Machine Learning (2024) 113:6977–7010
https://doi.org/10.1007/s10994-024-06585-0

Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned

Jesse Davis[1,2] · Lotte Bransen[1,2] · Laurens Devos[1,2] · Arne Jaspers[3] · Wannes Meert[1,2] · Pieter Robberechts[1,2] · Jan Van Haaren[1,2,4] · Maaike Van Roy[1,2]

Received: 2 October 2023 / Revised: 12 June 2024 / Accepted: 13 June 2024 /
Published online: 17 July 2024

Abstract
There has been an explosion of data collected about sports. Because such data is extremely rich and complex, machine learning is increasingly being used to extract actionable insights from it. Typically, machine learning is used to build models and indicators that capture the skills, capabilities, and tendencies of athletes and teams. Such indicators and models are in turn used to inform decision-making at professional clubs. Designing these indicators requires paying careful attention to a number of subtle issues from a methodological and evaluation perspective. In this paper, we highlight these challenges in sports and discuss a variety of approaches for handling them. Methodologically, we highlight that dependencies affect how to perform data partitioning for evaluation as well as the need to consider contextual factors. From an evaluation perspective, we draw a distinction between evaluating the developed indicators themselves versus the underlying models that power them. We argue that both aspects must be considered, but that they require different approaches. We hope that this article helps bridge the gap between traditional sports expertise and modern data analytics by providing a structured framework with practical examples.

Keywords Sports analytics · Challenges with evaluation · Indicator evaluation · Model evaluation · Model verification · Reliability

1 Introduction

Over the past decades, data analytics has played a more influential role in a wide variety of sports, ranging from single-player sports such as chess (Silver et al., 2017) to team-based sports such as baseball, basketball, hockey, and soccer (Albert et al., 2017; Baumer et al., 2023) Nowadays, large amounts of data are being collected about both the physical states of athletes as well as technical performances. This is facilitated by a range of technologies including GNSS (Global Navigation Satellite System) trackers integrated with inertial measurement units (e.g., Catapult Sports), biometric sensors (e.g., heart rate monitors), optical tracking systems (e.g., TRACAB, Second Spectrum, HawkEye) and computer

Editor: Ulf Brefeld.

Extended author information available on the last page of the article

vision-assisted annotation of events from broadcast video (e.g., Stats Perform, StatsBomb). The volume, complexity, and richness of these data sources have made machine learning (ML) an increasingly important analysis tool. Consequently, ML is being used to inform decision-making in professional sports. On the one hand, it is used to extract actionable insights from the large volumes of data related to player performance, tactical approaches, and the physical status of players. On the other hand, it is used to partially automate tasks such as video analysis (Carling et al., 2005) that are typically done manually.

At a high level, ML plays a role in team sports in three areas:

1. **Player recruitment.** Ultimately, recruitment involves (1) assessing a player's skills and capabilities on a technical, tactical, physical and mental level and how they will evolve, (2) projecting how the player will fit within the team, and (3) forecasting how their financial valuation will develop. (c.f., Bransen et al. (2019), Decroos et al. (2019), Liu and Schulte (2018), Franks et al. (2015))
2. **Match preparation.** Preparing for a match requires performing an extensive analysis of the opposing team to understand their tendencies and tactics. This can be viewed as a SWOT analysis, which particularly focuses on the opportunities and threats. How can we punish the opponent? How can the opponent punish us? These findings are used by the coaching staff to prepare a game plan. Typically, such reports are prepared by analysts who spend many hours watching videos of upcoming opponents. The analysts must annotate footage and recognize reoccurring patterns, which is a very time-consuming task. Learned models can automatically identify patterns that are missed or not apparent to humans (e.g., subtle patterns in big data) (Shaw & Gopaladesikan, 2021), automate tasks (e.g., tagging of situations) (Bauer & Anzer, 2021; Miller & Bornn, 2017) that are done by human analysts, and give insights into players' skills.
3. **Management of a player's health and fitness.** Building up and maintaining a player's fitness level is crucial for achieving good performances (Halson, 2014; Bourdon et al., 2017). However, training and matches place athletes' bodies under tremendous stress. It is crucial to monitor fitness, have a sense of how much an athlete can do or, most importantly, when they need to rest and recover. Moreover, managing and preventing injuries is crucial for a team's stability and continuity which is linked to success (Raysmith & Drew, 2016; Podlog et al., 2015; Eirale et al., 2013; Williams et al., 2016).

One of the most common uses of ML for addressing the aforementioned tasks is developing novel indicators for quantifying performances. Typically, machine-learned models are trained on large historical databases of past matches. Afterwards, the indicator is derived from the model, as the indicator cannot be used directly as a target for training because it is not in the data,[1] One prominent example of such an indicator is expected goals (xG) (Green, 2012), which is used in soccer and ice hockey to quantify the quality of the scoring opportunities that a team or player created. The underlying model is a binary classifier that predicts the outcome of a shot based on features such as the distance and angle to the goal, the assist type and the passage of play (Fig. 1). xG is typically a more consistent measure of performance than actual goals, which are extremely important in these sports but also very rare. Even shots are relatively infrequent, and their conversion is subject to

[1] This is related to self-supervision (Balestriero et al., 2023) where models are initially trained on an auxiliary or pretext classification task using pseudo-labels which help to initialize the model parameters.

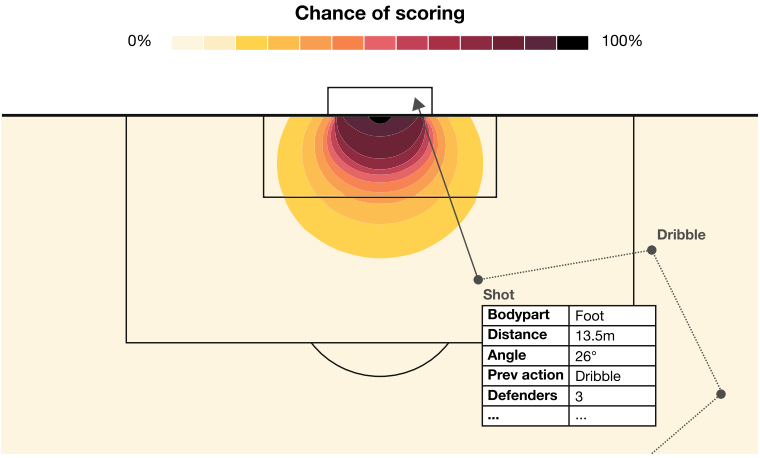

Fig. 1 The expected goals (xG) metric is based on a classification model that predicts the probability of scoring a goal from a particular shot. The colors show how moving closer to the goal would increase the probability

variance. The idea of xG is to separate the ability to get into good scoring positions from the inherent randomness (e.g., deflections) of converting them into goals.

Typically, an indicator should satisfy several properties. First, it should provide insights that are not currently available (Franks et al., 2016). For example, xG tells us something beyond looking at goals scored, namely how often a player gets into good shooting positions. Second, the indicator should be based on domain knowledge and concepts from sports such that it is intuitive and easy for non-ML experts to understand. Finally, the domain experts need to trust the indicator. This often boils down to being able to contextualize when the indicator is useful and ensuring some level of robustness in its value (i.e., it should not wildly fluctuate).

These desiderata illustrate that a key challenge in developing indicators is how to evaluate them: none of the desiderata naturally align with the standard performance metrics used to evaluate learned models. This does not imply that standard evaluation metrics are not important. In particular, ensuring that probability estimates are well-calibrated is crucial in many sports analytics tasks. It is simply that one must both evaluate the indicator itself and the models used to compute the indicator's value.

The importance of AI and ML within sports has led to much closer collaborations between data scientists and practitioners with backgrounds in sports science, video analysis, and scouting, many of whom are also seeking to enhance their own data science skills. This paper aims to provide a structured framework with practical examples to facilitate better understanding and collaboration among individuals from diverse backgrounds who are working on sports data. By bridging the gap between traditional sports expertise and modern data analytics, we hope to 'assist' more effective interdisciplinary partnerships and advancements in the field. We will focus on methodological and evaluation issues because they are neither thoroughly discussed in the literature nor extensively supported in common tools despite the fact that they consume substantially more time than, e.g., setting up training models, in the data science pipeline (Munson, 2011).

Concretely, our goal is four-fold. First, we will highlight some challenges that arise when analyzing sports data. Second, we discuss some methodological considerations that

arise due to the nature of sports data and the tasks considered that affect how an evaluation should be executed. Third, we will discuss the various ways that indicator evaluation has been approached. Fourth, we will overview how the learned models that the indicators rely upon have been evaluated. While we will discuss standard evaluation metrics, we will argue that a broader perspective on evaluation is warranted that also considers aspects such as explainability and quantifying the uncertainty in a model's predictions. Moreover, we believe it is essential to employ reasoning techniques to understand how the model may perform when confronted with instances that are not in the training, validation or testing data. This paper focuses on the context of professional soccer, where we have substantial experience. However, we believe the lessons and insights are applicable to other team sports, or other domains than sports.

2 Common sports data and analytics tasks

This section serves as a short, high-level primer on the data collected from sports matches as well as typical styles of performance indicators and tactical analyses.

2.1 Data

While there are a variety of sources of data collected about sports, we will discuss four broad categories: physical data, match sheet data, play-by-play data and (optical) tracking data. Such data is available for a variety of sports (e.g., American football, basketball, and ice hockey), but for consistency the visual examples will come from soccer.

During training and matches, athletes often wear a GNSS tracker with inertial measurement unit (IMU) technology (e.g., from Catapult Sports). These systems measure various physical parameters such as distance covered, number of high-speed sprints, and high-intensity accelerations. These parameters are often augmented with questionnaire data (Buchheit et al., 2016) to obtain subjective measurements such as the rating of perceived exertion (RPE) (Borg, 1982) for a training session. Such approaches are used to monitor and eventually optimize an athlete's fitness level to ensure their availability and ability to compete.

Match sheet data or box scores are the most common and widely available source of data. These include aggregate statistics for the team and/or the players on a per-match basis. For example, in soccer this could include information such as line-ups, substitutions, cards, goals and assists. More advanced details such as the number of passes attempted and completed per player or team may also be included. Moreover, for a number of sports, this data has been collected for decades.

Play-by-play or event stream data tracks actions that occur with the ball. Each such action is annotated with information such as the type of the action, the start and end locations of the action, the result of the action (i.e., successful or not), the time at which the action was performed, the player who performed the action, and the team of which the acting player is a part of. Figure 2 illustrates six actions that are part of the game between Brazil and Belgium at the 2018 World Cup as they were recorded in the event stream data format. This data is collected for a variety of sports by vendors such as StatsBomb[2] and Stats

[2] https://www.statsbomb.com

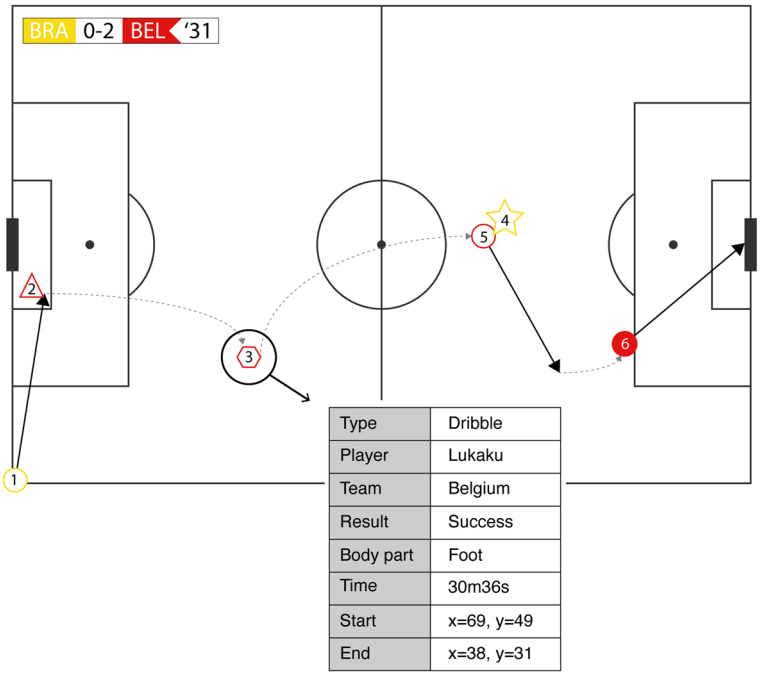

Fig. 2 The sequence of actions leading up to Belgium's second goal during the 2018 World Cup quarter-final. Each on-the-ball action is annotated with a couple of attributes, as illustrated for Lukaku's dribble. *(Data source: StatsBomb)*

Perform[3] who typically employ human annotators to collect the data by watching broadcast video (Bialik, 2014a).

