3** The meta data and first eight frames from the ball data in the position data file of match J03WMX.

Fig. 5 shows the trajectories of all players and the ball five seconds before and at an exemplary time frame, the first successful shot on target in match J03WOH (Table 1).

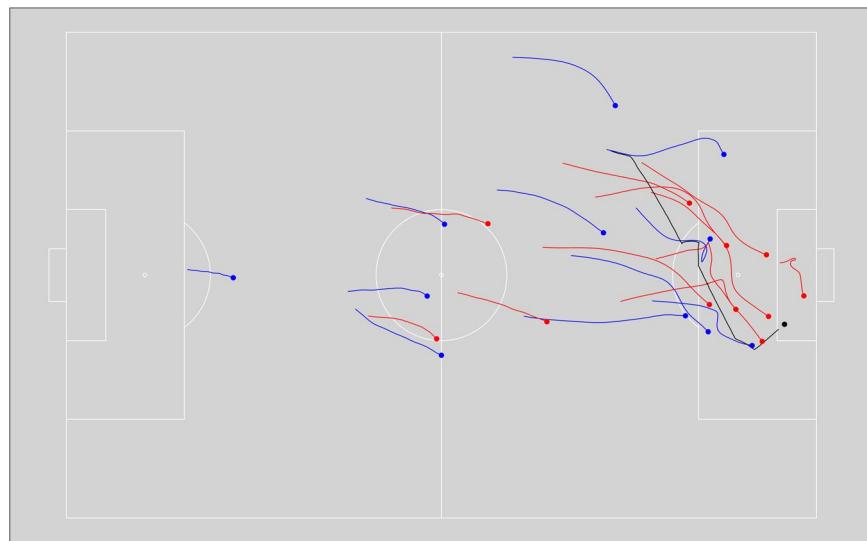
**Expected goals.** The dataset contains a total of 171 shots with an xG attribute. In total, 19 goals were scored in all seven matches with sum of the xG of 18.42. The xG range between 0.01 and 0.85 with an average xG of  $0.11 \pm 0.15$ , similar to the average shot-to-goal conversion rate<sup>38</sup>. However, the median xG is 0.05, indicating that more shots are taken at lower xG values. Figure 6 presents all shot locations, categorized by the shot type, their success and xG. Generally, most shots were performed from inside the box, and most goals were scored from inside the box. Although this small sample analysis already indicates patterns in scoring behavior and the xG only slightly underestimates the actual goals scored, the number of matches in this dataset are not sufficient to representatively evaluate or compare different xG models. However, novel models can be evaluated on this dataset to guarantee transparency on the used methods and their output.

**Position data.** Generally, position data from optical tracking systems are prone to three errors: (i) inaccurate projection of the coordinates, (ii) missing data due to occlusions, (iii) false assignment of players identities (ID swaps)<sup>30</sup>.

Erroneous projection may happen due to issues in camera settings, pitch detection, player detection<sup>30</sup>, or external factors, like smoke from pyrotechnics. The resulting error has been investigated in a recent validation study<sup>29</sup>. An infrared camera-based motion capture system VICON (Vicon Motion Systems Ltd., Oxford, UK) was used as the reference system as it is usually referred to as the gold standard for optical tracking systems. Both systems were set up to cover a  $30 \times 30$  m area inside a soccer stadium. Inside the area, a soccer-specific parkour

