PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach

LUCA PAPPALARDO, ISTI-CNR, Italy
PAOLO CINTIA and PAOLO FERRAGINA, Department of Computer Science, University of Pisa, Italy
EMANUELE MASSUCCO, Wyscout, Italy
DINO PEDRESCHI, Department of Computer Science, University of Pisa, Italy
FOSCA GIANNOTTI, ISTI-CNR, Italy

The problem of evaluating the performance of soccer players is attracting the interest of many companies and the scientific community, thanks to the availability of massive data capturing all the events generated during a match (e.g., tackles, passes, shots, etc.). Unfortunately, there is no consolidated and widely accepted metric for measuring performance quality in all of its facets. In this article, we design and implement PlayeRank, a data-driven framework that offers a principled multi-dimensional and role-aware evaluation of the performance of soccer players. We build our framework by deploying a massive dataset of soccer-logs and consisting of millions of match events pertaining to four seasons of 18 prominent soccer competitions. By comparing PlayeRank to known algorithms for performance evaluation in soccer, and by exploiting a dataset of players' evaluations made by professional soccer scouts, we show that PlayeRank significantly outperforms the competitors. We also explore the ratings produced by PlayeRank and discover interesting patterns about the nature of excellent performances and what distinguishes the top players from the others. At the end, we explore some applications of PlayeRank—i.e. searching players and player versatility—showing its flexibility and efficiency, which makes it worth to be used in the design of a scalable platform for soccer analytics.

CCS Concepts: • **Information systems** → *Information retrieval*; *Learning to rank*; • **Computing methodologies** → *Machine learning approaches*;

Additional Key Words and Phrases: Sports analytics, soccer analytics, football analytics, clustering, searching, ranking, multi-dimensional analysis, predictive modelling, data science, big data

ACM Reference format:
Luca Pappalardo, Paolo Cintia, Paolo Ferragina, Emanuele Massucco, Dino Pedreschi, and Fosca Giannotti. 2019. PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach. *ACM Trans. Intell. Syst. Technol.* 10, 5, Article 59 (September 2019), 27 pages.
https://doi.org/10.1145/3343172

This work has been partially funded by EU project SoBigData RI, grant #654024.
Authors' addresses: L. Pappalardo (corresponding author) and F. Giannotti, ISTI-CNR, Via G. Moruzzi, 1, Pisa, Italy, 56124; emails: {luca.pappalardo, fosca.giannotti}@isti.cnr.it; P. Cintia (corresponding author), P. Ferragina, and D. Pedreschi, Department of Computer Science, University of Pisa, Largo B. Pontecorso 3, Pisa, Italy, 56127; emails: paolo.cintia@isti.cnr.it, {ferragin, pedre}@di.unipi.it; E. Massucco, Wyscout, Chiavari, Italy; email: emanuele.massucco@wyscout.com.

2157-6904/2019/09-ART59
https://doi.org/10.1145/3343172

1 INTRODUCTION

Rankings of soccer players and data-driven evaluations of their performance are becoming more central in the soccer industry [4, 14, 15, 20, 32, 34, 40, 46]. On the one hand, many sports companies, websites, and television broadcasters—such as Opta, WhoScored.com, and Sky, as well as the plethora of online platforms for fantasy football and e-sports—widely use soccer statistics to compare the performance of professional players with the purpose of increasing fan engagement via critical analyses, insights, and scoring patterns. On the other hand, coaches and team managers are interested in analytic tools to support tactical analysis and monitor the quality of their players during individual matches or entire seasons [11, 27]. Not least, soccer scouts and performance analysts are continuously looking for data-driven tools to improve the retrieval of talented players with desired characteristics, based on evaluation criteria that take into account the complexity and the multi-dimensional nature of soccer performance. While selecting talents on the entire space of soccer players is unfeasible (if not impossible!) for humans, data-driven performance scores help select a small subset of the best players who meet specific constraints (e.g., age, performance features and trends, roles). This allows scouts and performance analysts to analyze a smaller set of players, thus saving considerable time and economic resources while broadening scouting operations and career opportunities of talented players.

The problem of data-driven evaluation of player performance is gaining interest in the scientific community, too, thanks to the availability of massive data streams generated by (semi-)automated sensing technologies [4, 10, 15, 28–30, 40, 46, 48], such as the so-called soccer-logs that detail all the spatio-temporal events related to players during a match (e.g., tackles, passes, fouls, shots, dribbles, etc.). Ranking players means defining a relation of order between them with respect to *some measure* of their performance over a sequence of matches. In turn, measuring performance means computing a *data-driven performance rating* that quantifies the quality of a player's performance in a specific match. This is a complex task, since there is no objective and shared definition of performance quality, which is an inherently multidimensional concept [36]. Several data-driven ranking and evaluation algorithms have been proposed in the literature to date, but they suffer from three main limitations.

First, existing approaches are *mono-dimensional*, in the sense that they propose metrics that evaluate the player's performance by focusing on one single aspect (mostly, passes or shots [6, 12, 23, 24, 39]), thus missing to exploit the richness of attached meta-information provided by soccer-logs. Conversely, soccer scouts search for a talented player based on "metrics" that combine many relevant aspects of their performance, from defensive skills to possession and attacking skills. Since mono-dimensional approaches cannot meet this requirement, there is the need for a framework capable to exploit a comprehensive evaluation of performance based on the richness of the meta-information available in soccer-logs. Second, existing approaches evaluate performance without taking into account the specificity of each player's *role* on the field (e.g., right back, left wing), so they compare players that comply with different tasks [6, 12, 23, 24, 39]. Since it is meaningless to compare players who comply with different tasks and considering that a player can change roles from match to match and even within the same match, there is the need for an automatic framework capable of assigning a role to players based on their positions during a match or a fraction of it. Third, missing a gold-standard dataset, existing approaches in the literature report judgments that consist mainly of informal interpretations based on some simplistic metrics (e.g., market value or goals scored [6, 45, 47]). It is important instead to evaluate the goodness of ranking and performance evaluation algorithms in a quantitative and thorough manner, through datasets built with the help of human experts as done for example for the evaluation of recommender systems in Information Retrieval.

This article presents the results of a joint research among academic computer scientists and data scientists of Wyscout,[1] the leading company for soccer scouting. The goal has been to study the limitations of existing approaches and develop PlayeRank, a new-generation data-driven framework for the performance evaluation and the ranking of players in soccer. PlayeRank offers a principled multi-dimensional and role-aware evaluation of the performance of soccer players, driven only by the massive and standardized soccer-logs currently produced by several sports analytics companies (i.e., Wyscout, Opta, Stats). PlayeRank is designed around the orchestration of the solutions to three main phases: a learning phase, a rating phase, and a final ranking phase. PlayeRank models the *performance of a soccer player in a match* as a multidimensional vector of features extracted from soccer-logs. In the learning phase, PlayeRank performs two main sub-tasks: (i) the *extraction of feature weights*: since we do not have a ground-truth for "learning" the mapping from the performance features to the players' performance quality, we turn this problem into a classification problem between the multidimensional vector of features, aggregated over all players' of a *team*, and the result this team achieved in a match; (ii) the *training of a role detector*: given that there are different player roles in soccer, we identify, in an unsupervised way, a set of roles from the match events available in the soccer-logs. In the subsequent rating phase, a player's performance quality in a match is evaluated as the scalar product between the previously computed feature weights and the values these features get in that match played by that player. In the final ranking phase, PlayeRank computes a set of *role-based rankings* for the available players by taking into account their performance ratings and their role(s) as they were computed in the two phases before.

To validate our framework, we instantiated it over a massive dataset of soccer-logs that is unique in the large number of logged matches and players and for the length of the period of observation. In fact, it includes 31M of events covering around 20K matches and 21K players in the last four seasons of 18 soccer competitions: Spanish first division, English first division, Italian first division, German first division, French first division, Portuguese first division, Turkish first division, Greek first division, Austrian first division, Swiss first division, Russian first division, Dutch first division, Argentinian first division, Brazilian first division, European Champions League, Europa League, World Cup 2018, and European Cup 2016. Then we performed an extensive experimental analysis advised by a group of professional soccer scouts that showed that PlayeRank is robust in agreeing with a ranking of players given by these experts, with an improvement up to 30% (relative) and 21% (absolute) with respect to the current state-of-the-art algorithms [6, 12].