Optical tracking data reports the locations of all the players and the ball multiple times per second (typically between 10 and 25 Hz). This data is collected using a fixed installation in a team's stadium using high-resolution cameras. Such a setup is expensive and typically only used in top leagues. There is now also extensive work on tracking solutions based on broadcast video (Johnson, 2020; Arbués Sangüesa, 2021). Figure 3 shows a short sequence of broadcast tracking data.

2.2 Physical indicators

Ultimately, the goal of monitoring physical indicators is to provide insight into the physical status of an athlete. These are on the one hand used for optimizing physical fitness and on the other hand reducing/predicting injury risk.

Optimizing fitness is also sometimes referred to as load management. Typically, research distinguishes between the external load, which is the amount and type of activities an athlete performs, and the internal load, which is the body's physiological response to the activity (Impellizzeri et al., 2005). The external load is typically measured using sensor systems such as GNSS, accelerometers and local positioning systems. As such there has

[3] https://www.statsperform.com

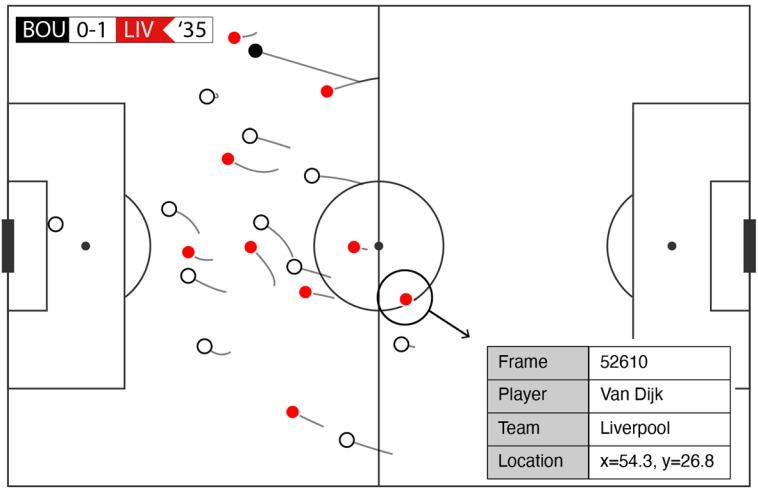

Fig. 3 Illustration of broadcast tracking data for the first goal of Liverpool against Bournemouth on Dec 7, 2019. The black lines represent each player's and the ball's trajectories during a period of 1.5 s. Each trajectory contains the position, a team identifier and a player identifier at every time step. *(Data source: Last Row)*

been substantial computational work (i.e., signal processing) to design indicators from raw data. From a sports science perspective, there is also substantial interest in understanding the link between internal and external load because this can help to evaluate whether someone is getting fitter or experiencing problems, and possibly has an elevated injury risk. A number of studies have relied on machine learning for exploring this question (Bartlett et al., 2017; Jaspers et al., 2018). More generally, load management is important because if you can predict the impact of load on physical fitness it can help to periodize the training load (i.e., prescribe more or less load on certain days) (Jeffries et al., 2022; Vanrenterghem et al., 2017).

One prominent task is to learn models to assess injury risk (Rossi et al., 2018; de Leeuw et al., 2022; Bornn et al., 2019). These methods focus on predicting the probability of a near-term injury based on an athlete's recent training regime and possibly, individual characteristics (Windt & Gabbett, 2017). They typically focus on overuse injuries because an acute trauma (e.g., bone fracture or concussion) often arises from unforeseeable factors (e.g., how and when an athlete gets hit or tackled).

More generally, learned models are often used as proxies for physical parameters that are difficult to collect. For example, athletes in sports such as basketball and running often undergo biomechanical analyses, which require extensive lab setups (e.g., motion capture systems and force plates). Therefore, there is interest in performing such analyses in the wild using cheaper and more portable sensors such as inertial measurement units (Dorschky et al., 2023). Similarly, VO_2Max serves as a gold standard for aerobic fitness, which is essential for endurance sports. However, the tests for measuring it are arduous and performing them can disrupt an athlete's training such that it cannot be measured often. Thus, there is work on trying to predict this value based on less strenuous activities (de Leeuw et al., 2023; De Brabandere et al., 2018). Finally, teams may want to estimate parameters (e.g., total distance, sprint distance, number of accelerations) from players

on other teams, and there is now interest in trying to derive these from broadcast tracking data (Mortensen & Bornn, 2020).

2.3 Individual performance indicators

Quantifying a player's proficiency across a variety of skills is a central problem in player recruitment. Data-driven performance indicators can be helpful in this process to construct more comprehensive player profiles. These performance indicators can be directly derived from the data or estimated through machine-learned models (Kovalchik, 2023).

Derived indicators are typically based on counting how often a player performs a certain type of action (e.g., passes) or aggregating over action attributes (e.g., average pass length) (Quang Nguyen & Matthews, 2024; Llana et al., 2022). For example, a count of the number of passes that bring the ball closer to the opposition goal (i.e., *progressive passes*) allows the separation of players whose passes are mostly safe and cautious from players whose passes disproportionately drive the ball forward towards the opposition goal. A limitation of these derived indicators is that they generally cannot take the context in which actions are executed into account. This means that two progressive passes are considered equivalent regardless of the number of defenders they overtake. Model-based indicators, on the other hand, can learn from historical data how different contexts can impact the effectiveness and value of specific actions.

Model-based indicators can be further divided into two sub-categories. The first type focuses on a single action such as a pass or shot. The second type takes a holistic approach by developing a unifying framework that can value a wide range of action types.

Single action. Single action indicators typically take the form of expected value-based statistics: they measure the expected chance that a typical player would successfully execute the considered action in a specific game context. For example, the aforementioned xG model in soccer assigns a probability to each shot that represents its chance of directly resulting in a goal. These models are learned using standard probabilistic classifiers such as logistic regression or tree ensembles from large historical datasets of shots. Each shot is described by the game context from when it was taken, and how this is represented is the key difference among existing models (Green, 2012; Lucey et al., 2015; Robberechts & Davis, 2020).

Such indicators exist for a variety of sports including American football (e.g., expected completion percentage for quarterbacks and expected yards after the catch for receivers),[4] basketball (e.g., expected field goal percentage (Sarlis & Tjortjis, 2020)), and ice hockey (e.g., expected goals (Macdonald, 2012)).

All actions. Instead of building bespoke models for each action, these indicators use the same framework to aggregate a player's contributions over a set of action types. Regardless of sport, almost all approaches exploit the fact that each action a_i changes the game state from s_i to s_{i+1} (as illustrated in Fig. 4). These approaches value the contribution of an action a_i as:

$$C(s_i, a_i) = V(s_{i+1}) - V(s_i), \tag{1}$$

where $V(.)$ is the value of a game state and s_{i+1} is the game state that results from executing action a_i in game state s_i.

[4] See https://nextgenstats.nfl.com/glossary for the definitions of these indicators.

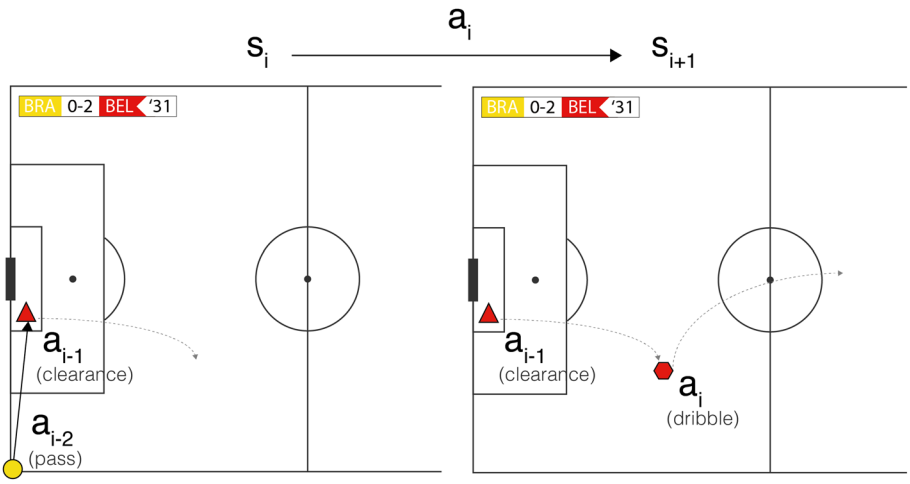

Fig. 4 Lukaku's dribble (a_i) changes the game state from the pre-action state s_i to the post-action state s_{i+1}

Approaches differ on how they value game states, with two dominant paradigms emerging: scoring-based and win-based. Scoring-based approaches take a narrower possession-based view. These approaches value a game state by estimating the probability that the team possessing the ball will score. In soccer, this may entail looking at the near-term probability of a goal in the next 10 actions or 10 s (Decroos et al., 2019) or the long-term probability of scoring the next goal (Fernández et al., 2021). Win-based approaches look at valuing actions by assessing a team's chance of winning the match in each game state. That is, these approaches look at the difference in in-game win-probability between two consecutive game states (Pettigrew, 2015; Burke, 2010; Robberechts et al., 2021; Bouey, 2013). Such models have been developed for many sports, including basketball (Cervone et al., 2014), American football (Romer, 2006), ice hockey (Routley & Schulte, 2015; Liu & Schulte, 2018) and rugby (Kempton et al., 2016).

2.4 Tactical analyses

Tactics are short-term behaviors that are used to achieve a strategic objective such as winning or scoring a goal. Identifying and understanding the strengths and weaknesses of an opponent's tactics is a key problem in match preparation. Moreover, in player recruitment it is important for a team to understand if a player's skill set fits within the team's tactical blueprint. At a high level, AI/ML is used for tactical analyses in two ways: to discover patterns and to evaluate the efficacy of a tactic.

Discovering patterns is a broad task that may range from simply trying to understand where on the field certain players tend to operate and who tends to pass to whom, to more complicated analyses that involve identifying sequences of reoccurring actions. Typically, techniques such as clustering, non-negative matrix factorization, and pattern mining are used to find such reoccurring behaviors (Wang et al., 2015; Decroos & Davis, 2019; Bekkers & Dabadghao, 2019; Anzer et al., 2022; Andrienko et al., 2019; Miller & Bornn, 2017).

Evaluating the efficacy of tactics is an equally broad task that can generally be split up into two parts: evaluating the efficacy of (1) a current and (2) a counterfactual tactic.

Assessing the efficacy of currently employed tactics is typically done by focusing on a specific tactic (e.g., counterattack, pressing) and relating it to other success indicators (e.g., goals, wins) (Fernandez-Navarro et al., 2019; Merckx et al., 2021). In contrast, assessing the efficacy of counterfactual tactics is more challenging as it entails understanding what would happen *if* a team (or player) employed different tactics than those that were observed. This is extremely interesting and challenging from an AI/ML and evaluation perspective as it involves both (1) accurately modeling the current behavior of teams, and (2) reasoning in a counterfactual way about alternative behaviors. Such approaches have been developed in basketball and soccer to assess counterfactual shot (Sandholtz & Bornn, 2020; Van Roy et al., 2021, 2023) and movement (Le et al., 2017; Lowe, 2013) tactics. Moreover, work in soccer has also considered causal analyses of throw-ins (Epasinghege Dona & Swartz, 2024) and crossing (Wu et al., 2021) tactics.