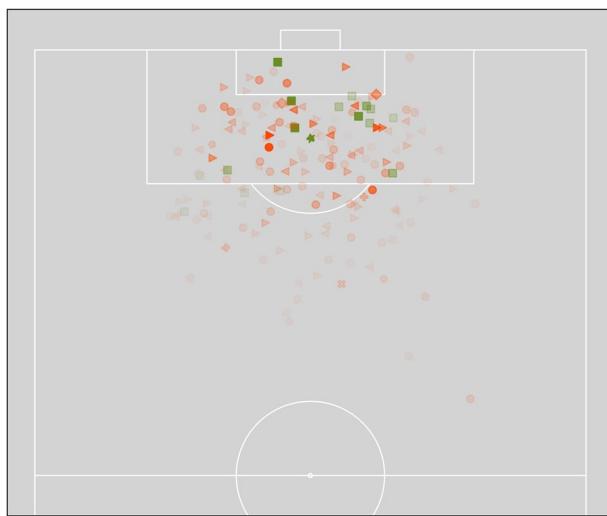


**Fig. 4** Kernel-density plot of different playing positions from the home team in match J03WOH.



**Fig. 5** Trajectories of players for the five seconds until the first successful shot on target. Blue: Attacking team; Red: Defending team; Black: Ball;  $\sim$ : coordinates before shot;  $\bullet$ : coordinate at the timestamp of the shot.

was designed to test typical movements, including accelerations, decelerations, sprints, changes of directions, and curved running paths. Additionally, a small-sided game mode was played by ten athletes. The root mean square error (RMSE) was calculated for the raw positions, velocities and acceleration data. Results showed a RSME range from 0.06 to 0.18 m, 0.03 to 0.09  $\text{m}\cdot\text{s}^{-1}$ , and 0.06 to 0.27  $\text{m}\cdot\text{s}^{-2}$  for positions, velocities, and accelerations, respectively. It should be noted that the tracking error of TRACAB is greater when players are running at greater velocities. This heteroscedastic distribution of errors can also be observed for sensor-based electronic player tracking systems (EPTS)<sup>39,40</sup> systems. Therefore, appropriate data cleaning algorithms, like low-pass



**Fig. 6** Expected goals analysis. Each marker corresponds to a shot event at this location: Green = successful shot; Red = Unsuccessfull shot. Marker shape corresponds to the shot type:  $\circ$  = ShotWide;  $>$  = SavedShot;  $<$  = BlockedShot;  $\square$  = SuccessfulShot;  $\diamond$  = OtherShot;  $*$  = Penalty;  $\lozenge$  = ShotWoodWork;  $+$  = FreeKick\_BlockedShot;  $X$  = FreeKick\_SavedShot;  $\bigcirc$ ,  $\bigtriangleup$  = FreeKick\_ShotWide. Higher xG values correspond to lower transparency.

filtering should be applied before further analysis, especially when deriving velocities and accelerations in high-intensity situations<sup>41</sup>.

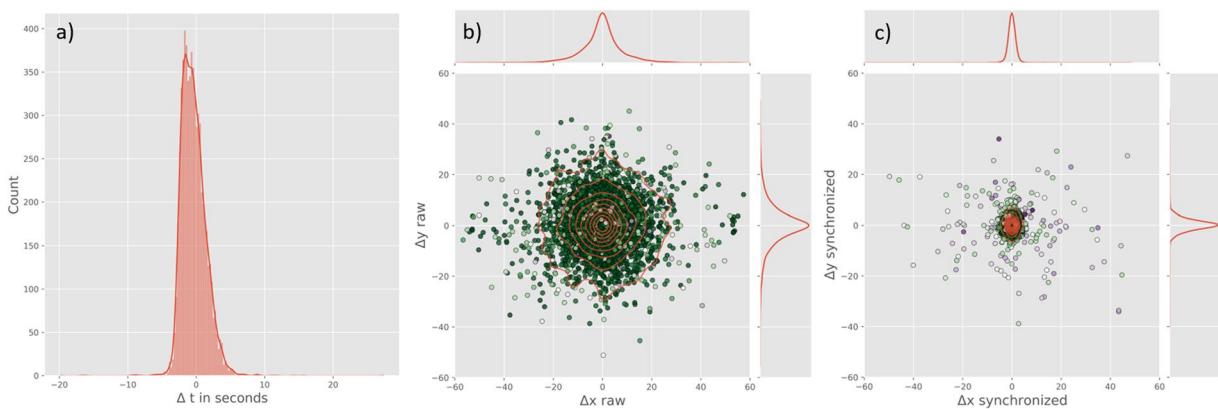
Missing data usually occur due to occlusion of athletes during situations of high spatial density (e.g., corners)<sup>29,30</sup>. To resolve occlusions, the combination of multiple camera angles is necessary. The TRACAB Gen5 system uses 16 cameras and covers each area from multiple angles. During their validation study, Linke *et al.*<sup>29</sup> found no erroneous or missing frames. Further, all data are visually controlled by live operators for plausibility and undergo a final quality check after the game.