One of the main characteristics of PlayeRank is that, by providing a score that meaningfully synthesizes a player's performance quality in a match or in a series of matches, it enables the analysis of the statistical properties of soccer performance. In this regard, the analysis of the performance ratings resulting from PlayeRank, for all the players and all the matches in our dataset, revealed several interesting patterns. First, on the basis of the players' average position during a match, the role detector finds *eight* main roles (Section 4.3) and enables the investigation of the notion of *player's versatility*, defined as his ability to change role from match-to-match (Section 6.2). Second, the analysis of feature weights reveals that there is no significant difference among the 18 competitions, with the only exception of the competitions played by national teams (Section 4.4). Third, the distribution of player ratings changes by role, thus suggesting that the performance of a player in a match highly depends on the zone of the soccer field he is assigned to (Section 4.5). This is an important aspect that will be exploited to design a novel search engine for soccer players (Section 6.1). Fourth, we find that the distribution of performance ratings is strongly peaked around its average, indicating that "outlier" performances are rare (Section 4.5). In particular, these outlier performances are unevenly distributed across the players: While the majority of players achieve

[1] https://wyscout.com/.

a few excellent performances, a tiny fraction of players achieve many excellent performances. Moreover, we find that top players do not always play in an excellent way but, nonetheless, they achieve excellent performances more frequently than the other players (Section 4.5).

In conclusion, our study shows that PlayeRank is an innovative data-driven framework that goes beyond the state-of-the-art results in the evaluation and ranking of soccer players. This study also provides the first thorough, and somewhat surprising, characterization of soccer performance. The last section will start from PlayeRank to present a set of new challenging problems in soccer analytics that we state and comment to stimulate the research interest from the community of data scientists.

The Python source code of PlayeRank can be found at https://github.com/mesosbrodleto/playerank. An online app to interact with the ratings generated by PlayeRank can be found at http://playerank.d4science.org/. A dataset containing the PlayeRank scores for all players and matches for one season of several soccer competitions can be found at https://doi.org/10.6084/m9.figshare.9361148.v1.

2 RELATED WORKS

The availability of massive data portraying soccer performance has facilitated recent advances in soccer analytics. The so-called soccer-logs [4, 15, 40, 46], capturing all the events occurring during a match, are one of the most common data formats and have been used to analyze many aspects of soccer, both at team [8, 11, 25, 35, 50] and individual levels [6, 12, 33]. Among all the open problems in soccer analytics, the data-driven evaluation of a player's performance quality is the most challenging one, given the absence of a ground-truth for that performance evaluation.

Data-driven Evaluation of Performance. While many metrics have been proposed to capture specific aspects of soccer performance (e.g., expected goals, pass accuracy, etc.), just a few approaches evaluate a player's performance quality in a systemic way. The flow centrality (FC) metric proposed by Duch et al. [12], one of the first attempts in this setting, is defined as the fraction of times a player intervenes in pass chains that end in a shot. Based on this metric, they rank all players in UEFA European Championship 2008 and observe that 8 players in their top-20 list belong to the UEFA's top-20 list that was released just after the competition. Being based merely on pass centrality, as the authors themselves highlight in the article, the FC metric mostly makes sense for midfielders and forwards. Brooks et al. [6] develop the Pass Shot Value (PSV), a metric to estimate the importance of a pass for generating a shot. They represent a pass as a vector of 360 features describing the vicinity of a field zone to the pass's origin and destination. Then, they use a supervised machine-learning model to predict whether or not a given pass results in a shot. The feature weights resulting from the model training are used to compute PSV as the sum of the feature weights associated with the pass's origin and destination. They finally used soccer-logs to rank players in the Spanish first division season 2012–'13 according to their average PSV, showing that it correlates with the rankings based on assists and goals. Unfortunately, as the authors highlight in the article, PSV is strongly biased towards offensive-oriented players. Moreover, PSV is a pass-based metric that thus omits all the other kinds of events observed during a soccer match and lacks of a proper validation. Instead of proposing their own algorithm for performance quality evaluation, Nsolo et al. [33] extract performance metrics from soccer-logs to predict the WhoScored.com performance rating with a machine-learning approach. The resulting model is more accurate for specific roles (e.g., forwards) and competitions (e.g., English first division) when predicting if a player is in the top 10%, 25%, or 50% of the WhoScored.com ranking.

The problem of evaluating players' performance has received attention in other team sports, too, such as hockey, basketball, and baseball. In hockey, Schulte and Zhao proposed the Scoring

Impact metric [44] to rank ice hockey players in the NHL, depending on his team's chance of scoring the next goal. In basketball, the Performance Efficiency Rating[2] is a widely used metric to assess players' performance by deploying basketball-logs (e.g., pass completed, shots achieved). In baseball, a plethora of statistical metrics has been proposed to evaluate the performance of players and teams [2].

Rating Systems for Sports Teams. Many studies focus on developing the so-called *rating systems*, such as Elo and TrueSkill [19, 21], which rank teams or players based on their past victories/defeats and the estimated strength of the opponent. Therefore, they do not take into account player-observed match events nor other quantitative aspects of individual and collective performance [35]. As a result, unlike PlayeRank, such rating systems are unable to provide an explicit characterization of the performance of a player as well as to discern his contribution in a match.

Relations between Performance and Market Value. Another strand of literature focuses on quantifying the relation between proxies of a player's quality, such as market value, wage, or popularity, and his performance on the field. Stanojevic and Gyarmati [45] use soccer-logs to infer the relation between a player's typical performance and his market value as estimated by crowds. They find a large discrepancy between estimated and real market values, due to the lack of important information such as injury-proneness and commercialization capacity. Müller et al. [31] develop a similar approach and use soccer-logs, as well as players' popularity data and market values in the preceding years, to estimate a player transfer fee. They show that for the low- and medium-priced players the estimated market values are comparable to estimations by the crowd, while the latter performs better for the high-priced players. Torgler and Schmidt [47] investigate what shapes performance in soccer, represented as a player number of goals and assists. They find that salary, age, and team effects have a statistically significant impact on a player's performance on the field.

Position of Our Work. Despite an increasing interest in this research field, our review of the state-of-the-art highlights that there is no validated framework allowing for a multi-dimensional and role-aware evaluation of soccer performance quality. In this article, we overcome this issue by proposing PlayeRank, a framework that deploys the events described by soccer-logs to evaluate a player's performance quality and a player's role in a match. In contrast to FC and PSV, which lack a proper validation with domain experts, we test the framework against a humanly labeled dataset we have specifically built for the purpose of evaluating soccer players' performance. Finally, and for the first time in the literature, we shed some light on the statistical patterns that characterize soccer player performance by providing a novel and thorough analysis that exploits PlayeRank scores and the large and unique dataset of competitions, teams, and players available to us.

3 THE PLAYERANK FRAMEWORK

Figure 1 shows how the PlayeRank framework operates. It is designed to work with soccer-logs, in which a match consists of a sequence of events encoded as a tuple: $\langle id, type, position, timestamp \rangle$, where *id* is the identifier of the player that originated/refers to this event, *type* is the event type (i.e., passes, shots, goals, tackles, etc.), *position* and *timestamp* denote the spatio-temporal coordinates of the event over the soccer field. PlayeRank assumes that soccer-logs are stored into a database, which is updated with new events after each soccer match (Figure 1(a)).

The key task addressed by PlayeRank is the *"evaluation of the performance quality of a player u in a soccer match m."* This consists of computing a numerical rating $r(u, m)$, called *performance rating*, that aims at capturing the quality of the performance of u in m given *only* the set of events related to that player in that match. This is a complex task, because of the many events observed in

[2] https://www.basketball-reference.com/about/per.html.