3 Challenges with evaluation

The nature of sports data and the tasks typically considered within sports analytics and science pose many challenges from an evaluation and analysis perspective.

3.1 Lack of ground truth

For many variables of interest, there are simply very few or even no labels, which arises when analyzing both match and physical data. When analyzing matches, a team's specific tactical plan is unknown to outside observers. One can make educated guesses on a high level, but often not for fine-grained decisions. Similarly, when trying to assign ratings to players' actions in a soccer match, there is no variable that directly records this. In fact, in this case, no such objective rating even exists.

Physical parameters can also be difficult to collect. For example, if one is interested in measuring fatigue[5] during a match or training session, some measures are invasive (e.g., blood lactate or creatine kinase). Similarly, in endurance sports such as distance running and cycling, monitoring athletes' aerobic fitness levels is important, which is often measured in terms of the maximal oxygen uptake (VO_2Max) (Joyner, 1991). However, the test to measure this variable is extremely strenuous and disrupts training regimes, so it can only be measured sporadically.

3.2 Credit assignment

It is often unclear why an action succeeded or failed. For example, did a pass not reach a teammate because the passer mishit the ball or did their teammate simply make the wrong run? Similarly, for those actions that are observed, we are unsure why they arose. For example, does a player make a lot of tackles in a soccer match because they are covering for a teammate who is constantly out of position? Or is the player a weak defender that is being targeted by the opposing team?

[5] Note that there are different types of fatigue that could be monitored such as musculoskeletal or cardiovascular fatigue.

3.3 Noisy features and/or labels

When monitoring the health status of players, teams often partially rely on questionnaires (Buchheit et al., 2016) and subjective measures like the rating of perceived exertion (Borg, 1982). Players respond to such questionnaires in different ways, with some being more honest than others. There is a risk for deception (e.g., players want to play, and may downplay injuries). There are also well-known challenges when working with subjective data. Similarly, play-by-play data is often collected by human annotators, who make mistakes and they might not always be fully aligned on the definitions for certain types of events or qualifiers (Bialik, 2014b). For example, crosses that (accidentally) result in a goal are often labeled as shots, and vice versa, errant shots can be labeled as crosses. Also, some annotations are inherently subjective. For instance, Opta's "big chance" and Wyscout's "smart pass" qualifiers. Moreover, in some cases they suffer from biases because scorekeepers often are employed by a team (van Bommel & Bornn, 2017). The definitions of events and actions can also change over time. Finally, data quality issues are also present in (optical) tracking data. For example, occlusions of players and low-resolution camera set-ups make it difficult to accurately estimate the locations of the players and the ball. This results in errors in the tracking data, which are not only problematic in itself, but are also propagated to other metrics (e.g., speed and acceleration (Wu & Swartz, 2023) and further derived indicators) that are calculated based on this data.

3.4 Small sample sizes

There may only be limited data about teams and players. For example, a top flight soccer team plays between 34 and 38 league games in a season and will perform between 1500 and 3000 on-the-ball actions in a game.[6] Even top players do not appear every game and sit out matches strategically for rest.

3.5 Non-stationary data

The sample size issues are compounded by the fact that sports is a very non-stationary setting, meaning data that is more than one or two seasons old may not be relevant. On a team level, playing styles tend to vary over time due to changes in playing and management personnel. On a player level, skills evolve over time, often improving until a player reaches their peak, prior to an age-related decline. More generally, tactics evolve and change.

3.6 Counterfactuals

Many evaluation questions in sports involve reasoning about outcomes that did not occur. This is most notable in the case of defense, where defensive tactics are often aimed at preventing dangerous actions from arising such as wide-open three-point shots in the NBA or one vs. the goalie in soccer. Unfortunately, it is hard to know why certain actions were or were not taken. For example, it is difficult to estimate whether the goalie would have saved the shot if they had been positioned slightly differently Similarly, evaluating tactics also

[6] The number depends on what is annotated in the data (e.g., pressure events) and modeling choices such as whether a pass receival is treated as a separate action.

involves counterfactual reasoning as a coach is often interested in knowing what would have happened if another policy had been followed, such as shooting more or less often from outside the penalty box in soccer.

3.7 Interventions

The data is observational and teams constantly make decisions that affect what is observed. This is particularly true for injury risk assessment and load management, where the staff will alter players' training regimes or keep them out of a match if they are worried about the risk of injury. This would affect any collected indicators about a player's health status. This also has a knock-on effect of creating ambiguous labels. If a player is too injured, then there is a clear label. However, if a player does not play, we do not know whether the player was healthy (e.g., planned rotation) or held out because the staff believes they are close to injury and whether this assessment was justified or not. Managers also change tactics during the course of the game, depending on the score and the team's performance. For example, in soccer, a manager may substitute a forward for a defender when a team is defending a lead but would switch a defender for a forward when the team is trailing.

4 Methodological considerations

From a methodological perspective, sports data present a number of challenges that if not properly accounted for will affect the obtained results. These include how the data is partitioned in the evaluation, how biases in the data are accounted for, and how contextual factors are considered.

4.1 Data partitioning considerations

The bedrock of empirical evaluation in machine learning lies in the use of an independent test set to assess a model's ability to generalize to unseen data. Typically, this is done by a cross-validation scheme where the data are randomly partitioned into a number of folds. However, this standard approach is not necessarily well-suited to the data or tasks that characterize sports analytics tasks.

On the data level, the problem is that sports data are often not independent and identically distributed. Consequently, the standard approach of randomly partitioning examples is not applicable as is. Two notable cases are temporal dependencies and subject consistency.

Temporal dependencies. Sports are inherently temporal. On the physical side, training sessions are ordered: each individual session is planned based both on activities that happened in the (recent) past (e.g., recovery after a match) and those that are planned in the near future (e.g., light workout before a match). Similarly, seasons, matches within a season, and events within a match are all ordered. In temporal data, the data does not change enough from one measurement to the next, introducing strong correlations.

Subject consistency. The same subject will act similarly within and across measurements. For example, when looking at segments of running data,

the parameters extracted from segments will be correlated over one run. Moreover, a runner's parameters will also be correlated across multiple different runs.

In both cases, an incorrect partitioning of data would mean that we cannot be sure whether we are predicting the intended target, or simply the strong proxy that exists (i.e., predict the same value as in the previous time segment or the previous value observed for the same person). For both, we will have to alter the cross-validation scheme to take groups of dependent instances into account.

Coping with temporal dependencies requires considering temporal groupings when partitioning the data. Consequently, the cross-validation scheme must involve a partition into non-overlapping periods (Bergmeir et al., 2018). One approach is to employ a single split, which could be in the middle of the season (e.g., Rossi et al. (2018)) or between seasons (e.g., Jaspers et al. (2018)). One complication is that seasons may start and end at different times for different leagues (e.g., many European soccer leagues run from August to May whereas the US league runs from February to November. Another alternative is to employ a walk-forward cross-validation scheme. In this setup, there is an initial temporal split and predictions are made for a small test set consisting of the examples after the split point. Then, the temporal split point is moved forward. Thus, the previous test data is added to the training set, the model is retrained, and predictions are made for a new test set.

Coping with subject consistency means that we might want to force all data from one subject to be either in the training set or the testing set but not both. Therefore, the partitioning of data into folds needs to happen on the subject level and not the example level. This is sometimes called subject-level cross-validation (Dehghani et al., 2019).

On the task level, a practical issue that arises with sports models is understanding how applicable they are to new settings where we have no available data such as a different competition, or a different gender. Concretely, it may be important to answer questions such as:

- Will a running fatigue prediction model be applicable to a new athlete (Op De Beéck et al., 2018)?
- Will a model for predicting the expected length of a ski jump make accurate predictions for jumps made on a different hill (Link et al., 2022)?
- Will an expected goals model for soccer trained on data from men's leagues be applicable to a women's league (Bransen & Davis, 2021)?

Addressing such questions requires carefully partitioning the data. If the goal is to generalize to new athletes, then the aforementioned subject-level cross-validation should be employed. Similarly, if the goal is to generalize to new venues, then all data collected from a venue should appear in the same fold (i.e., a venue-level partitioning). When looking at generalizing across different competitions, then a leave-one-competition-out scheme would be appropriate.

Finally, while there is agreement that sports data is not stationary, there has been little work on investigating if data goes out of date (e.g., Robberechts and Davis (2020)). That is, when should models be retrained and what data, if any, should be excluded from training. As more data becomes available, it will be more important to perform (public) studies to answer these questions.

4.2 Data biases

When looking at technical performance metrics derived from data collected about matches, there are a number of (subtle) biases in the data that can affect the model's performance. Namely, good teams and good players tend to perform more actions. For example, exceptional shooters in basketball (e.g., Stephen Curry, Damian Lillard) take more deep three-point shots (Goldsberry, 2019) and three-point shots in general than weaker shooters. Similarly, in soccer, good finishers are likely to take more shots. Hence, models trained to predict the expected three-point percentage in basketball or expected goals in soccer will be trained from data that contain correlated observations (i.e., multiple shots from the same player) and the sample sizes per player can vary widely. This is not necessarily problematic if the objective is to obtain a well-calibrated probability estimate. However, another use of these metrics is to measure skill. Typically, this is done by comparing the true outcomes to the expected outcomes predicted by the model. In the case of soccer, there is evidence that in this use case there is a small but persistent bias that makes the actual and expected goals closer for excellent finishers than it really is (Davis and Robberechts, 2023). That is, these metrics may underestimate the skill of excellent finishers. A classic way to cope with the issue of repeated measurements and interdependencies is to consider hierarchical or multilevel models (Tureen & Olthof, 2022; Yurko et al., 2019). These types of models allow for accounting for group structure (e.g., position groups, teams) as well as repeated measurements about individuals. Another idea is to use a post-processing technique such as multicalibration (Hebert-Johnson et al., 2018), which attempts to calibrate a model's predictions for subpopulations that may be defined by complex intersections of many attributes. This introduces a grouping effect that enables explicitly representing various player types based on their characteristics within a single model (Davis & Robberechts, 2024).

Similarly, good teams may behave differently. For example, in soccer this manifests itself in terms of a stronger team's ability to retain possession of the ball for longer (StatsBomb, 2021). This could raise issues when training action-value models because they rely on constructing features that capture the game state. Many of these models hand-craft features that make use of information about prior actions (e.g., Decroos et al. (2019)), which could introduce correlations with team strength. These correlations may result in over- or undervaluing certain actions.

4.3 Accounting for contextual factors

Contextual factors are important on two different levels. First, they affect the features that are used as input for the learned models that form the basis of indicators like expected goals. Second, they are important to consider when proposing the final indicator that is built on top of the learned model's predictions such as an action-value metric.

4.3.1 Contextualizing feature values

While it is well known that standard preprocessing techniques such as standardization or min-max normalization are often important to apply prior to learning, there are several subtler transformations that can be useful to apply in the context of sports. Questionnaire data collected as part of physical monitoring often involves subjective measurements, which people use differently. In these situations, it is often beneficial to model deviations from a

baseline value such as a user's average rating or in the case of repeated measures during an activity to the first reported value (Op De Beéck et al., 2018). This is analogous to issues that arose in the Netflix challenge, where an important contribution was the importance of isolating the user-movie interaction component of a rating by deriving a new variable that controlled for factors related to the user, the movie, and time (Koren, 2008, 2010).

Finally, it can be important to have enough contextual variables to discern what is important. This requires adapting a feature to the value of another feature. There are a surprising array of examples as to where this may be necessary.