The player identities are assigned manually before the start of data collection. Swaps of identities may happen if two players are in close proximity to each other and the system confuses their identity. Such swaps are resolved by the live operators or during the final quality control.

**Synchronicity between event and position data.** Both, the position data and the events contain an ISO 8601 timestamp that can be used to map events onto the position data and vice versa. However, due to the human error in the manual annotation process of the event data, unsystematic errors arise when simply aligning the respective timestamps. This problem is well known in the community and has been addressed by several synchronization approaches<sup>11,37,42</sup>. Common to these approaches is the definition of cost functions which describe specific features for a given event. For example, at the instance of a passing event, the distance between the ball and the passing player is small and subsequently increases, accompanied by an acceleration peak of the ball. The cost functions are aggregated in a reasonable time window around the timestamp attributed to the event and the position data frame with the maximum value gets synchronized with the event. Since this approach can unintentionally swap the order of events, Kwiatkowski and Clark<sup>43</sup> further suggested to use the Needleman-Wunsch algorithm<sup>44</sup> to synchronize the position data and event timeseries. Van Roy and colleagues<sup>45</sup> evaluated this approach and suggested a certainty value based on the unweighted sum of the cost functions.

We applied this approach using the software package *DataBallPy* (v0.5.3) for Python to synchronize all ( $n = 6251$ ) passes and shots in this dataset. For the cost functions, a time difference between the event timestamp and the position data timestamp was designed that converges to one, outside of a window of ten seconds. Further, the distance between the coordinates of the event data and the ball's position data, the distance between the player involved in the event and the ball, the rate of change in distance between the player and the ball, and the ball acceleration are modeled as sigmoid functions.

An evaluation of the synchronization is visible in Fig. 7. The difference between the unsynchronized and synchronized timestamps is normally distributed around an average of  $-0.37 \pm 1.82$  s. The maximum difference between the unsynchronized and synchronized timestamp is 27.34 s. Similarly, the spatial difference between the coordinates in the event data and the ball coordinates in the position data at the time of the event were evaluated (Fig. 7). The differences are normally distributed around zero for both, the unsynchronized and synchronized event data. The average distance of the unsynchronized is  $9.37 \pm 8.39$  m with a maximum distance of 58.61 m. The synchronization reduced the distance to  $2.61 \pm 3.60$  m with a maximum distance of 75.35 m. Generally, events with high distance may be explainable by misclassifications through edge cases in the cost functions, or erroneous information in the event data (e.g., the wrong player attributed with a pass). However, as this analysis did not consider a “ground-truth” by analyzing the video footage of the matches, the underlying mechanisms are speculative. A systematic analysis of the accuracy of different synchronization algorithms would be beyond the scope for this paper, but future work can use this dataset as a benchmark for that purpose.



**Fig. 7** Evaluation of the event-synchronization. (a) Histogram of the differences between the event timestamp before and after the synchronization; (b) Jointplot of differences between the coordinates attributed in the event data and the ball position at the event timestamp before the synchronization; (c) Jointplot of differences between the coordinates attributed in the event data and the ball position at the event timestamp after the synchronization. A kernel density estimation is visible as red lines. The certainty score is color-coded for each observation with dark green, white, and purple corresponding to high, moderate, and low values, respectively.

### Usage Notes

The *floodlight*<sup>32</sup> software package (v0.4.0) for the Python programming language has been used to process and visualize the data for this paper. The package specializes in routines for parsing, processing, manipulating, and plotting data from a variety of sports data providers with a focus on scientific computing. The raw data can be downloaded from the repository and imported via the ‘*floodlight.io.dfl*’ module. Alternatively, it can be directly accessed via the ‘*floodlight.io.datasets*’ module.