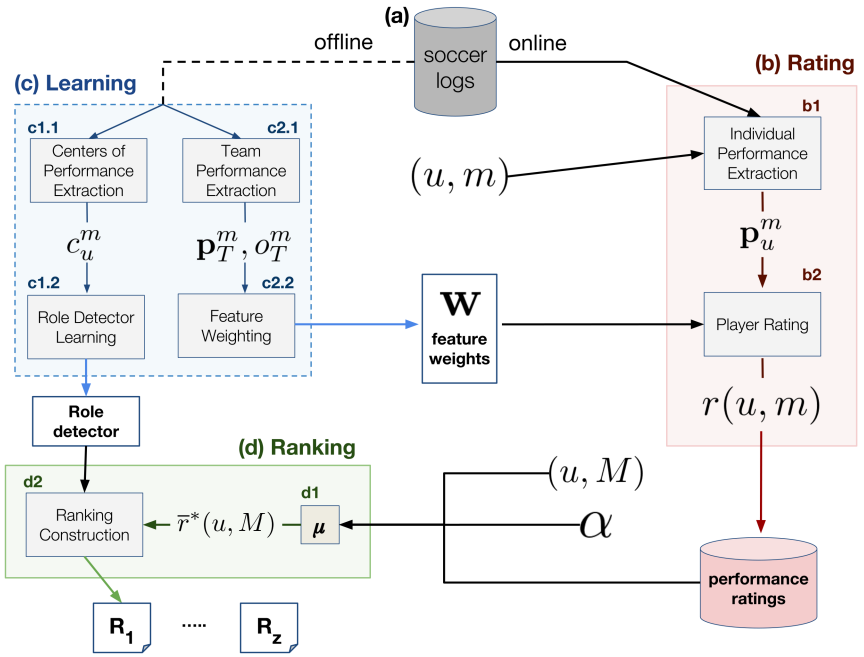


Fig. 1. Schema of the PlayeRank framework. Starting from a database of soccer-logs (a), it consists of three main phases. The learning phase (c) is an "offline" procedure: It must be executed at least once before the other phases, since it generates information used in the other two phases, but then it can be updated separately. The rating (b) and the ranking phases (d) are online procedures, i.e., they are executed every time a new match is available in the database of soccer-logs. We refer to the text for the notation used in the figure.

a match, the interactions among players within the same team or against players of the opponent team, and the fact that players' performance is inextricably bound to the performance of their team and possibly of the opponent team. PlayeRank addresses such complexity by means of a procedure that hinges onto a massive database of soccer-logs and consists of three phases: a rating phase, a ranking phase, and a learning phase.

3.1 Rating Phase

The rating phase (Figure 1, step b) is responsible for the computation of the performance rating $r(u, m)$, and it is run for each player u every time a new match m becomes available in the soccer-logs database. This phase exploits information computed "offline" and consists of two main steps: individual performance extraction (Figure 1, step b1) and player rating (Figure 1, step b2).

Individual Performance Extraction. Given that a match m is represented as a set of events, PlayeRank *models* the performance of a player u in m by means of an n-dimensional feature vector $\mathbf{p}_u^m = [x_1, \ldots, x_n]$, where x_i is a feature that describes a specific aspect of u's behavior in match m and is computed from the set of events played by u in that match. In our experiments in Section 4, we provide an example of $n = 76$ features extracted from our dataset. Some features simply count some events (e.g., number of fouls, number of passes, etc.), some others are at a finer level in that they distinguish the outcome of those events—i.e., if they were "accurate" or "not accurate."

Player Rating. The evaluation of the performance of a player u *in a single match* m is computed as the scalar product between the values of the features referring to match m and the feature weights

$\mathbf{w}$ computed during the learning phase (Figure 1, step c2.2, described in Section 3.3). Each feature weight models the importance of that feature in the evaluation of the performance quality of any player. Formally speaking, given the multi-dimensional vector of features $\mathbf{p}_u^m = [x_1, \ldots, x_n]$ and their weights $\mathbf{w}$, PlayeRank evaluates the performance of a player u in a match m as follows:

$$r(u, m) = \frac{1}{R} \sum_{i=1}^{n} w_i \times x_i. \tag{1}$$

The quantity $r(u, m)$ is called the *performance rating* of u in match m, where R is a normalization constant such that $r(u, m) \in [0, 1]$. We decided to not include the number of goals scored in a match into the set of features (for reasons that are explained in Section 3.3 (learning phase)), but, since goals could themselves be important to evaluate the performance of some (offensive) players, PlayeRank can be adapted to manage goals, too, via an *adjusted-performance rating*, defined as:

$$r^*(u, m) = \alpha \times norm\_goals + (1 - \alpha) \times r(u, m), \tag{2}$$

where *norm_goals* indicates the number of goals scored by u in match m normalized in the range $[0, 1]$, and $\alpha \in [0, 1]$ is a parameter indicating the importance given to goals into the new rating. Clearly, $r^*(u, m) = r(u, m)$ when $\alpha = 0$, and $r^*(u, m) = norm\_goals$ when $\alpha = 1$.

Finally, PlayeRank computes the rating of a *player* u over a series of matches $M = (m_1, \ldots, m_g)$ by aggregating u's ratings over those matches according to a function $\mu(r(u, m_1), \ldots, r(u, m_g))$, which, in this article, is set to the Exponential Weighted Smoothing Average (EWMA). This way, the performance quality of player u after g matches is computed as:

$$\overline{r}(u, M) = \overline{r}(u, m_g) = \beta \times r(u, m_g) + (1 - \beta) \times \overline{r}(u, m_{g-1}), \tag{3}$$

where β is a proper smoothing factor set in the range $[0, 1]$. In other words, the performance quality of player u after g matches, i.e., $\overline{r}(u, m_g)$, is computed as the weighted average of the rating $r(u, m_g)$ reported by u in the last match m_g and the previous smoothed ratings $\overline{r}(u, m_{g-1})$. This way, we are weighting more the recent performances of players. Similarly, the *goal-adjusted rating* $\overline{r}^*(u, m_g)$ of u given a series of g matches is computed as the EWMA of his adjusted performance ratings. The quantity $\overline{r}(u, M)$ is called the *player rating* of player u given M, while $\overline{r}^*(u, M)$ is called the *adjusted-player rating* of player u given M.

Example. Let us assume that performance of soccer players is described by $n = 3$ features – number of passes, number of shots, and number of yellow cards. Then, the performance of the player "Ronaldo" who made in the match "Juve-Roma" 25 passes and 2 shots and got 1 yellow card is described as:

$$\mathbf{p}_{\text{Ronaldo}}^{\text{(Juve-Roma)}} = [\underbrace{25}_{\text{passes}}, \underbrace{2}_{\text{shots}}, \underbrace{1}_{\text{cards}}].$$

Let us assume now that the feature weights computed during the learning phase are

$$\mathbf{w} = [\underbrace{0.05}_{\text{passes}}, \underbrace{0.5}_{\text{shots}}, \underbrace{-0.2}_{\text{passes}}],$$

so yellow cards are weighted negatively, whereas shots are weighted positively and five times more than passes, and that R is the normalization factor. Then, the performance rating of player "Ronaldo" in match "Juve-Roma" is computed as:

$$r(\text{Ronaldo, Juve-Roma}) = \frac{\overbrace{(25 \times 0.05)}^{\text{passes}} + \overbrace{(2 \times 0.5)}^{\text{shots}} + \overbrace{(1 \times -0.2)}^{\text{cards}}}{R} = \frac{2.05}{R}.$$

3.2 Ranking Phase

Based on the *players ratings* computed in the previous phase, PlayeRank constructs a set of *role-based rankings* $R_1, \ldots, R_z$, each corresponding to one of the z roles identified by a *role detector* (Figure 1, step c1.2, described in Section 3.3), an algorithm previously trained during the learning phase that assigns to one or more roles each player u in a match m. PlayeRank assigns a player u to R_i if he has at least x% of the matches in M assigned to role i, where x is a parameter chosen by the user. In our experiments in Section 4, we select $x = 40\%$, a choice dictated by the fact that arguably a soccer player may be assigned to at most two roles. Experiments showed this threshold is robust; however, this parameter can be chosen by the user when running PlayeRank, possibly increasing the number of assigned roles per player. Depending on the value of the threshold x, a player can appear in more than one ranking and with different ranks, since they depend on $\overline{r}(u, M)$.

3.3 Learning Phase

The learning phase (Figure 1, step c) is executed "offline" and generates information used in the rating and the ranking phases. It consists of two steps: feature weighting and role detector training.

Feature Weighting. Performance evaluation is a difficult task, because we do not have an objective evaluation of the performance $\mathbf{p}_u^m$ of each player u. This technically means that we do not have a ground-truth dataset to learn a *relation* between performance features and performance quality of u in match m. However, the outcome of a match may be considered a natural proxy for evaluating performance quality at *team* level. Therefore, we overcome that limitation by proposing a *supervised* approach: We determine the impact of the n chosen features onto a player performance by looking in turn at the team-wise contribution of these features to the match outcome.