- In continuous monitoring of a physical parameter, it is often useful to consider deviations from a feature's value, e.g., the beginning of a training session or test (e.g., Op De Beéck et al. (2018)).
- The cumulative training load in a month is influenced by the number of days in the month (Hyndman & Athanasopoulos, 2023).
- Sizes of fields can vary in sports such as soccer and baseball, and considering these effects can be important in certain use cases. For example, in soccer, any metric that involves a spatial component such as pitch control (a bigger pitch means that the total possible pitch control is also higher) will be affected (Fernandez & Bornn, 2018). Similarly, the number of attempted crosses tends to be higher on wider pitches than on narrower pitches. The changes in dimensions also affect definitions of common zones such as the offensive or final third.
- A team's tactical approach such as whether a soccer team plays with a back three or back four can also affect physical indicators (Fifield, 2022; Modric et al., 2022; Baptista et al., 2019).
- Physical parameters are also affected by environmental variables. For example, outside temperature is relevant when considering body measurements like temperature and altitude difference is important to assess heart rate when running. Typically, such information needs to come from domain knowledge.
- Players' physical and technical indicators can also be affected by whether they are playing home or away as well as by the condition of the field that they are playing on (Chmura et al., 2021; Gollan et al., 2020).

4.3.2 Contextualizing an indicator

There is a general agreement in sports analytics that counting statistics ignore a number of important contextual factors such as the amount of playing time a player receives or the pace of the game. This is typically addressed by normalizing an indicator. Because metrics typically rely on aggregating predictions in some way, it is important to think about the most appropriate way to do this as it will affect the evaluation of the metric. This can be done in several ways.

Per action: An indicator can be averaged per action or action type. This can help overcome issues such as certain players performing more or less actions where summing can skew results (e.g., midfielders perform more actions than other players in soccer). However, this can lead to sample size issues, meaning that including some notion of variance is informative.

Per minutes played: Often metrics are prorated to a set amount of time. For example, in soccer this is often per 90 min of play (e.g., Decroos et al. (2019)). However, in soccer the length of the game varies due to the referee's discretionary added time at the end of each

half. Moreover, the amount of time the ball is in play varies from match to match. Therefore some approaches normalize metrics based on ball-in-play time (Llana et al., 2022).[7]

Per possession: This approach reports an average per possession (Oliver, 2002). However, in many sports (e.g., basketball, soccer) possessions may not be recorded directly in the data. Hence, they must be derived using some heuristics (Phatak et al., 2022).

At a higher level, there are also game state, opponent, and seasonal effects that may need to be accounted for. First, there can be time-varying effects within a match. For example, the goal scoring rates increase throughout the course of a match in soccer (Bransen & Davis, 2021). Similarly, substitutes in soccer tend to generate more value (i.e., as determined by an action-value model) than starters.

Second, there can also be score difference effects. When the score is lopsided, teams or players may not play as hard; or younger, less experienced players may play. Similarly, actions carry more impact if they can influence the result of a game. These effects can be captured by using win probability-based metrics (e.g., Robberechts et al. (2021), Pettigrew (2015)) or by contextualizing performances (e.g., by pressure level (Bransen et al., 2019)).

Third, the strength of an opponent clearly will affect player and team performance. Therefore, it can be useful to include opponent-based adjustments. For example, Pelechrinis et al. (2018) set up an optimization problem to account for an opponent's defensive strength when evaluating expected points added for American football.

Finally, there can also be effects due to where a match falls in a season. Towards the end of the season, if a team is out of contention, players may be strategically rested (e.g., to improve draft odds, to let younger players gain experience). Moreover, simply having nothing to play for may affect players (Bransen et al., 2019). Additionally, injuries and fatigue tend to accumulate over the season, hence a team's and possibly even the overall strength of a competition may vary. For example, in soccer, the goal scoring rates also tend to slightly increase over the course of a season (Bransen & Davis, 2021).

5 Evaluating an indicator

A novel indicator should capture something about a player's (or team's) technical performance or capabilities during a match. This objective does not align with typical metrics considered in machine learning. Moreover, while an indicator is based on a learned model, how the model's output is used in the indicator is not necessarily the same as the target variable that the model was trained to predict. Consequently, evaluating a novel indicator's usefulness is difficult as it is unclear what it should be compared against. This problem is addressed in multiple different ways in the literature.

5.1 Correlation with existing success indicators

In all sports, a variety of indicators exist that denote whether a player (or team) is considered or perceived to be good.

When evaluating individual players, there are a wealth of existing indicators that are commonly reported and used. First, there are indirect indicators such as a player's market value, salary, playing time, or draft position. Second, there are indicators derived from competitions such as goals and assists in soccer (or ice hockey). It is therefore possible to

[7] The company SkillCorner normalizes some physical metrics per 60 min of ball-in-play time.

design an evaluation by looking at the correlation between each indicator's value for every player (Pettigrew, 2015; Routley & Schulte, 2015; Liu & Schulte, 2018). Alternatively, it is possible to produce two rank-ordered lists of players (or teams): one based on an existing success indicator and another based on a newly designed indicator. Then the correlation between rankings can be computed.

Arguably, an evaluation that strives for high correlations with existing indicators misses the point: the goal is to design indicators that provide insights that current ones do not. If a new indicator simply yields the same ranking as looking at goals, then it does not provide any new information. Moreover, some existing success indicators capture information that is not related to performance. For example, salary can be tied to draft position and years of service. Similarly, a soccer player's market value or transfer fee also encompasses their commercial appeal, and more importantly, the duration of their current contract (McHale & Holmes, 2023). Even playing time is not necessarily merit-based.

Other work tries to associate performance and/or presence in the game with winning. This is appealing as the ultimate goal is to win a game.[8] For example, indicators can be based on correlating how often certain actions are performed with match outcomes, points scored, or score differentials (McHale & Scarf, 2007; McHale et al., 2012; Pappalardo et al., 2019). An alternative approach is to build a predictive model based on the indicators and see if it can be used to predict the outcomes of future matches (Hvattum, 2020).

5.2 The face validity test

When evaluating indicators about player performance, one advantage is that there is typically consensus on who are among the very top players. While experts, pundits, and fans may debate the relative merits of Lionel Messi and Cristiano Ronaldo, there is little debate that they fall in the very top set of offensive players. An offensive metric where neither of those players scores well, is not likely to convince practitioners. In other words, if a metric blatantly contradicts most conventional wisdom, there is likely a problem with it. This is also called face validity (Jacobs & Wallach, 2021). Of course, some unexpected or more surprising names could appear towards the top of such a ranking, but one would be wary if all such names were surprising.

Unfortunately, this style of evaluation is most suited to analyzing offensive contributions. In general, there is more consensus on the offensive performances of individual players than their defensive performances, as good defense is a collective endeavor and more heavily reliant on tactics.

5.3 Make a prediction

While backtesting indicators (and models) is clearly a key component of development, sports does offer the possibility for real-world predictions on unseen data. One can predict, and most importantly publish, the outcomes of matches or tournaments prior to their start. In fact, there have been several competitions designed around this principle (Dubitzky et al., 2019) or people who have collected predictions online.[9]

[8] This is not always the case: sometimes teams play for draws, rest players for strategic reasons, prioritize getting young players experience or try to lose to improve draft position.

[9] https://twitter.com/TonyElHabr/status/1414619621659971588

This is even possible for player indicators, and is often done in the work on quantifying player performance (Decroos et al., 2019; Liu & Schulte, 2018). Decroos et al. (2019) included lists of the top under-21 players in several of the major European soccer leagues for the 2017/2018 season. It is interesting to look back on the list, and see that there were both hits and misses. For example, the list had some players who were less heralded then such as Mason Mount and Mikel Oyarzabal, who are now key players. Similarly, it had several recognized prospects such as Kylian Mbappé, Trent Alexander-Arnold, and Frenkie de Jong who have ascended. Finally, there were misses like Jonjoe Kenny and David Neres. While one has to wait, it does give an immutable forecast that can be evaluated.

Because they do not allow for immediate results, such evaluations tend to be done infrequently. However, we believe this is something that should be done more often. It avoids the possibility of cherry-picking results and overfitting by forcing one to commit to a result with an unknown outcome. This may also encourage more critical thinking about the utility of the developed indicator. The caveat is that the predictions must be revisited and discussed in the future, which also implies that publication venues would be open to such submissions. Beyond the time delay, another drawback is that they involve sample sizes such as one match day, one tournament, or a short list of players.

5.4 Ask an expert

Developed indicators and approaches can be validated by comparing them to an external source provided by domain experts. This goes beyond the face validity test as it requires both deeper domain expertise and a more extensive evaluation such as comparing tactical patterns discovered by an automated system to those found by a video analyst. Pappalardo et al. (2019) compared a player ranking system they developed to rankings produced by scouts. Similarly, Dick et al. (2022) asked soccer coaches to rate how available players were to receive a pass in game situations and compared this assessment to a novel indicator they developed.

Ideally, such an expert-based evaluation considers aspects beyond model accuracy. Ultimately, an indicator should provide "value" to the workflow of practitioners. Hence, it is relevant to measure how much time it saves an analyst in his workflow, whether an indicator can provide relevant new insights and whether the expert can correctly interpret the model's output. This type of evaluation checks whether indicators fulfill the needs of users (i.e., usefulness and usability) and also arises in human-computer interaction (Xu, 2019).

However, this type of evaluation can be difficult as not all researchers have access to domain experts, particularly when it comes to high-level sports. Moreover, teams want to maintain a competitive advantage, so one may not be able to publish such an evaluation.

5.5 Reliability

Indicators are typically developed to measure a skill or capability such as shooting ability in basketball or offensive contributions. While these skills can and do change over a longer timeframe (multiple seasons), they typically are consistent within a season or even across two consecutive seasons. Therefore, an indicator should take on similar values in such a time frame.

One approach (Baron et al., 2024; Hvattum, 2020; Van Roy et al., 2020) to measure an indicator's reliability is to split the data set into two, and then compute the correlation between the indicators computed on each dataset. An example of such an

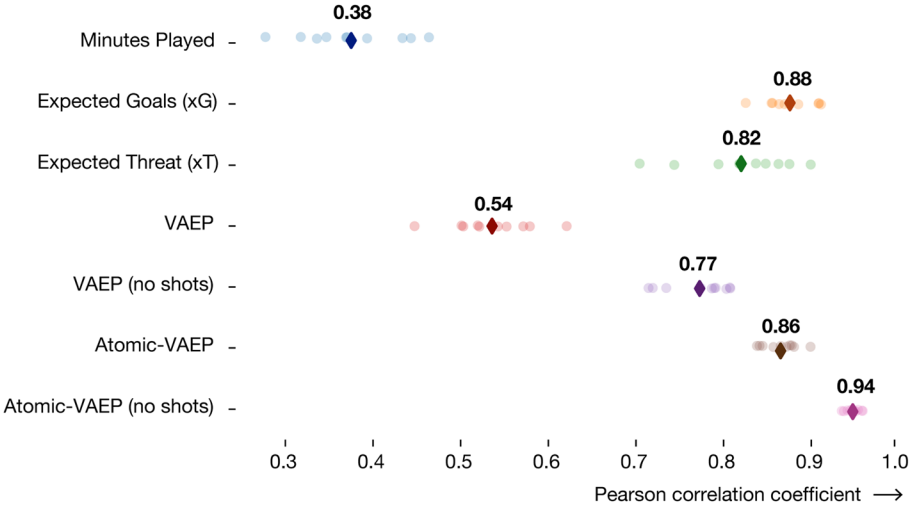

Fig. 5 Pearson correlation between player performance indicators for ten pairs of successive seasons in the English Premier League (2009/10–2019/20). The diamond shape indicates the mean correlation. The simple "minutes played" indicator is the least reliable, while the Atomic-VAEP (Decroos, 2020) indicator is more reliable than its VAEP (Decroos et al., 2019) predecessor and xT (Singh, 2019). As shots are infrequent and have a variable outcome, omitting them increases an indicator's reliability. The xT indicator does not value shots. Only players that played at least 900 min (the equivalent of ten games) in each of the successive seasons are included

evaluation is shown in Fig. 5. Methodologically, one consideration is how to partition the available data. Typically, one is concerned with respecting chronological orderings in temporal data. However, in this setting, such a division is likely suboptimal. First, games missed by injury will be clustered and players likely perform differently right when they come back. Second, the difficulty of a team's schedule is not uniformly spread over a season. Third, if the time horizon is long enough, there will be aging effects.