The package leverages on known packages, like *numpy*, *pandas*, or *scipy* for specialized core data objects, to represent event and position data, pitches and more. The *floodlight* package also provides out of the box methods for smoothing (e.g., Butterworth low-pass filter) calculating metrics for intensity (distance, velocity, acceleration, metabolic power), approximate entropy, geometric properties (team centroid, stretch index), space control (Voronoi tesselation), and plotting individual data points or trajectories. All implemented models have been used and evaluated in peer-reviewed publications. The package has an extensive documentation and getting-started tutorials for newcomers in the field of match analysis in sports and is suitable for professional, non-professional and teaching environments, alike (<https://floodlight.readthedocs.io>).

Since this dataset is limited to seven matches, the generalizability of findings in with regard to match analysis may be low. Although approaches to match analysis considering only single matches exist<sup>45,46</sup>, more recent domain specific analysis contain over 100 studies on average<sup>47,48</sup> with up to 1200 matches<sup>49</sup>. Since the outcome of single matches may be heavily influenced by individual decision making or luck<sup>35</sup> robust findings in match analysis should be based on representative sample sizes<sup>50,51</sup>. However, the strength of this dataset is that it can be utilized for a variety of reproducibility and benchmarking tasks currently lacking in the literature, especially methodological approaches including but not limited to space control, expected value, data synchronization, trajectory prediction, network analysis, trajectory clustering, formation detection, data mining, game segmentation, and visualization<sup>3</sup>.

### Code availability

The code used for the visualizations is available on GitHub (<https://github.com/spoho-datasience/idsse-data>). All figures have been produced with Python v3.10, *floodlight* v0.4.0, and *seaborn* v0.13.2

Received: 22 August 2024; Accepted: 20 January 2025;

Published online: 01 February 2025

### References

1. Rein, R. & Memmert, D. Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus* **5**, 1410 (2016).
2. Goes, F. R. *et al.* Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *Eur J Sport Sci* **21**, 481–496 (2021).
3. Gudmundsson, J. & Horton, M. Spatio-temporal analysis of team sports. *ACM Computing Surveys* **50**, 1–34 (2017).
4. Herold, M., Kempe, M., Bauer, P. & Meyer, T. Attacking key performance indicators in soccer: Current practice and perceptions from the elite to youth academy level. *J Sport Sci Med* **20**, 158 (2021).
5. Carling, C., Bradley, P., McCall, A. & Dupont, G. Match-to-match variability in high-speed running activity in a professional soccer team. *J Sport Sci* **34**, 2215–2223 (2016).
6. Bradley, P. S. *et al.* High-intensity running in English FA Premier League soccer matches. *J Sport Sci* **27**, 159–168 (2009).
7. Low, B. *et al.* A Systematic Review of Collective Tactical Behaviours in Football Using Positional Data. *Sports Med* **50**, 343–385, <https://doi.org/10.1007/s40279-019-01194-7> (2020).
8. Müller-Budack, E., Theiner, J., Rein, R. & Ewerth, R. ‘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 - MMSports '19* 25–33 <https://doi.org/10.1145/3347318.3355527> (ACM Press, Nice, France, 2019).