This idea is motivated by the fact that (i) a team's ultimate purpose in a match is to win by scoring one goal more than the opponent and (ii) some actions of players during a match have a higher impact on the chances of winning a match than others. For example, making a pass that puts a teammate in position to score a goal (assist) is intuitively more valuable than making a pass to a close teammate in the middle of the field. Conversely, getting a red card is intuitively less valuable than, say, winning a dribble against an opponent. Therefore, those actions that strongly increase (or decrease) the chances of winning a match must be weighted more during the evaluation, either positively or negatively. While soccer practitioners and fans have in mind an idea of what the most and the least valuable actions during a match are, it is important to develop a data-driven and automatic procedure that quantifies how valuable an action is with respect to increasing or decreasing the chances of winning a match.

PlayeRank implements this syllogism via a two-phase approach. In the first phase (Figure 1, step c2.1) it extracts the performance vector $\mathbf{p}_T^m$ of team T in match m and the outcome o_T^m of that match: where $o_T^m = 1$ indicates a victory for team T in match m and $o_T^m = 0$ indicates a non-victory (i.e., a defeat or a draw) for T. The team performance vector $\mathbf{p}_T^m = [x_1^{(T)}, \ldots, x_n^{(T)}]$ is obtained by summing the corresponding features over all the players U_T^m composing team T in match m:

$$\mathbf{p}_T^m[i] = \sum_{u \in U_T^m} \mathbf{p}_u^m[i].$$

In the second phase (Figure 1, step 2.2), PlayeRank solves a classification problem between the team performance vector $\mathbf{p}_T^m$ and the outcome o_T^m. This classification problem has been shown in Reference [35] to be meaningful, because there is a strong relation between the team performance vector and the match outcome. We use a linear classifier, such as the Linear Support Vector Classifier (LSVC), to solve the previous classification problem, and then we extract from the classifier the weights $\mathbf{w} = [w_1, \ldots, w_n]$ that quantify the influence of the features to the outcomes of soccer

matches, as explained above. These weights are then used in the rating phase (Figure 1, step b2) to compute the performance ratings of players.

Role Detector Training. As pointed out in References [38, 44], performance ratings are meaningful only when comparing players with similar *roles*. In soccer, each role corresponds to a different area of the playing field where a player is assigned responsibility relative to his teammates [3]. Different roles imply different tasks, hence it is meaningless to compare, for example, a player that is asked to create goal occasions and a player that is asked to prevent the opponents to score. Furthermore, a role is not a unique label, as a player's area of responsibility can change from one match to another and even within the same match. Given these premises, we decided to design and implement an algorithm able to detect the role associated with a player's performance in a match based on the soccer-logs. We observe that there do exist methods, originally designed for hockey [44], that compute the roles of players via an affinity clustering applied over a heatmap describing their presence in predefined zones of the field. But these approaches are arguably not effective in soccer, because it offers a lower density of match events w.r.t. hockey. Nonetheless we experimented and discarded the approach of Reference [44], because it produces on our dataset a clustering with a very low quality (i.e., silhouette score $ss < 0.2$).

Conversely, PlayeRank detects a player's role in a match by looking at his *average position*. This is motivated by the fact that a player's role is often defined as the position covered by the player relative to his teammates [3]. This is called the *center of performance* for u in m, and it is denoted as $\mathbf{c}_u^m = (\overline{x}_u^m, \overline{y}_u^m)$, where $\overline{x}_u^m$ and $\overline{y}_u^m$ are the average coordinates of u's events in match m, as they are extracted from the soccer-logs (step c1.1, Figure 1). Then PlayeRank deploys a k-means algorithm [18] to group the centers of performance of all players in all matches (step 1.2, Figure 1).

PlayeRank also accounts for the possibility of having "hybrid" roles where a center of performance is assigned to two or more clusters. This is useful in situations where the center of performance of a player u is between two or more clusters, and so the role of u in match m cannot be well characterized by just one single cluster. Therefore, PlayeRank aims at a finer classification of roles via a *soft clustering*. For every center of performance $\mathbf{c}_u^m$ occurring in some cluster C_i, PlayeRank computes its k-silhouette $s_k(\mathbf{c}_u^m)$ with respect to every other cluster C_k ($k \neq i$) as:

$$s_k\left(\mathbf{c}_u^m\right) = \frac{\overline{d_k}(\mathbf{c}_u^m) - \overline{d_i}(\mathbf{c}_u^m)}{max(\overline{d_i}(\mathbf{c}_u^m), \overline{d_k}(\mathbf{c}_u^m))}, \tag{4}$$

where $\overline{d_z}(\mathbf{c}_u^m)$ is the average distance between $\mathbf{c}_u^m$ and all other points in cluster C_z. PlayeRank assigns $\mathbf{c}_u^m$ to *every* cluster C_j for which $s_j(\mathbf{c}_u^m) \leq \delta_s$, where δ_s is a threshold indicating the tolerance to "hybrid" centers. If no such j does exist, $\mathbf{c}_u^m$ is assigned to the cluster C_i given by the partitioning computed by the k-means algorithm.

For the sake of completeness, we mention that in approaching the task of role classification, we have considered other, more sophisticated modeling of players' performance such as heatmaps (as in Reference [44]; see comments above) or events direction (as in Reference [3]). However, we found that the resulting clusters were of lower quality in terms of the silhouette score. Note that the PlayeRank works with any type of role classification. It is indeed sufficient to change the role detector module with the preferred algorithm or static role labelling; the remainder of the framework works the same.

4 EXPERIMENTAL RESULTS

We implemented the PlayeRank framework and executed it on a massive database of soccer-logs provided by the company Wyscout. In this section, we show experiments for each of the modules described in Section 3 and depicted in Figure 1.

Table 1. For Each Competition, We Describe the Corresponding Geographic Area, Total Number of Seasons, Matches, Events, and Players

competition	seasons	matches	events	players
Spanish first division	4	1,520	2,541,873	1,264
English first division	4	1,520	2,595,808	1,231
Italian first division	4	1,520	2,610,908	1499
German first division	4	1,124	2,075,483	1,042
French first division	4	1,520	2,592,708	1,288
Portuguese first division	4	1,124	1,720,393	1,227
Turkish first division	4	1,124	1,927,416	1,182
Greek first division	4	1,060	1,596,695	1,151
Austrian first division	4	720	1,162,696	593
Swiss first division	4	720	1,124,630	647
Russian first division	4	960	1,593,703	1,046
Dutch first division	4	1,248	2,021,164	1,177
Argentinian first division	4	1,538	2,450,170	1,870
Brazilian first division	4	1,437	2,326,690	1,790
European Champions League	3	653	995,363	3,577
Europa League	3	1,416	1,980,733	9,100
European Cup 2016	1	51	78,140	552
World Cup 2018	1	64	101,759	736
	64	19,619	31,496,332	(*)21,361

The dataset covers 18 competitions, for a total of 64 soccer seasons and around 20K matches, 31M events, and 21K players. (*) 21,361 indicates the number of distinct players, as some players play with their teams in both national and continental/international competitions.

4.1 Soccer-logs Database

We use a database of soccer-logs provided by Wyscout consisting of 31,496,332 events, describing 19,619 matches, 296 clubs, and 21,361 players of several seasons of 18 prominent competitions around the world (see Table 1): Spanish first division, English first division, Italian first division, German first division, French first division, Portuguese first division, Turkish first division, Greek first division, Austrian first division, Swiss first division, Russian first division, Dutch first division, Argentinian first division, Brazilian first division, European Champions League, Europa League, World Cup 2018, and European Cup 2016.

Each event records: (i) a unique event identifier; (ii) the type of the event; (iii) a time-stamp; (iv) the player related to the event; (v) the team of the player; (vi) the match in which the event is observed; (vii) the position on the soccer field, specified by a pair of integers in the range $[0, 100]$ indicating the percentage from the left corner of the attacking team; (viii) the event subtype and a list of tags, which enrich the event with additional information (see Table 3). We do not consider the *goalkeeping* events available from the Wyscout APIs, as we discard goalkeepers from the analysis.[3] Table 2 shows an example of an event in the Italian first division, corresponding to an accurate pass by player 3,344 (Rafinha) of team 3,161 (Internazionale) made at second 2.41 in the first half of match 2,576,335 (Lazio—Internazionale) started at position (49, 50) of the field. Figure 2 shows a pictorial representation of the events produced by player Lionel Messi during a match in the Spanish first division, where each event is drawn at the position of the field where it has occurred.