Franks et al. (2016) propose a metric to capture an indicator's stability. It tries to assess how much an indicator's value depends on context (e.g., a team's tactical system, quality of teammates) and changes in skill (e.g., improvement through practice). It does so by looking at the variance of indicators using a within-season bootstrap procedure.

Another approach (Decroos & Davis, 2019) is to look at consecutive seasons and pose the evaluation as a nearest neighbors problem. That is, based on the indicators computed from one season of data for a specific player, find a rank-ordered list of the most similar players in the subsequent (or preceding) season. The robustness of the indicator is then related to the position of the chosen player in the ranking.

6 Evaluating a model

Indicators are typically constructed based on the predictions provided by machine-learned models. Thus, the quality of the underlying models directly affects the meaningfulness and reliability of an indicator. Evaluating the models used to produce the indicator is a multi-faceted endeavor. On the one hand, it is necessary to employ standard evaluation metrics to ensure that the model is sufficiently performant. Indeed, many tasks in sports rely on having sufficiently well-calibrated probability estimates (Silva Filho et al., 2023). On the other hand, standard metrics are insufficient to fully characterize all relevant aspects of a model's performance. For example, being able to explain or interpret how a model behaves and identify cases when its predictions may not be trustworthy are also important aspects of evaluation. However, this is difficult to capture with metrics that just assess the aggregate correctness of a model's predictions. Moreover, techniques such as verification allow detecting issues in the model that standard evaluation metrics fail to identify. This is because verification can reason about how the model will perform in all situations and not just those that arise in training, validation or testing data. For example, a labeling bias in the data (also present in the testing data) may surface when reasoning about the behavior of the model in scenarios for which experts have strong intuitions based on domain knowledge.

6.1 Standard evaluation metrics

Different metrics are used to assess models depending on whether the task is classification (e.g., predicting whether a shot will be scored in soccer) or regression (e.g., predicting rushing yards in NFL). Classification metrics evaluate how well a model distinguishes between different classes, focusing on the accuracy of the classification but also on the calibration of the model's confidence in these predictions. Regression metrics, on the other hand, assess the accuracy of continuous predictions, emphasizing error magnitudes and variance explained.

6.1.1 Classification

Many novel indicators involve using a classification model that makes probabilistic predictions. A range of metrics is commonly used for optimizing and evaluating these classifiers:

Accuracy simply counts how often the classifier correctly predicts the true label. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where TP stands for true positives, TN for true negatives, FP for false positives and FN for false negatives.

Precision measures how many of the cases that are predicted to belong to the positive class actually turned out to be positive. Precision is useful in cases where false positives are a higher concern than false negatives.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall measures how many of the actual positive cases the model was able to predict correctly. Recall is a useful metric in cases where false negatives are of a higher concern than

false positives. For example, it is important in injury prediction where it is less harmful to raise a false alarm than that players with an actual injury risk go undetected.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F1 score computes the harmonic mean of the precision and recall metrics. The F1 score could be an effective evaluation metric when false positives and false negatives are equally costly.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

AUROC is calculated as the area under (AU) the receiver operator curve (ROC), which shows the trade-off between recall and the false positive rate ($FPR = FP/(TN + FP)$) across different decision thresholds. It is essentially a ranking metric and assesses whether the model ranks the positive examples ahead of the negative examples.

Logarithmic loss measures the uncertainty of the prediction based on how much it varies from the actual outcome. The logarithmic loss is computed as

$$LL = \frac{1}{|c|} \sum_{i=1}^{|c|} y_i \log \hat{p}_i \tag{6}$$

with $|c|$ the number of classes, $\hat{p}_i$ the probability forecast of class i and y_i a binary indicator of whether or not class i is the true outcome. A perfect classifier would have a logarithmic loss of precisely zero. Less ideal classifiers have progressively larger values.

Brier Score measures the accuracy of probabilistic predictions as the mean squared difference between the predicted probabilities and the actual outcomes:

$$BS = \frac{1}{|c|} \sum_{i=1}^{|c|} (\hat{p}_i - y_i)^2 \tag{7}$$

with $|c|$ the number of classes, $\hat{p}_i$ the probability forecast of class i and y_i a binary indicator of whether or not class i is the true outcome.

Ranked Probability Score (RPS) was introduced by Epstein (1969) to evaluate probability forecasts of ranked categories. It is related to the Brier score, but it explicitly accounts for the ordinal structure of the predictions. For example, when predicting win-draw-loss match outcomes in soccer, it can be taken into account that predicting a tie when the actual outcome is a loss is considered a better prediction than a win. It can be defined as

$$RPS = \frac{1}{|c| - 1} \sum_{k=1}^{|c|-1} \left(\sum_{l=1}^{k} (\hat{p}_l - y_l) \right)^2 \tag{8}$$

As the RPS is an error measure, a lower value corresponds to a better fit.

The selection of an evaluation metric should be tailored to the particular situation in which the predictive model is intended to be employed. When the dataset is not significantly imbalanced, the AUROC metric is most suitable for classification tasks or ranking examples based on their probability of being either positive or negative. For example, when searching for the top-k tactical patterns that are most likely to result in a goal. However, if the dataset displays a significant imbalance and/or the primary concern revolves around the

positive class, the F1 score or Precision-Recall curve should be considered. Additionally, F1 (or precision and recall) has the advantage of being more straightforward to interpret and convey to sports practitioners.

When we care about using the actual values of the probabilities, the choice is between the Brier score or RPS (depending on whether the target variable is ordered) and logarithmic loss. Brier score and logarithmic loss are similar in the sense that they are both proper scoring rules and can both only be minimized by reducing the individual prediction errors. However, they differ in how they aggregate the individual prediction errors. To illustrate this difference and to more easily compare the two metrics, let $e_i = |\hat{p}_i - y_i|$ be the prediction error for example i. Using this definition and the multiplication rule for logarithms, we can simplify the formulas for the Brier score and logarithmic loss to:

$$BS = \frac{1}{|c|} \sum_{i=1}^{|c|} e_i^2 \tag{9}$$

$$LL = \frac{1}{|c|} \sum_{i=1}^{|c|} \log\left(1 - e_i\right) = \frac{1}{|c|} \log\left(\prod_{i=1}^{|c|} 1 - e_i\right) \tag{10}$$

Hence, the Brier score combines individual prediction errors by summing them while the logarithmic loss combines individual prediction errors by multiplying them. This insight is the reason we recommend using the Brier score to build a predictive model if the resulting probabilities will be summed (e.g., for computing player ratings) or subtracted (e.g., modeling decision-making) (Decroos & Davis, 2020).

Proper scoring rules might not be sufficient to evaluate if a model is calibrated since a high discriminative power can obscure a poor calibration. Therefore, they should always be combined with reliability diagrams (Niculescu-Mizil & Caruana, 2005) or the multi-class expected calibration error (ECE) (Guo et al., 2017).

Also, it is important to remember that the aforementioned metrics (with the exception of AUROC) depend on the class distribution, and hence their values need to be interpreted in this context. This is important as scoring rates can vary by competition (e.g., men's leagues vs. women's leagues) (Pappalardo et al., 2021).

6.1.2 Regression

Regression evaluation metrics can be broadly categorized into two types: error-based metrics and variance-explained metrics.

Error-based metrics assess the quality of a model as a function of the errors between the actual values y_i and predicted values $\hat{y}_i$:

$$\frac{1}{N} \sum_{i=1}^{N} f(y_i - \hat{y}_i) \tag{11}$$

They differ in how they measure the differences between y_i and $\hat{y}_i$. The Mean Absolute Error (MAE) takes the absolute value of the errors, providing a straightforward interpretation of the typical prediction error in the same units as the data. The Mean Squared Error (MSE) calculates the average of the squared errors, penalizing larger errors more heavily than smaller ones. Finally, the Mean Absolute Percentage Error (MAPE) measures the average absolute percentage error between predicted and actual values. It is useful

for understanding prediction accuracy in relative terms, especially when dealing with data where values vary widely.

Variance-explained metrics such as the coefficient of determination (R^2 *score*) measure the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides an indication of how well the model fits the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2} \tag{12}$$

where N is the number of predictions, y_i is the actual value, $\hat{y}_i$ is the predicted value and $\bar{y}$ is the mean of the actual values. An R^2 value of 1 indicates a perfect fit, while an R^2 value of 0 indicates that the model does not explain any of the variance.

When evaluating the fit of a regression model, it is useful to compute both error-based metrics and the R^2 value since they provide distinct insights. On one hand, error-based metrics inform us about the average error between the regression model's predicted values and the actual values. On the other hand, R^2 indicates the extent to which the predictor variables can account for the variability observed in the response variable.

6.1.3 Other tasks

Finally, certain applications might require specialized metrics. For example, when forecasting match results for betting purposes, neither the actual probabilities nor the most likely outcome are as critical. More importantly, it is essential to identify betting opportunities that are undervalued. Economic measures are more appropriate in that case (Wunderlich & Memmert, 2020).

6.2 Sensitivity analyses and feature importances

To promote interactions with domain experts and to gain insights into the domain, there has been interest in using the techniques developed in explainable AI to understand the importance and uncertainty of various features. Such analysis can rely on techniques such as SHAP (Lundberg & Lee, 2017) or LIME (Ribeiro et al., 2016) to derive post-hoc feature importances for individual predictions. For example, Anzer and Bauer (2021) use SHAP to analyze the influence of different features on expected goals models. However, these methods have limitations, as they only offer insights into how individual features influence the predictions of a machine learning model. Sun et al. (2020) addressed this issue by investigating mimic learning, a technique to translate black-box models into interpretable ones through model distillation. Another approach is to consider interpretability from the start. This can be done by considering models such as variants of generalized additive models (Caruana et al., 2015; Nori et al., 2019). This enables showing partial dependency plots, which can give intuitions about which features are important. These have been used to explore both expected goals models (Bransen & Davis, 2021) and action-value models (Decroos & Davis, 2020).

While Bayesian approaches have also been used extensively for sports models, these typically measure uncertainty surrounding the learned model's parameters and not its predictions. There are some exceptions, such as the work by Kwiatkowski (2017) which employed Bayesian logistic regression to train an expected goals model for soccer. In order to assess finishing skill, this work included an indicator variable for each player in the

feature set of an expected goals model. The sign and magnitude of the weight associated with a player's variable represented a player's finishing skill with the range of possible values representing the confidence in the assessment.

Finally, one idea is to take a more domain-knowledge-driven approach and formulate features in a way that directly relates to concepts that are understandable by practitioners. For example, Van Haaren (2021) trained an expected goals model that encoded the location of the shot based on zones or areas that are typically used by coaches. Similarly, Fernandez-Navarro et al. (2016) considered typical tactical indicators that were supplied by a data provider in a principle component analysis to understand playing style.

6.3 Quantifying uncertainty at prediction time

When a learned model is operating in the wild, it is important that the user trusts its predictions. Thus, one very relevant question is whether the model makes systematic mistakes. That is, are there certain regions of the instance space where the model consistently performs worse. This may be because there was little training data available for certain, rare situations, or some data encountered in practice is very dissimilar to the training data (e.g., noise, annotation errors, data drift, a rare match situation).

To increase trust, one recent trend in machine learning is for learned models to quantify the uncertainty associated with their predictions (Hüllermeier & Waegeman, 2021). This can take several different forms, such as providing a range of plausible values instead of a point prediction (Li et al., 2021; Friedman, 2001; Khosravi et al., 2011), assigning a confidence score to the prediction (Hyndman & Athanasopoulos, 2023; Papadopoulos et al., 2007), or flagging test instances where the model is at a heightened risk of making a misprediction (Cortes et al., 2016; Hendrickx et al., 2024).

While these can help promote trust in a model, there has been less work in applying these techniques in the context of sports analytics. One exception is the work of Van Roy and Davis (2023) which reasoned about predictions for an action-value model in soccer based on tree ensembles. Specifically, they applied an approach that can flag when a tree ensemble makes a prediction that is dissimilar to those observed during training time (Devos et al., 2023). It is known that learned models struggle to extrapolate beyond the training data (Seo et al., 2000) and even when confronted with subtly different inputs, a learned model may behave in unexpected ways (Szegedy et al., 2014). In the context of soccer, different data may arise for two reasons. One, the data is often annotated by humans who make mistakes. Two, there could be a rare or fluke play that occurred in a match (e.g., a shot from midfield). Van Roy and Davis (2023) were able to automatically detect both types of issues.