9. Bialkowski, A. *et al.* Large-scale analysis of soccer matches using spatiotemporal tracking data. in *2014 IEEE International Conference on Data Mining* 725–730 <https://doi.org/10.1109/ICDM.2014.133> (IEEE, Shenzhen, China, 2014).
10. Low, B., Rein, R., Schwab, S. & Memmert, D. Defending in 4-4-2 or 5-3-2 formation? Small differences in footballers' collective tactical behaviours. *J Sport Sci* **40**, 351–363 (2022).
11. Bauer, P. & Anzer, G. A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Front Sports Act Liv* **3**, 53 (2021).
12. Link, D., Lang, S. & Seidenschwarz, P. Real time quantification of dangerousness in football using spatiotemporal tracking data. *PLoS ONE* **11**, e0168768 (2016).
13. Wunderlich, F. *et al.* Assessing machine learning and data imputation approaches to handle the issue of data sparsity in sports forecasting. *Mach Learn* **114**, 48, <https://doi.org/10.1007/s10994-024-06651-7> (2025).
14. Raabe, D., Nabben, R. & Memmert, D. Graph representations for the analysis of multi-agent spatiotemporal sports data. *Appl Intell* **53** (2023).
15. Dick, U. & Brefeld, U. Learning to rate player positioning in soccer. *Big Data* **7**, 71–82 (2019).
16. Giancola, S., Amine, M., Dghaily, T. & Ghanem, B. SoccerNet: A scalable dataset for action spotting in soccer videos. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 1792–179210, <https://doi.org/10.1109/CVPRW.2018.00223> (IEEE, Salt Lake City, UT, 2018).
17. Delige, A. *et al.* SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 4503–4514, <https://doi.org/10.1109/CVPRW53098.2021.00508> (IEEE, Nashville, TN, USA, 2021).
18. Jiang, T. *et al.* WorldPose: A world cup dataset for global 3D human pose estimation. in *European Conference on Computer Vision* (2025).
19. Somers, V. *et al.* SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 3293–3305, <https://doi.org/10.1109/CVPRW63382.2024.00334> (IEEE, Seattle, WA, USA, 2024).
20. Pappalardo, L. *et al.* A public data set of spatio-temporal match events in soccer competitions. *Sci Data* **6**, 236 (2019).
21. StatsBomb Open Data. <https://github.com/statsbomb/open-data> (2024).
22. Theiner, J. *et al.* Extraction of positional player data from broadcast soccer videos. in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 1463–1473 <https://doi.org/10.1109/WACV51458.2022.00153> (IEEE, Waikoloa, HI, USA, 2022).
23. Goes & Kempe, M. WCSF Solution Room. [https://github.com/MatKempeGroningen/WCSF\\_SolutionRoom](https://github.com/MatKempeGroningen/WCSF_SolutionRoom) (2023).
24. Pettersen, S. A. *et al.* Soccer video and player position dataset. in *Proceedings of the 5th ACM Multimedia Systems Conference* 18–23 <https://doi.org/10.1145/2557642.2563677> (ACM, Singapore Singapore, 2014).
25. Skillcorner SkillCorner Open Data. <https://github.com/SkillCorner/opendata>.
26. Biermann, H. *et al.* A unified taxonomy and multimodal dataset for events in invasion games. in *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports* 1–10, <https://doi.org/10.1145/3475722.3482792> (ACM, Virtual Event China, 2021).
27. Bassek, M., Raabe, D., Banning, A., Memmert, D. & Rein, R. Analysis of contextualized intensity in Men's elite handball using graph-based deep learning. *J Sport Sci* **41**, 1299–1308 (2023).
28. Die Offiziellen Spieldaten | DFL Deutsche Fußball Liga. <https://www.dfl.de/de/hintergrund/spieldaten/der-ursprung-innovativer-statistiken-die-offiziellen-spieldaten/>.
29. Linke, D., Link, D. & Lames, M. Football-specific validity of TRACAB's optical video tracking systems. *PLoS ONE* **15**, e0230179 (2020).
30. Manafirad, M., Ebadi, H. & Abrishami Moghaddam, H. A survey on player tracking in soccer videos. *Comput Vis Image Underst* **159**, 19–46 (2017).
31. Bassek, M., Rein, R., Weber, H. & Memmert, D. An integrated dataset of spatiotemporal and event data in elite soccer. *figshare* <https://doi.org/10.6084/m9.figshare.28196177> (2025).