[3]Goalkeepers would need a dedicated analysis, since it is the only role having different game rules w.r.t. to all other players.

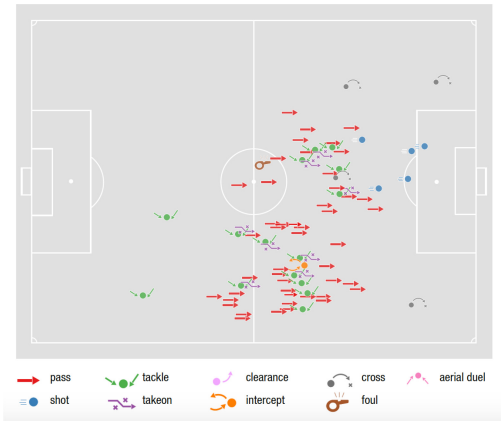


Fig. 2. Example of events observed for Lionel Messi (FC Barcelona) during a match in the Spanish first division, season 2015/2016. Each event is shown on the field at the position where it has occurred with a marker indicating the type of the event.

Table 2. Example of Event in the Dataset Corresponding to an Accurate Pass by Player 3,344 (Rafinha) of Team 3,161 (Internazionale) Made at Second 2.41 of Match 2,576,335 (Lazio—Internazionale) Started at Position (49, 50) of the Field


```
{"id": 253668302,
"eventName": "Pass",
"eventSec": 2.41,
"playerId": 3344,
"matchId": 2576335,
"teamId": 3161,
"positions":
     [{"x": 49, "y": 50}],
"subEventId": 85,
"subEventName": "Simple pass",
"tags":
     [{"id": 1801}]}
```


Table 3. Event Types with Their Possible Subtypes and Tags

type	subtype	tags
pass	cross, simple pass	accurate, not accurate, key pass, opportunity, assist, (goal)
foul		no card, yellow, red, 2nd yellow
shot		accurate, not accurate, block, opportunity, assist, (goal)
duel	air duel, dribbles, tackles, ground loose ball	accurate, not accurate
free kick	corner, shot, goal kick, throw in, penalty, simple kick	accurate, not accurate, key pass, opportunity, assist, (goal)
offside	acceleration, clearance, simple touch	counter attack, dangerous ball lost, missed
touch		ball, interception, opportunity, assist, (goal)

For further details, see the Wyscout API documentation at: https://apidocs.wyscout.com/.

In our dataset, a match consists of an average of about 1,600 events, and for each player there are about 57 observed events per match (Figures 3(a)–(b)), with an average inter-time between two consecutive events of 3.45s (Figure 3(c)). Passes are the most frequent events, accounting for around 50% of the total events (Figure 3(d)). Wyscout soccer-logs adhere to a standard format for storing events collected by semi-automatic systems [15, 40, 46] and do not include off-ball actions. Moreover, given the existing literature on the analysis of soccer matches [6, 8, 16, 17, 35, 36], we can state that the dataset we use in our experiments is unique in the large number of events, matches and players considered, and for the length of the period of observation.

4.2 Performance Extraction

We compute the players' performance vectors by a two-step procedure: First, we define a feature for every possible combination of type, subtype, and tag shown in Table 3. For example, given

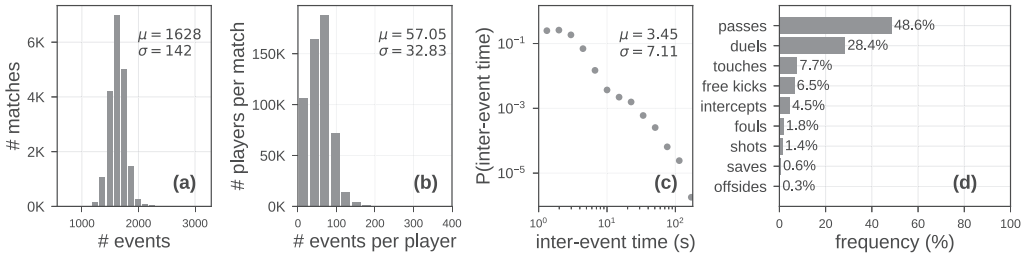


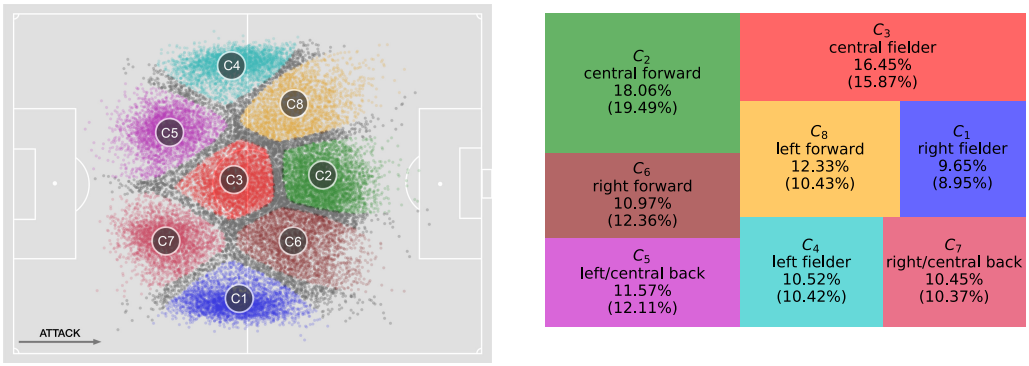
Fig. 3. (a) Distribution of the number of events per match (μ = average, σ = st. deviation). On average, a match has 1,628 events. (b) Distribution of the number of events produced by one player per match. On average, a player generates around 57 events per match. (c) Distribution of inter-event times, defined as the time (in seconds) between two consecutive events in a match. On average, there are around three seconds between an event and the next one in a match. (d) Frequency of events per type. Passes are the most frequent event, accounting for about 48% of the events in a match.

the *foul* type, we obtain four features: *foul no card*, *foul yellow*, *foul red,* and *foul 2nd yellow*. We discard the *goal* tag, since we have implicitly considered the goals as the outcome of a performance during the learning phase. Nevertheless, goals can be still included in the performance rating by Equation (2) in Section 3.1. Eventually, we extracted 76 features from the Wyscout soccer-logs and normalized them in the range $[0, 1]$ to guarantee that all features are expressed in the same scale (see Table 7 for a list of all the features). We tried more sophisticated features by considering the field zones where events have occurred or the fraction of the match when they have occurred, but we did not find any significant difference w.r.t. the results presented below. We finally build the performance vector $\mathbf{p}_u^m$ for a player u in match m by counting the number of events of a given type, subtype, and tag combination observed for u in m. For example, the number of fouls without card made by u in m compose the value of feature *foul no card* of u in m.

4.3 Role Detection

To discover roles, we execute the role-detection algorithm of Section 3.3 by varying $k = 1, \ldots, 20$ and specifying $\delta_s = 0.1$, which implies that 5% of the centers are classified as hybrids.[4] We observe that $k = 8$ provides the best clustering in terms of silhouette score ($ss = 0.43$) and that these results are stable across several executions of the experiment where different sets of centroids are used to initialize the k-means algorithm. Figure 4(a) shows the result of the 8-means clustering. We asked professional soccer scouts, employed by Wyscout, to provide an interpretation of the 8 clusters with terms suitable for soccer practitioners. An explanation for the clusters C1–C8, as well as a set of players typically in each role, are provided in Table 4. It is worth noting that, while there are 10 players in a team (excluding the goalkeeper), the clustering algorithm detected 8 roles. This means that there is at least one cluster (i.e., role) having more than one player in each team. Moreover the correspondence with classic roles is not perfect in that two players classified in two different classic roles can appear in the same cluster, and vice versa. Figure 4(b) shows how the performances and the players are distributed among the roles, where each player is assigned to the role he covers most frequently during the available seasons. We find that role C_2 (central forward) is the most common role, covering 18% of performances and 19% of players, followed by role C_3 (central fielder), covering 16% of performances and 15% of players. All other roles are almost equally populated.

[4]The experiments showed that the number of hybrid centers increases linearly with δ_s, from none to all centers.