6.4 Verification: reasoning about learned models

Reasoning techniques such as verification are a powerful alternative to traditional aggregated metrics to evaluate a learned model (Straccia & Pratesi, 2022). Verification goes beyond standard metrics by investigating how a model will behave in practice. That is, it allows the user to (interactively) query the model to answer hypothetical questions about examples that are not in the train, validation or test sets. Moreover, in certain instances, it can provide formal guarantees about how the model will behave. Consequently, it helps a user gain insight into what the model has learned from the data. The following are examples of types of questions that a user may have that reasoning can address:

- Is a model robust to small changes to the inputs? For example, does a small change in the time of the game and the position of the ball significantly change the probability that a shot will result in a goal? This relates to adversarial examples (c.f. image recognition).
- Given a specific example of interest, can one or more attributes be (slightly) changed so that the indicator is maximized? This is often called a *counterfactual* explanation, e.g., *if* the goalie would have been positioned closer to the near post, how would that have affected the estimated probability of the shot resulting in a goal? We want to emphasize that, this is not a causal counterfactual (because the considered models are not causal models).
- Does the model behave as expected in scenarios where we have strong intuitions based on domain knowledge? For example, one can analyze what values the model can predict for shots that are taken from a very tight angle or very far away from the goal. One can then check whether the predictions for the generated game situations are realistic.

Typical aggregated test metrics do not reveal the answers to these questions. Nevertheless, the answers can be very valuable because they provide insights into the model and can reveal problems with the model or the data.

To better understand how verification goes beyond standard metrics, consider the question: all other features being held constant, does a small change in the time of the game significantly change the probability that a shot will result in a goal? By making predictions (i.e., applying forward reasoning or deduction to derive a target value from observed feature values), it is possible to check that this never happens for all pairs of examples in the train, validation and test set. However, it does not allow ruling out the possibility that two such shots could arise in future games where the model offers wildly different predictions despite the fact they agree on all features except for time. Verification approaches on the other hand can answer such questions by combining backwards and forwards reasoning (Devos et al., 2021; Kwiatkowska et al., 2011; Russell et al., 2015; Kantchelian et al., 2016; Katz et al., 2017). Given a desired target value (i.e., prediction), and possibly some constraints on the values that the features can take on, a verification algorithm either constructs one or more instances that satisfy the constraints, or it proves that no such instance exists. Importantly, the constructed instances are not restricted to being in the currently available data. Consequently, it provides formal guarantees, like theorem provers (e.g., as used in satisfiability testing).

We have used verification to evaluate soccer models in two novel ways. First, we show how it is possible to debug the training data and pinpoint labeling errors (or inconsistencies). Second, we identify scenarios where the model produces unexpected and undesired predictions. These are shortcomings in the model itself. We use Veritas (Devos et al., 2021) to analyze two previously mentioned soccer analytics models: xG and the VAEP holistic action-value model.

First, we analyzed an xG model in order to identify "what are the optimal locations to shoot from outside the penalty box?" We used Veritas to generate 200 examples of shots from outside the penalty box that would have the highest probability of resulting in a goal, which are shown as a heatmap in Fig. 6. The cluster in front of the goal is expected as it corresponds to the areas most advantageous to shoot from. The locations near the corners of the pitch are unexpected. We looked at the shots from the 5 m square area touching the corner and counted 11 shots and 8 goals, yielding an extremely high 72% conversion rate. This reveals an unexpected labeling behavior by the human annotators. Given the distance to the goal and the tight angle, one would expect a much lower conversion rate. This is

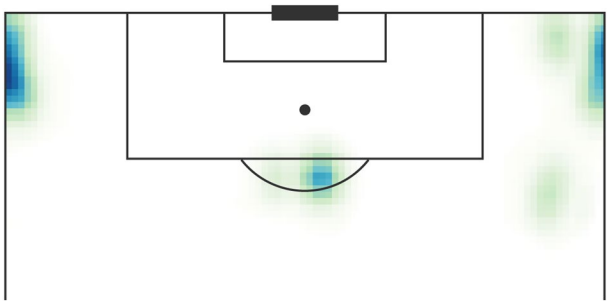

Fig. 6 A heatmap showing where VERITAS generates instances of shots from outside the penalty box with the highest xG values

likely because the human annotators are only labeling actions as a shot in the rare situations where the action results in a goal or a save. Otherwise, the actions are labeled as a pass or a cross.

Second, we analyzed VAEP (Decroos et al., 2019), a holistic action-value model for soccer. The models underlying this indicator look at a short sequence of consecutive game actions and predict the probability of a goal in the next 10 actions. Unlike xG models, all possible actions (passes, dribbles, tackles, ...) are considered, not just shots. For the data in an unseen test set, the model produces well-calibrated probability estimates in aggregate. However, we used verification to look for specific scenarios where the model performs badly. VERITAS was able to construct several instances that did not appear in the training, validation or testing data that, while unlikely, exhibited uncredible behavior by the model. Concretely, VERITAS generated instances where all the values of all features were fixed except for the time in the match, and found that the probability of scoring varied dramatically according to match time. Figure 7 shows this variability for one such instance. The probability gradually increases over time, which is not necessarily unexpected as scoring rates tend to slightly increase as a match progresses. However, about 27 min into the first half the probability of scoring dramatically spikes. Clearly, this behavior is undesirable: we would not expect such large variations. Moreover, this behavior did not arise when investigating the data we had access to. This suggests that time should probably be handled differently in the model, e.g., by discretizing it to make it less fine-grained.

Such an evaluation is still challenging. One has to know what to look for, which typically requires significant domain expertise or access to a domain expert. However, such an evaluation neither requires expertise in machine learning nor deep knowledge of the underlying model. Arguably, this is a big advance because it enables domain experts themselves to validate models and assess liability. Moreover, the process is exploratory: there is a huge space of scenarios to consider and the questions have to be iteratively refined.

7 Discussion

Applying an appropriate methodology and evaluation framework when using learning systems in the context of sports largely relies on expertise gained through experience. Our goal in writing this article was to (1) elucidate and summarize some of the key issues and (2) provide advice and examples about how to tackle these problems. Ideally, this will benefit both data scientists starting to work in this area and practitioners who hope to make more use of automated analysis.

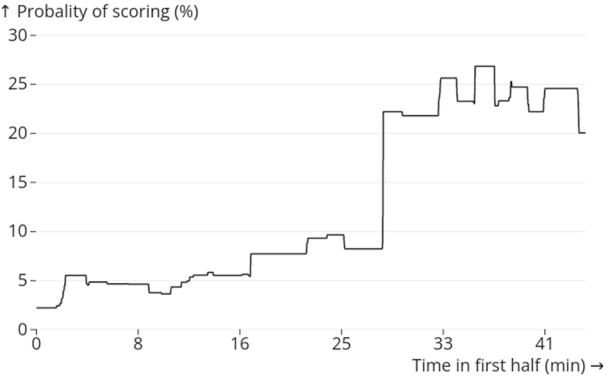

Fig. 7 For specific action sequences, the time remaining in the game has a large variable effect on the probability of scoring in the VAEP action model. This variability is unexpected and reveals a robustness issue with the model

Methodologically, it is important to critically think about the data at hand. First, it is necessary to identify what dependencies exist in the data and in what situations the models will be used. This will inform how the data is partitioned to learn and evaluate the models that form the basis of the developed indicators. Second, there can be biases in the data and this can inform both which features and model classes should (not) be considered. Third, one must consider contextual factors, both when designing features and aggregating model outputs to produce an indicator's final value.

Because constructing novel indicators of performance involves combining the outputs of learned models, it is important to differentiate between evaluating the validity of the indicator itself and the quality of the models underlying it. Both aspects must be considered, but there are important differences in how to conduct the evaluations.

When evaluating the validity of novel indicators of performance, we would like to caution against looking at correlations to other success metrics as we believe that a high correlation to an existing indicator fails the central goal: gaining new insights. In this, involving a domain expert in the evaluation can be extremely insightful. We also believe that the reliability and stability of indicators is important, and should be more widely studied. The best approach for evaluation is often a specific problem. We encourage researchers to critically think about the end objective of their indicators and explicitly state and discuss their added benefit over existing ones. It is particularly useful to ground such a discussion in how the indicators may be used in practice. Moreover, we hope that the field will continue to discuss best practices.

When evaluating the models used to construct the underlying systems and indicators, we believe that a variety of different factors must be considered. While standard metrics measuring some notion of predictive accuracy are important, only considering them is insufficient. It is crucial to have broader insights into a model's behavior, particularly since this helps facilitate interactions with domain experts. Consequently, it is necessary to consider aspects such as interpretability, explanations and trust. We believe that evaluating models by *reasoning about their behavior* is crucial: this changes the focus from a purely data-based evaluation perspective which is the current norm to one that considers the effect of potential future data on the model. Critically reflecting on what situations a model will work well in and which situations it may struggle in, helps build trust and set appropriate expectations.

Still, using reasoning is not a magic solution. When a reasoner identifies unexpected behaviors, there are at least two possible causes. One cause is errors in the training data

which are picked up by the model and warp the decision boundary in unexpected ways (e.g., Fig. 6). Some errors can be found by inspecting the data, but given the nature of the data, it can be challenging to know where to look. The other cause is peculiarities with the model itself, the learning algorithm that constructed the model, or the biases resulting from the model representation (e.g., Fig. 7). Traditional evaluation metrics are completely oblivious to these issues. They can only be discovered by reasoning about the model. Unfortunately, it remains difficult to correct a model that has picked up on an unwanted pattern. For example, the time's effect on the probability of scoring can only be resolved via representing the feature in a different way, relearning the model, and reassessing its performance. Alas, this is an iterative guess-and-check approach. We believe that reasoning approaches to evaluation are only in their infancy and need to be further explored.

While this paper grounded its discussion of methodology and evaluation in the context of sports, we do feel that some of the insights and advice are relevant for other application domains where machine learning is applied. For example, evaluation challenges also arise in prognostics, especially when it is impossible to directly collect data about a target such as time until failure. In both domains, we do not want to let the athlete nor machine be damaged beyond repair. Also, we perform multiple actions to avoid failure, making it difficult to attribute value to individual actions or identify root causes. Another example is how to deal with subjective ratings provided by users, which often occurs when monitoring players' fitness and was also a key issue in the Netflix challenge. Finally, in other application domains such as finance and health where models are currently deployed, it is also important to evaluate the robustness of learned models.

Author Contributions Conceptualization: JD, LB, PR, LD, WM, AJ, JVH, MVR; Methodology: JD, PR, LD; Writing—original draft preparation: JD, PR, LD, WM; Writing—review and editing: JD, LB, PR, LD, WM, AJ, JVH, MVR; Funding acquisition: WM, JD; Resources: WM, JD; Supervision: JD.

Funding This research received funding from The European Union's Horizon Europe Research and Innovation program under the grant agreement TUPLES No. 101070149 (LD, JD), Interuniversity Special Research Fund (IBOF/21/075), the Flemish Government under the "Onderzoeksprogramma Artifciële Intelligentie (AI) Vlaanderen" program, and the Research Foundation—Flanders (LD: 1SB1322N).

Data availability Not applicable.

Code availability Not applicable.

Declaration

Conflict of interest Not applicable.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Albert, J., Glickman, M.E., & Swartz TB, et al (2017). Handbook of Statistical Methods and Analyses in Sports. Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Chapman & Hall.

Andrienko, G., Andrienko, N., Anzer, G., et al. (2019). Constructing spaces and times for tactical analysis in football. *IEEE Transactions on Visualization and Computer Graphics, 27*(4), 2280–2297.

Anzer, G., & Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (Soccer). *Frontiers in Sports and Active Living, 3*, 624475.