32. Raabe, D. *et al.* floodlight - A high-level, data-driven sports analytics framework. *J Open Source Softw* **7**, 4588 (2022).
33. Country coefficients | UEFA rankings | UEFA.com. <https://www.uefa.com/nationalassociations/uefarankings/country/?year=2023>.
34. Catalogue of Definitions Official Match Data Version 5.1. (DFL Deutsche Fußball Liga GmbH, 2020).
35. Wunderlich, F., Seck, A. & Memmert, D. The influence of randomness on goals in football decreases over time. An empirical analysis of randomness involved in goal scoring in the English Premier League. *J Sport Sci* **39**, 2322–2337 (2021).
36. Andrienko, G. *et al.* Visual analysis of pressure in football. *Data Min Knowl Disc* **31**, 1793–1839 (2017).
37. Biermann, H. *et al.* Synchronization of passes in event and spatiotemporal soccer data. *Sci Rep* **13**, 15878 (2023).
38. Lamas, L., Senatore, J. V. & Fellingham, G. Two steps for scoring a point: Creating and converting opportunities in invasion team sports. *PLoS ONE* **15**, e0240419 (2020).
39. Luteberget, L. S. & Gilgjen, M. Validation methods for global and local positioning-based athlete monitoring systems in team sports: A scoping review. *BMJ Open Sport Exerc* **6**, e000794 (2020).
40. Blauberger, P., Marzilger, R. & Lames, M. Validation of player and ball tracking with a local positioning system. *Sensors* **21**, 1465 (2021).
41. Ellens, S., Middleton, K., Gastin, P. B. & Varley, M. C. Techniques to derive and clean acceleration and deceleration data of athlete tracking technologies in team sports: A scoping review. *J Sport Sci* **40**, 1772–1800 (2022).
42. Van Roy, M., Cascioli, L. & Davis, J. ETSY: A rule-based approach to Event and Tracking data SYnchronization. in *Machine Learning and Data Mining for Sports Analytics* (eds. Brefeld, U., Davis, J., Van Haaren, J. & Zimmermann, A.) vol. 2035 11–23 (Springer Nature Switzerland, Cham, 2024).
43. Kwiatkowski, M. & Clark, A. The right way to synchronize event and tracking data. <https://kwiatkowski.io/sync.soccer>.
44. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443–453 (1970).
45. Vilar, L., Araújo, D., Davids, K. & Bar-Yam, Y. Science of winning soccer: Emergent pattern-forming dynamics in association football. *J Syst Sci Complex* **26**, 73–84 (2013).
46. Frencken, W., Poel, H. D., Visscher, C. & Lemmink, K. Variability of inter-team distances associated with match events in elite-standard soccer. *J Sport Sci* **30**, 1207–1213 (2012).
47. Forcher, L., Forcher, L., Altmann, S., Jekauc, D. & Kempe, M. The success factors of rest defense in soccer – A mixed-methods approach of expert interviews, tracking data, and machine learning. *J Sport Sci Med* **22**, 707–725 (2023).
48. Lepschy, H., Wäsche, H. & Woll, A. How to be successful in football: A systematic review. *Open Sports Sci J* **11**, 3–23 (2018).
49. Stöckl, M., Seidl, T., Marley, D. & Power, P. Making offensive play predictable - Using a graph convolutional network to understand defensive performance in soccer. in *MIT SLOAN Sports Analytics Conference* 1–19 (2021).
50. Mehta, S., Bassek, M., Garnica-Caparrós, M. & Memmert, D. "Chop and Change": Examining the occurrence of squad rotation and its effect on team performance in top European football leagues. *Int J Sports Sci Coach* **19**, 2467–2475 (2024).
51. Raabe, D., Biermann, H., Bassek, M., Memmert, D. & Rein, R. The dual problem of space: Relative player positioning determines attacking success in elite men's football. *J Sport Sci* **42**, 1821–1830 (2024).
52. Definitionskatalog Offizielle Spieldaten Version 3.0. (DFL Deutsche Fußball Liga GmbH).

## Acknowledgements

This work has been funded by the German Research Foundation (DFG) under grant number 522904388.

## Author contributions

M.B. processed the data, made plots and wrote the manuscript. R.R. wrote and reviewed the manuscript. H.W. collected the data and reviewed the manuscript. D.M. provided funding and reviewed the manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to M.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025