(a) (b)

Fig. 4. (a) Grouping of the centers of performance in the clusters $C_1, \ldots, C_8$. Each color identifies a different cluster (role); grey points indicate hybrid centers of performance. Table 4 shows an interpretation of clusters given by professional soccer scouts. (b) Distribution of the 8 roles discovered by the role detector across performances and players (in parenthesis) within our dataset. Each player is assigned to the role he covers most frequently during the available seasons.

Table 4. Interpretation of the 8 Clusters Detected and Examples of Players Assigned to Each Cluster

	cluster	description	examples
C1	right fielder	plays on the right as a wing, back, or both	S. Roberto, Danilo
C2	central forward	plays in the center, close to the opponent's goal	Messi, Suárez
C3	central fielder	plays in the center	Kroos, Pjanić
C4	left fielder	plays on the left as a wing, back, or both	Nolito, J. Alba
C5	left central back	plays close to his own goal, on the left	Bartra, Maguire
C6	right forward	plays on the right, close to the opponent's area	Robben, Dembélé
C7	right central back	plays close to his own goal, on the right	J. Martínez, Matip
C8	left forward	plays on the left, close to the opponent's area	Neymar, Insigne

4.4 Feature Weighting

As discussed in Section 3.3, PlayeRank turns the problem of estimating the 76 feature weights into a classification problem between a team performance vector and a match outcome. We instantiate this problem by creating, for each match m, two examples $\mathbf{p}^m_{T_1}$ and $\mathbf{p}^m_{T_2}$, which correspond to the performance vectors of two playing teams T_1 and T_2, and the match outcome label o^m_T that is 1 if a team wins and 0 otherwise. The resulting dataset consists of 19,619 examples, 80% of which are used to train a Linear Support Vector Classifier (LSVC). We have selected the cost parameters that had the maximum average Area Under the Receiver Operating Characteristic Curve (AUC) on a 5-fold cross-validation. We validate LCSV on the remaining 20% of the examples, finding an $AUC = 0.89$, $F1 = 0.81$, accuracy= 0.82.[5] This result is significantly better than the predictive results of two baseline classifiers: (i) a classifier that always predicts the most frequent match outcome (i.e., non-victory, $AUC = 0.50$, $F1 = 0.48$, accuracy $= 0.62$); (ii) a classifier that chooses the label at random based on the distribution of victories and non-victories ($AUC = 0.50$, $F1 = 0.53$,

[5]The AUC is the expectation that a uniformly drawn random positive is ranked before a uniformly drawn random negative. The accuracy is the number of correct predictions over the total. The F1 score is the harmonic average of precision and recall (precision is the number of correct positive results divided by the number of the classifier's positive results; recall is the number of correct positive results over the total positive results).

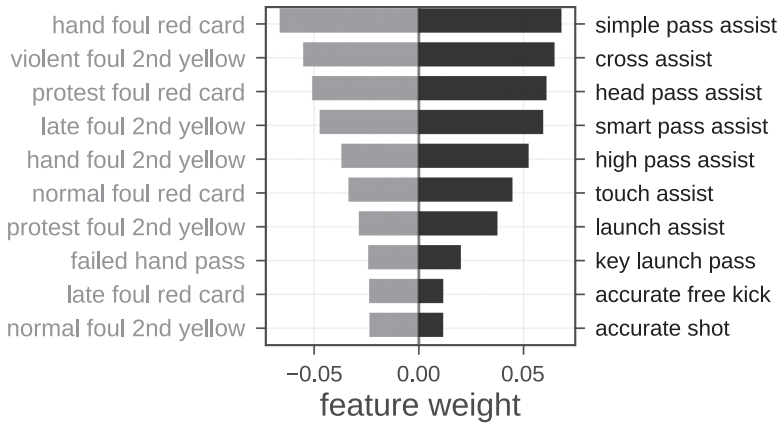


Fig. 5. Top-10 (black bars, on the right) and bottom-10 (grey bars, on the left) features according to the value of the weights extracted from the LSVC trained on all competitions together.

accuracy = 0.53). We also experimented with different labelling of o_T^m by defining either $o_T^m = 0$ in the case of defeat and $o_T^m = 1$ otherwise, or by defining a ternary classification problem where $o_T^m = 1$ indicates a victory, $o_T^m = 0$ a defeat, and $o_T^m = 2$ a draw. In all these cases, we did not find any significant difference in the feature weights described below, so we chose to deploy the binary classification problem above.

Figure 5 shows the top-10 (black bars) and the bottom-10 (grey bars) feature weights $\mathbf{w} = [w_1, \ldots, w_n]$ resulting from LSVC (see Appendix B for details). We find that assist-based features are the most important ones, followed by the number of key passes and the accuracy of shots. In contrast, getting a red/yellow card gets a strong negative weight, especially for hand and violent fouls. It is interesting to notice that, though these choices are pretty natural for who is skilled in soccer-player evaluations, PlayeRank derived them automatically by just looking at the massive soccer-logs in our dataset.

For the sake of completeness of our experimental results, we also repeated the classification task separately: (i) competition-by-competition, i.e., we created 18 LSVCs, each one trained on the matches of one competition only; (ii) role by role, i.e., we created 8 LSVCs, each one trained on the examples created from players tagged with one role only (Section 4.3).

Competition-based Weights. We extracted for each of the 18 competitions the corresponding set of weights $\mathbf{w}^{(j)} = [w_1^{(j)}, \ldots, w_n^{(j)}]$ $(j = 1, \ldots, 18)$ and quantified the difference between the weights $\mathbf{w}$ extracted from all competitions and the $\mathbf{w}^{(j)}$s via the Normalized Root-Mean-Square Error:

$$NRMSE(\mathbf{w}, \mathbf{w}^{(j)}) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(w_i - w_i^{(j)})^2}}{\max \mathbf{w} - \min \mathbf{w}}, \tag{5}$$

where $\max \mathbf{w}$ and $\min \mathbf{w}$ are the maximum and the minimum weights in $\mathbf{w}$, respectively. We found that the average *NRMSE* is around 6% and that 16 out of 18 competitions have *NRMSE* < 7% (Figure 6(b)), indicating that the difference between any $\mathbf{w}^{(j)}$ and $\mathbf{w}$ is small, and hence the relation between team performance and match outcome is in most of the cases independent of the specific competition for clubs considered. Only for competitions involving national teams, such as the European Cup 2016 and the World Cup 2018, the *NRMSE* is higher, 17% and 20%, respectively (Figure 6(b)). This can be due either to the fact that these two competitions have a few matches (51 and 64, respectively; see Table 1) or that while all the other competitions refer to soccer clubs,