Anzer, G., Brefeld, U., & Bauer, P., et al. (2022). Detection of tactical patterns using semi-supervised graph neural networks. In: MIT Sloan Sports Analytics Conference.

Arbués Sangüesa, A. (2021). A journey of computer vision in sports: from tracking to orientation-base metrics. PhD thesis, Universitat Pompeu Fabra.

Balestriero, R., Ibrahim, M., & Sobal, V., et al. (2023). A cookbook of self-supervised learning. arXiv:2304.12210.

Baptista, I., Johansen, D., Figueiredo, P., et al. (2019). A comparison of match-physical demands between different tactical systems: 1–4-5-1 vs 1–3-5-2. *PLOS ONE, 14*(4), 1–12. https://doi.org/10.1371/journal.pone.0214952

Baron, E., Sandholtz, N., Chan, T., et al. (2024). Miss it like Messi: Extracting value from off-target shots in soccer. *Journal of Quantitative Analysis in Sports, 20*(1), 37–50.

Bartlett, J., O'Connor, F., & Naa, Pitchford. (2017). Relationships between internal and external training load in team sports athletes: Evidence for an individualised approach. *International Journal of Sports Physiology and Performance, 12*(2), 230–234.

Bauer, P., & Anzer, G. (2021). Data-driven detection of counterpressing in professional football. *Data Mining and Knowledge Discovery, 35*, 2009–2049.

Baumer, B. S., Matthews, G. J., & Nguyen, Q. (2023). Big ideas in sports analytics and statistical tools for their investigation. *Wiley Interdisciplinary Reviews Computational Statistics, 15*(6), e1612.

Bekkers, J., & Dabadghao, S. S. (2019). Flow motifs in soccer: What can passing behavior tell us? *Journal of Systems Architecture, 5*, 299–311.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis, 120*, 70–83.

Bialik, C. (2014a). The people tracking every touch, pass and tackle in the world cup. https://fivethirtyeight.com/features/the-people-tracking-every-touch-pass-and-tackle-in-the-world-cup/.

Bialik, C. (2014b). Statkeepers call the shots, but they can't agree on them. https://fivethirtyeight.com/features/statkeepers-call-the-shots-but-they-cant-agree-on-them/.

van Bommel, M., & Bornn, L. (2017). Adjusting for scorekeeper bias in NBA box scores. *Data Mining and Knowledge Discovery, 31*(6), 1622–1642.

Borg, G. (1982). Psychophysical bases of perceived exertion. *Medicine Science in Sports Exercise, 14*(5), 377–381.

Bornn, L., Ward, P., & Norman, D. (2019). Training schedule confounds the relationship between acute:chronic workload ratio and injury. In: MIT Sloan Sports Analytics Conference.

Bouey, M. (2013). NBA win probability added. https://www.inpredictable.com/2013/06/nba-win-probability-added.html.

Bourdon, P. C., Cardinale, M., Murray, A., et al. (2017). Monitoring athlete training loads: Consensus statement. *International Journal of Sports Physiology and Performance, 12*(S2), 161–170.

Bransen, L., & Davis, J. (2021). Women's football analyzed: Interpretable expected goals models for women. In: *Proceedings of the AI for Sports Analytics Workshop*.

Bransen, L., Robberechts, P., & Van Haaren, J., et al. (2019). Choke or shine? quantifying soccer players' abilities to perform under mental pressure. In: MIT Sloan Sports Analytics Conference.

Buchheit, M., Cholley, Y., & Lambert, P. (2016). Psychometric and physiological responses to a preseason competitive camp in the heat with a 6-hour time difference in elite soccer players. *International Journal of Sports Physiology and Performance, 11*(2), 176–181.

Burke, B. (2010). WPA explained. http://archive.advancedfootballanalytics.com/2010/01/win-probability-added-wpa-explained.html.

Carling, C., Williams, A.M., & Reilly, T. (2005). Handbook of soccer match analysis: A Systematic Approach to Improving Performance. Routledge.
Caruana, R., Lou, Y., & Gehrke, J., et al. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p 1721–1730, https://doi.org/10.1145/2783258.2788613.
Cervone, D., D'Amour, A., & Bornn, L., et al. (2014). POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data. In: MIT Sloan Sports Analytics Conference.
Chmura, P., Liu, H., & Andrzejewski, M., et al. (2021). Is there meaningful influence from situational and environmental factors on the physical and technical activity of elite football players? Evidence from the data of 5 consecutive seasons of the German bundesliga. PLoS One 16(3).
Cortes, C., DeSalvo, G., & Mohri, M. (2016). Learning with rejection. In: *Proceedings of The 27th International Conference on Algorithmic Learning Theory* (ALT 2016).
Davis, J., & Robberechts, P. (2023). Expected metrics as a measure of skill: Reflections on finishing in soccer. In: *Proceedings of 10th Workshop on Machine Learning and Data Mining for Sports Analytics*.
Davis, J., & Robberechts, P. (2024). Biases in expected goals models confound finishing ability. arXiv:2401.09940.
De Brabandere, A., Op De Beéck, T., Schütte, K. H., et al. (2018). Data fusion of body-worn accelerometers and heart rate to predict vo2max during submaximal running. *PloS One, 13*(6), e0199509.
Decroos, T. (2020). Soccer analytics meets artificial intelligence: Learning value and style from soccer event stream data. PhD thesis.
Decroos, T., & Davis, J. (2019). Player vectors: Characterizing soccer players' playing style from match event streams. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp 569–584.
Decroos, T., & Davis, J. (2020). Interpretable prediction of goals in soccer. In: AAAI 2020 Workshop on AI in Team Sports.
Decroos, T., Bransen, L., & Van Haaren, J., et al. (2019). Actions speak louder than goals: valuing player actions in soccer. In: *Proceedings of 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp 1851–1861.
Dehghani, A., Glatard, T., & Shihab, E. (2019). Subject cross validation in human activity recognition. arXiv preprint arXiv:1904.02666.
Devos, L., Meert, W., & Davis, J. (2021). Versatile verification of tree ensembles. In: *Proceedings of the 38th International Conference on Machine Learning*, pp 2654–2664.
Devos, L., Perini, L., & Meert, W., et al. (2023). Adversarial example detection in deployed tree ensembles. In: *Proceeding of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp 120–136.
Dick, U., Link, D., & Brefeld, U. (2022). Who can receive the pass? A computational model for quantifying availability in soccer. *Data Mining and Knowledge Discovery, 36*(3), 987–1014.
Dorschky, E., Camomilla, V., Davis, J., et al. (2023). Perspective on "in the wild'' movement analysis using machine learning. *Human Movement Science, 87*, 103042.
Dubitzky, W., Lopes, P., Davis, J., et al. (2019). The open international soccer database for machine learning. *Machine Learning, 108*(1), 9–28.
Eirale, C., Tol, J., Farooq, A., et al. (2013). Low injury rate strongly correlates with team success in Gatari professional football. *British Journal of Sports Medicine, 47*(12), 807–8.
Epasinghege Dona, N., & Swartz, T. (2024). Causal analysis of tactics in soccer: The case of throw-ins. *IMA Journal of Management Mathematics, 35*(1), 111–126.
Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology and Climatology, 8*(6), 985–987.
Fernandez, J., & Bornn, L. (2018). Wide open spaces: A statistical technique for measuring space creation in professional soccer. In: MIT Sloan Sports Analytics Conference.
Fernández, J., Bornn, L., & Cervone, D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning, 110*(6), 1389–1427.
Fernandez-Navarro, J., Fradua, L., Zubillaga, A., et al. (2016). Attacking and defensive styles of play in soccer: Analysis of Spanish and English elite teams. *Journal of Sports Sciences, 34*(24), 2195–2204. https://doi.org/10.1080/02640414.2016.1169309
Fernandez-Navarro, J., Fradua, L., Zubillaga, A., et al. (2019). Evaluating the effectiveness of styles of play in elite soccer. *International Journal of Sports Science & Coaching, 14*(4), 514–527.
Fifield, D. (2022). The art of playing in a back three compared to a back four, told by those who have done it. https://www.nytimes.com/athletic/3679252/2022/10/18/back-three-compared-to-back-four/.

Franks, A., Miller, A., & Bornn, L., et al. (2015). Counterpoints: Advanced defensive metrics for NBA basketball. In: MIT Sloan Sports Analytics Conference.
Franks, A. M., D'Amour, A., Cervone, D., et al. (2016). Meta-analytics: Tools for understanding the statistical properties of sports metrics. *Journal of Quantitative Analysis in Sports, 12*(4), 151–165.
Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics pp 1189–1232.
Goldsberry, K. (2019). How deep, audacious 3-pointers are taking over the NBA. https://www.espn.com/nba/story/_/id/28312678/how-deep-audacious-3-pointers-taking-nba.
Gollan, S., Bellenger, C., & Norton, K. (2020). Contextual factors impact styles of play in the English Premier League. *Journal of Sports Science and Medicine, 19*(1), 78–83.
Green, S. (2012). Assessing the performance of Premier League goalscorers. https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/.
Guo, C., Pleiss, G., & Sun, Y., et al. (2017). On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning*, pp 1321–1330.
Halson, S. L. (2014). Monitoring training load to understand fatigue in athletes. *Sports Medicine, 44*(2), 139–147.
Hebert-Johnson, U., Kim, M., & Reingold, O., et al. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In: *Proceedings of the 35th International Conference on Machine Learning*, p 1939–1948.
Hendrickx, K., Perini, L., Van der Plas, D., et al. (2024). Machine learning with a reject option: A survey. *Machine Learning, 113*(5), 3073–3110.
Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. *Machine Learning, 110*(3), 457–506.
Hvattum, L. M. (2020). Offensive and defensive plus-minus player ratings for soccer. *Applied Sciences, 10*(20), 7345.
Hyndman, R., & Athanasopoulos, G. (2023). *Forecasting: Principles and Practice* (3rd ed.). OTexts: Melbourne, Australia.
Impellizzeri, F., Rampinini, E., & Marcora, S. (2005). Physiological assessment of aerobic training in soccer. *Journal Sports Science, 23*(6), 583–592.
Jacobs, A.Z., & Wallach, H. (2021). Measurement and fairness. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, p 375–385.
Jaspers, A., Op De Beéck, T., Brink, M. S., et al. (2018). Relationships between the external and internal training load in professional soccer: What can we learn from machine learning? *International Journal of Sports Physiology and Performance, 13*(5), 625–630.
Jeffries, A., Marcora, S., Coutts, A., et al. (2022). Development of a revised conceptual framework of physical training for use in research and practice. *Sports Medicine, 52*, 709–724.
Johnson, N. (2020). Extracting player tracking data from video using non-stationary cameras and a combination of computer vision techniques. In: MIT Sloan Sports Analytics Conference.
Joyner, M. J. (1991). Modeling: optimal marathon performance on the basis of physiological factors. *Journal of Applied Physiology, 70*(2), 683–687.
Kantchelian, A., Tygar, J.D., & Joseph, A. (2016). Evasion and hardening of tree ensemble classifiers. In: *Proceeding of the 33rd International Conference on Machine Learning*, pp 2387–2396.
Katz, G., Barrett, C., & Dill, D.L., et al. (2017). Reluplex: An efficient smt solver for verifying deep neural networks. In: Computer Aided Verification, pp 97–117.
Kempton, T., Kennedy, N., & Coutts, A. J. (2016). The expected value of possession in professional rugby league match-play. *Journal of Sports Sciences, 34*(7), 645–650.
Khosravi, A., Nahavandi, S., Creighton, D., et al. (2011). Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks, 22*(9), 1341–1356.
Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 426–434.
Koren, Y. (2010). Collaborative filtering with temporal dynamics. *Communications of the ACM, 53*(4), 89–97.
Kovalchik, S. A. (2023). Player tracking data in sports. *Annual Review of Statistics and Its Application, 10*(1), 677–697.
Kwiatkowska, M., Norman, G., & Parker, D. (2011). PRISM 4.0: Verification of probabilistic real-time systems. In: *Proceeding of the 23rd International Conference on Computer Aided Verification*, pp 585–591.
Kwiatkowski, M. (2017). Quantifying finishing skill. https://statsbomb.com/articles/soccer/quantifying-finishing-skill/.