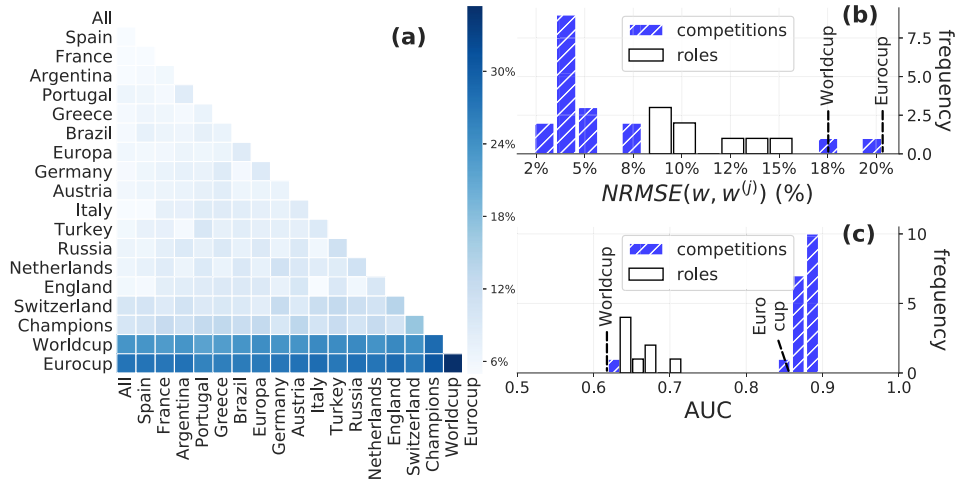


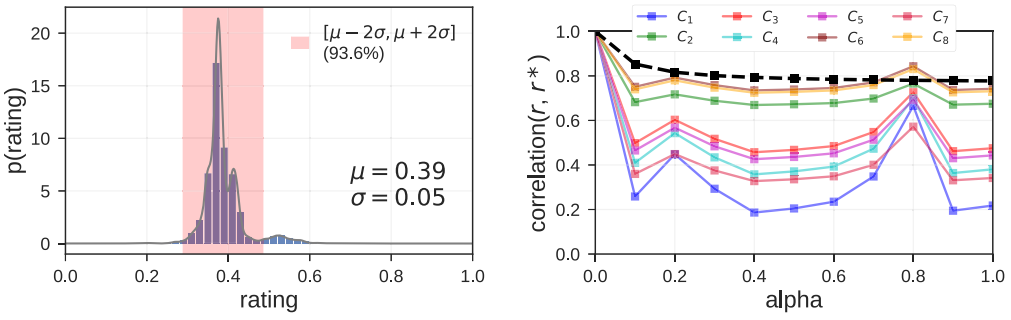
Fig. 6. (a) Heatmap indicating the Normalized Root Mean Squared Error (NRMSE) between the set of feature weights $\mathbf{w}^{(j)}$ of each competition j and the overall set of feature weights $\mathbf{w}$. (b) Distribution of $NRMSE(\mathbf{w}, \mathbf{w}^{(j)})$, expressed in percentage, indicating the normalized error between $\mathbf{w}$ and a competition's set of weights (blue bars) and a role's set of weights (white bars). (c) Distribution of AUC of the LSVCs trained on the 18 competitions separately (blue bars) and the 8 roles separately (white bars).

the European Cup 2016 and the World Cup 2018 are competitions for national teams, which are generally more unpredictable [8, 9]. Figure 6(c) indicates that the accuracy of the LSVC trained on the 64 matches of the World Cup 2018 is lower than the accuracy of the other models, suggesting that the number of matches in a competition influences the accuracy. However, the accuracy of the LSVC model trained on the European Cup 2016 is close to the accuracy of all other models, suggesting that the difference in the weights can be also due to the specific nature of the competition.

Role-based Weights. We repeated the classification task separately role-by-role by aggregating the players' feature role-by-role. We found that: (i) the accuracy of the LSVCs trained on the roles separately are lower than the accuracy of the models trained on the competitions, though the role-based model's accuracy is still higher than the model trained on the World Cup 2018 (Figure 6(c)); (ii) the *NRMSE* between each role's set of weights and the set of weights trained on all competitions together is lower than 15% (Figure 6(b)). This indicates that there is a small variation between the competition-based and the role-based sets of weights. For this reason, we will just use $\mathbf{w}$ (i.e., the set of weights computed at match level including all competitions) in the computation of the ratings in the following sections.

4.5 Player Ratings and Rankings

Given $\mathbf{w}$, we compute the performance rating $r(u, m)$ for each player u in each match m and then explore their distribution. As Figure 7(a) shows, the distribution is strongly peaked around its average ($\mu = 0.39$), indicating that "outlier" performances (i.e., $r(u, m) \notin [\mu - 2\sigma, \mu + 2\sigma]$, σ is the standard deviation) are rare. In particular, excellent performances (i.e., $r(u, m) > \mu + 2\sigma$), accounting for just 5% of the total, are unevenly distributed across the players. Indeed, the probability density function of the number of excellent performances per player is decreasing and long-tailed (Figure 9, ALL): While the majority of players achieve a few excellent performances, a tiny fraction of players achieve up to 40 excellent performances during the five years. This trend is observed also when we split performances by the player's role, highlighting the presence of a general pattern (Figure 9, C_1–C_8).



(a) (b)

Fig. 7. (a) Distribution of performance ratings. It is strongly peaked around the average $\mu = 0.39$, while outliers are rare. Most of the ratings (≈94%) are within the range $[\mu - 2\sigma, \mu + 2\sigma]$ (σ is the standard deviation). (b) Correlation between player ratings and adjusted-player ratings as α varies in the range $[0, 1]$. The dashed curve refers to all players together, the solid to the 8 roles.

As an example, let us consider all performances of role C_8 (left forward): Most of the players achieve excellence just once, while a few players achieve as many as 30 (Neymar, 21% of his performances), 16 (L. Insigne, 14%), and 15 (E. Hazard, 10%) excellent performances. Moreover, we find that a correlation exists between a player's average performance rating and the variability of his ratings (Figure 10): The stronger a player is (i.e., the higher his average performance rating), the more variable his performance ratings are (i.e., the higher is the standard deviation of his ratings). In other words, the best players do not play excellence in every match, they just achieve excellence more frequently than the other players. Taken together, Figures 7(a), 9, and 10 indicate that: (i) excellent performances are rare (≈5% of the total); (ii) just 11% of the players achieve excellent performances at least once; (iii) while a small set of players repeatedly achieve excellence, all other players do a few times, suggesting that the best players do not always play excellently but they just achieve it more frequently; (iv) excellent performances are at most 21% (Neymar) and on average 9% of all players reach excellence at least once.

By aggregating the performance rating of each player u over the whole series M of matches, we compute the player rating $\overline{r}(u, M)$. Figure 8 visualizes the distribution of these player ratings by grouping the players on the x-axis according to their roles. We recall that we are assigning a player to a role if he plays at least 40% of the matches in that role, meaning that a player may be assigned to at most 2 roles among the 8 roles detected. We observe a different distribution of ratings according to the players' roles, both in terms of range of values and their concentration. This fully justifies the design of the role-detection module in the PlayeRank framework. In fact, we notice that the top-ranked player of cluster C_4, Marcelo, gets a player rating that is below the average of the ratings of clusters C_6 or C_8 (Figure 8).

Table 5 reports the top-10 players grouped by the 8 roles. Although PlayeRank is fully data-driven, it is able to place the most popular players at the top of some ranking. For example, Lionel Messi (Barcelona) is the best player in cluster C_6 (see Figure 4), followed by other renowned players such as Thomas Müller (Bayern Munich) and Mohamed Salah (Liverpool). Instead, the best player in cluster C_2 (central forward) is Luís Suárez (Barcelona), preceding Cristiano Ronaldo (Juventus), Jonas (Benfica), and Benzema (Real Madrid). Other renowned players are at the top of their role's ranking, such as Neymar (PSG, C_8, left forward) and Marcelo (Real Madrid, C_4, left-fielder).

What it is surprising in these role-based rankings is that they have been derived by PlayeRank without considering the number of goals scored by players when building the performance

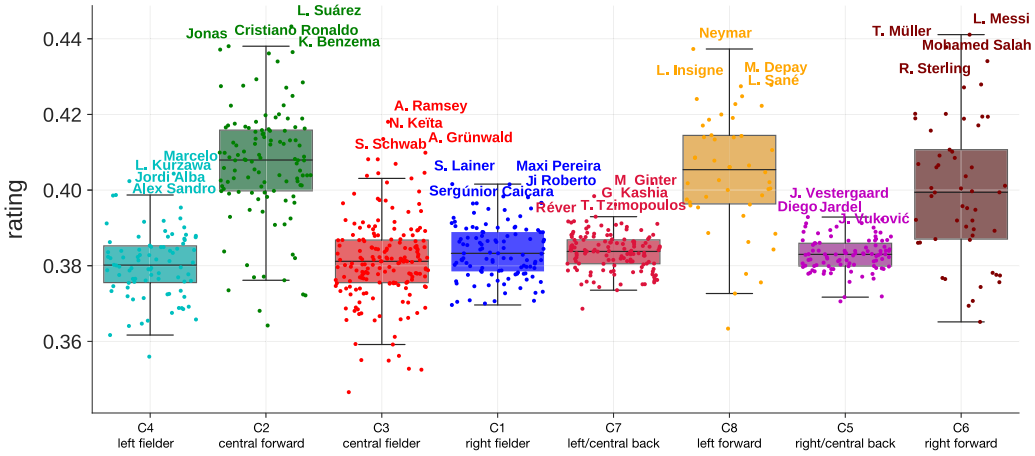


Fig. 8. Distribution of player ratings per role. Each boxplot represents a cluster (role) and each point (circle) indicates a player's rating, computed across all the performances in the last four seasons of the 18 competitions. The points are jittered by adding random noise to the x-axis value to help visualization. For each cluster, the player's name at the top of the corresponding role-based rankings are shown.