Le, H.M., Yue, Y., & Carr, P., et al. (2017). Coordinated multi-agent imitation learning. In: *Proceedings of the 34th International Conference on Machine Learning*, pp 1995–2003.
de Leeuw, A. W., van der Zwaard, S., van Baar, R., et al. (2022). Personalized machine learning approach to injury monitoring in elite volleyball players. *European Journal of sport Science, 22*, 511–520.
de Leeuw, A. W., Heijboer, M., Verdonck, T., et al. (2023). Exploiting sensor data in professional road cycling: personalized data-driven approach for frequent fitness monitoring. *Data Mining and Knowledge Discovery, 37*, 1125–1153.
Li, R., Reich, B. J., & Bondell, H. D. (2021). Deep distribution regression. *Computational Statistics & Data Analysis, 159*, 107203.
Link, J., Schwinn, L., & Pulsmeyer, F., et al. (2022). xlength: Predicting expected ski jump length shortly after take-off using deep learning. Sensors 22(21). https://doi.org/10.3390/s22218474, https://www.mdpi.com/1424-8220/22/21/8474.
Liu, G., & Schulte, O. (2018). Deep reinforcement learning in ice hockey for context-aware player evaluation. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp 3442–3448.
Llana, S., Burriel, B., Madrero, P., et al. (2022). Is it worth the effort? Understanding and contextualizing physical metrics in soccer. https://doi.org/10.48550/arXiv.2204.02313, arXiv:2204.02313.
Lowe, Z. (2013). Lights, cameras, revolution. https://grantland.com/features/the-toronto-raptors-sportvu-cameras-nba-analytical-revolution/.
Lucey, P., Bialkowski, A., Monfort, M., et al. (2015). Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In: MIT Sloan Sports Analytics Conference.
Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems.
Macdonald, B. (2012). An expected goals model for evaluating NHL teams and players. In: MIT Sloan Sports Analytics Conference.
McHale, I., & Scarf, P. (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica, 61*(4), 432–445.
McHale, I., Scarf, P., & Folker, D. (2012). On the development of a soccer player performance rating system for the English Premier League. *Interfaces, 42*(4), 339–351.
McHale, I. G., & Holmes, B. (2023). Estimating transfer fees of professional footballers using advanced performance metrics and machine learning. *European Journal of Operational Research, 306*(1), 389–399.
Merckx, S., Robberechts, P., & Euvrard, Y., et al. (2021). Measuring the effectiveness of pressing in soccer. In: Workshop on Machine Learning and Data Mining for Sports Analytics.
Miller, A., & Bornn, L. (2017). Possession sketches: Mapping NBA strategies. In: MIT Sloan Sports Analytics Conference.
Modric, T., Versic, S., & Winter, C., et al. (2022). The effect of team formation on match running performance in UEFA Champions League matches: Implications for position-specific conditioning. Science and Medicine in Football pp 1–8. https://doi.org/10.1080/24733938.2022.2123952.
Mortensen, J., & Bornn, L. (2020). Estimating locomotor demands during team play from broadcast-derived tracking data. arXiv preprint arXiv:2001.07692.
Munson, M. A. (2011). A study on the importance of and time spent on different modeling steps. *SIGKDD Explorations, 13*(2), 65–71.
Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine learning*, p 625–632.
Nori, H., Jenkins, S., & Koch, P., et al. (2019). Interpretml: A unified framework for machine learning interpretability. https://doi.org/10.48550/arXiv.1909.09223, arXiv:1909.09223.
Oliver, D. (2002). Basketball on Paper. Brassey's, Inc.
Op De Beéck, T., Meert, W., & Schütte, K., et al. (2018). Fatigue prediction in outdoor runners via machine learning and sensor fusion. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp 606–615.
Papadopoulos, H., Vovk, V., & Gammerman, A. (2007). Conformal prediction with neural networks. In: *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, pp 388–395, https://doi.org/10.1109/ICTAI.2007.47.
Pappalardo, L., Cintia, P., Ferragina, P., et al. (2019). Playerank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology, 10*(5), 1–27.
Pappalardo, L., Rossi, A., & Natilli, M., et al. (2021). Explaining the difference between men's and women's football. PLoS ONE 16(8).

Pelechrinis, K., Winston, W., & Sagarin, J., et al. (2018). Evaluating nfl plays: Expected points adjusted for schedule. In: *Proceedings of the 5th Workshop on Machine Learning and Data Mining for Sports Analytics*.

Pettigrew, S. (2015). Assessing the offensive productivity of NHL players using in-game win probabilities. In: MIT Sloan Sports Analytics Conference.

Phatak, A. A., Mehta, S., Wieland, F. G., et al. (2022). Context is key: normalization as a novel approach to sport specific preprocessing of KPI's for match analysis in soccer. *Scientific Reports, 12*(1), 1117.

Podlog, L., Buhler, C. F., Pollack, H., et al. (2015). Time trends for injuries and illness, and their relation to performance in the NBA. *Journal of Science and Medicine in Sport, 18*(3), 278–82.

Quang Nguyen, R. Y., & Matthews, G. J. (2024). Here comes the strain: Analyzing defensive pass rush in American football with player tracking data. *The American Statistician, 78*(2), 199–208. https://doi.org/10.1080/00031305.2023.2242442

Raysmith, B. P., & Drew, M. K. (2016). Performance success or failure is influenced by weeks lost to injury and illness in elite Australian track and field athletes: A 5-year prospective study. *Journal of Science and Medicine in Sport, 19*(10), 778–83.

Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p 1135–1144, https://doi.org/10.1145/2939672.2939778.

Robberechts, P., & Davis, J. (2020). How data availability affects the ability to learn good xG models. In: Workshop on Machine Learning and Data Mining for Sports Analytics, pp 17–27.

Robberechts, P., Van Haaren, J., & Davis, J. (2021). A Bayesian approach to in-game win probability in soccer. In: Proceedings of 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 3512–3521.

Romer, D. (2006). Do firms maximize? Evidence from professional football. *Journal of Political Economy, 114*(2), 340–365.

Rossi, A., Pappalardo, L., Cintia, P., et al. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. *PLOS ONE, 13*(7), 1–15. https://doi.org/10.1371/journal.pone.0201264

Routley, K., & Schulte, O. (2015). A Markov game model for valuing player actions in ice hockey. In: *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pp 782–791.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Magazine, 36*(4), 105–114.

Sandholtz, N., & Bornn, L. (2020). Markov decision processes with dynamic transition probabilities: An analysis of shooting strategies in basketball. *Annals of Applied Statistics, 14*(3), 1122–1145.

Sarlis, V., & Tjortjis, C. (2020). Sports analytics—evaluation of basketball players and team performance. *Information Systems, 93*, 101562.

Seo, S., Wallat, M., & Graepel, T., et al. (2000). Gaussian process regression: Active data selection and test point rejection. In: *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, vol 3. IEEE, pp 241–246, https://doi.org/10.1109/IJCNN.2000.861310, http://ieeexplore.ieee.org/document/861310/.

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

Silva Filho, T., Song, H., Perello-Nieto, M., et al. (2023). Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning, 112*(9), 3211–3260.

Silver, D., Hubert, T., & Schrittwieser, J., et al. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. https://doi.org/10.48550/arXiv.1712.01815, arXiv:1712.01815.

Singh, K. (2019). Introducing expected threat. https://karun.in/blog/expected-threat.html.

StatsBomb. (2021). Introducing On-Ball Value. https://statsbomb.com/articles/soccer/introducing-on-ball-value-obv/.

Straccia, U., & Pratesi, F. (2022). TAILOR handbook of trustworthy AI.

Sun, X., Davis, J., & Schulte, O., et al. (2020). Cracking the black box: Distilling deep sports analytics. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp 3154–3162.

Szegedy, C., Zaremba, W., & Sutskever, I., et al. (2014). Intriguing properties of neural networks. In: *Proceedings of the 2nd International Conference on Learning Representations*, arXiv:1312.6199.

Tureen, T., & Olthof, S. (2022). Estimated player impact (EPI): Quantifying the effects of individual players on football (soccer) actions using hierarchical statistical models. In: StatsBomb Conference.

Van Haaren, J. (2021). Why would I trust your numbers? On the explainability of expected values in soccer. In: *Proceedings of the AI for Sports Analytics Workshop*.

Van Roy, M., & Davis, J. (2023). Datadebugging: Enhancing trust in soccer action-value models by contextualization. In: *13th World Congress of Performance Analysis of Sport and 13th International Symposium on Computer Science in Sport*, pp 193–196.
Van Roy, M., Robberechts, P., & Decroos, T., et al. (2020). Valuing on-the-ball actions in soccer: A critical comparison of xT and VAEP. In: 2020 AAAI Workshop on AI in Team Sports.
Van Roy, M., Robberechts, P., & Yang, W.C., et al. (2021). Leaving goals on the pitch: Evaluating decision making in soccer. In: MIT Sloan Sports Analytics Conference.
Van Roy, M., Robberechts, P., Yang, W. C., et al. (2023). A Markov framework for learning and reasoning about strategies in professional soccer. *Journal of Artificial Intelligence Research, 77*, 517–562.
Vanrenterghem, J., Nedergaard, N., Robinson, M., et al. (2017). Training load monitoring in team sports: A novel framework separating physiological and biomechanical load-adaptation pathways. *Sports Medicine, 47*(11), 2135–2142.
Wang, Q., Zhu, H., & Hu, W., et al. (2015). Discerning tactical patterns for professional soccer teams: An enhanced topic model with applications. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp 2197–2206.
Williams, S., Trewartha, G., Kemp, S., et al. (2016). Time loss injuries compromise team success in elite rugby union: a 7-year prospective study. *British Journal of Sports Medicine, 50*(11), 651–656.
Windt, J., & Gabbett, T. (2017). How do training and competition workloads relate to injury? The workload-injury aetiology model. *British Journal of Sports Medicine, 51*(5), 428–435.
Wu, L. Y., & Swartz, T. B. (2023). The calculation of player speed from tracking data. *International Journal of Sports Science & Coaching, 18*(2), 516–522.
Wu, Y., Danielson, A., Hu, J., et al. (2021). A contextual analysis of crossing the ball in soccer. *Journal of Quantitative Analysis in Sports, 17*(1), 57–66.
Wunderlich, F., & Memmert, D. (2020). Are betting returns a useful measure of accuracy in (sports) forecasting? *International Journal of Forecasting, 36*(2), 713–722. https://doi.org/10.1016/j.ijforecast.2019.08.009
Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions, 26*(4), 42–46.
Yurko, R., Ventura, S., & Horowitz, M. (2019). nflWAR: A reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports, 15*(3), 163–183.

Authors and Affiliations

Jesse Davis[1,2] · Lotte Bransen[1,2] · Laurens Devos[1,2] · Arne Jaspers[3] · Wannes Meert[1,2] · Pieter Robberechts[1,2] · Jan Van Haaren[1,2,4] · Maaike Van Roy[1,2]

✉ Jesse Davis
jesse.davis@kuleuven.be

Lotte Bransen
lotte.bransen@kuleuven.be

Laurens Devos
laurens.devos@kuleuven.be

Arne Jaspers
arne.jaspers@kuleuven.be

Wannes Meert
wannes.meert@kuleuven.be

Pieter Robberechts
pieter.robberechts@kuleuven.be

Jan Van Haaren
jan.vanhaaren@kuleuven.be

Maaike Van Roy
maaike.vanroy@kuleuven.be

[1] Department of Computer Science, KU Leuven, 3000 Leuven, Belgium

[2] Leuven.AI-KU Leuven Institute for AI, 3000 Leuven, Belgium

[3] Department of Rehabilitation Sciences, KU Leuven, 3000 Leuven, Belgium

[4] Club Brugge, 8300 Westkapelle, Belgium