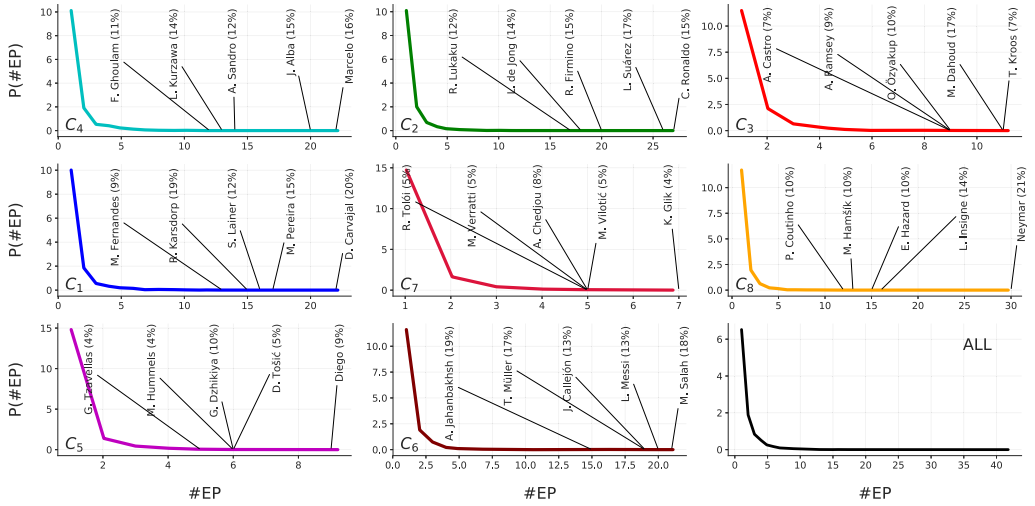


Fig. 9. Probability density function of the number of excellent performances #EP (i.e., $r(u,m) > \mu + 2\sigma$) per player, for each role ($C_1, \ldots, C_8$) and for all roles together (ALL). In each plot, we show the players who achieve the top-5 performances in the corresponding role. We observe that the probability density functions are decreasing and long-tailed, meaning that while the majority of players achieve a few excellent performances, a tiny fraction of players achieve up to dozens of excellent performances during the five years.

vector. Actually, we observe that in general the goal-adjusted ranking $\overline{r}^*(u,M)$ is consistent with $\overline{r}(u,M)$ for all values of α (Equation (2)): As the black dashed curve in Figure 7(b) shows, the correlation between the player rating and the adjusted-player rating slightly decreases with α, with values that are in general $\geq$0.8. However, when investigating how the correlation changes with α role-by-role, we find that while offensive-oriented roles such as C_2 (central forward), C_6 (right forward), and C_8 (left forward) show in general high correlations between those ratings

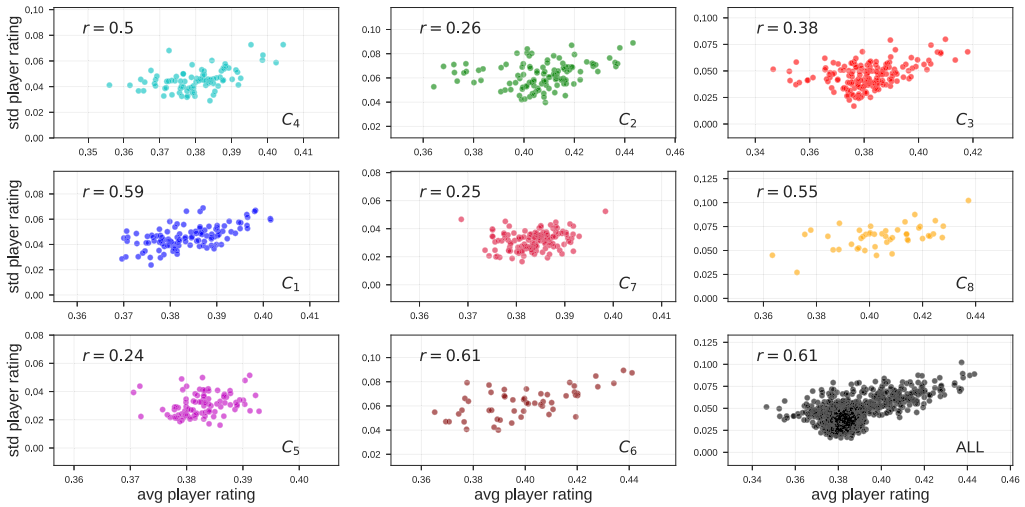


Fig. 10. Correlation between a player's average performance rating and his standard deviation for each role ($C_1, \ldots, C_8$) and all roles together (*ALL*). Here r indicates the Pearson correlation coefficient.

($\in [0.65, 0.85]$), roles C_3 (central-fielder), C_5 (left central back), and C_4 (left-fielder) show moderate correlations ($\in [0.40, 0.75]$); while role C_1 (right-fielder) shows low correlation ($\in [0.20, 0.65]$). This result suggests that the player rating of offensive players is not much influenced by the number of goals scored, presumably because they are already associated with events related to scoring.

5 VALIDATION OF PLAYERANK

Existing player-ranking approaches report judgments that consist mainly of informal interpretations based on some simplistic metrics (e.g., market value or goals scored [6, 45, 47]). It is important instead to evaluate the goodness of ranking and performance evaluation algorithms in a quantitative manner, through the help of human experts, as done for example for the evaluation of recommender systems in information retrieval.

We validated PlayeRank by creating and submitting a survey to three professional soccer talent scouts, employed by Wyscout, hence particularly skilled at evaluating and comparing soccer players. Our survey consisted of a set of pairs of players randomly generated by a two-step procedure, defined as follows: First, we randomly selected 35% of the players in the dataset. Second, for each selected player u, we cyclically iterated over the ranges $[1, 10]$, $[11, 20]$, and $[21, \infty]$ and selected one value, say x, for each of these ranges, and then picked the player being x positions above u and the one being x positions below u in the role-based ranking (if they exist). This generated a set P of 211 pairs involving 202 distinct players.

For each pair $(u_1, u_2) \in P$, each scout was asked to select the best player between u_1 and u_2 or to specify that the two players were equally valuable. For each such pair, we also computed the best player according to PlayeRank by declaring u_1 stronger than u_2 if u_1 precedes u_2 in the ranking. We discarded from P all pairs for which there is not a majority among the evaluations of the experts: namely, either all experts expressed equality or two experts disagreed in judging the best player and the third one expressed equality. As a result of this process, we discarded 8% of P's pairs.

Over the remaining P's pairs, we investigated two types of concordance among the scouts' evaluations: (i) the *majority concordance* c_{maj} defined as the fraction of the pairs for which PlayeRank agrees with at least two scouts; (ii) the *unanimous concordance* c_{una} defined as the

Table 5. Top-10 Players in Each Role-based Ranking, with the Corresponding Player Rating (r) Computed Across the Last Four Seasons of the 18 Competitions

	r	player	club		r	player	club		r	player	club
cluster 4 - left fielder	.404	Marcelo	R. Madrid	**cluster 2** - central foward	.404	L. Suárez	Barcelona	**cluster 3** - central fielder	.404	A. Ramsey	Arsenal
	.402	L. Kurzawa	PSG		.402	C. Ronaldo	Juventus		.402	N. Keïta	Leipzig
	.399	A. Sandro	Juventus		.399	Jonas	Benfica		.399	A. Grünwald	A. Wien
	.399	J. Alba	Barcelona		.399	K. Benzema	R. Madrid		.399	S. Schwab	R. Wien
	.395	J. Willems	Eintracht F.		.395	D. Mertens	Napoli		.395	van Overeem	AZ
	.393	D. Alaba	Bayern M.		.393	L. de Jong	PSV		.393	O. Özyakup	Beşiktaş
	.392	M. Alonso	Chelsea		.392	S. Agüero	Man City		.392	A. Dzagoev	CSKA M.
	.392	B. Davies	Tottenham		.392	S. Heung-Min	Tottenham		.392	C. Tolisso	Bayern M.
	.391	D. Kombarov	Spartak M.		.391	D. Alli	Tottenham		.391	Nainggolan	Roma
	.390	J. Brenet	PSV		.390	A. Dzyuba	Zenit		.390	B. N'Diaye	Stoke City
cluster 1 - right fielder	.402	S. Lainer	Salzburg	**cluster 7** - central back	.402	M. Ginter	Borussia M.	**cluster 8** - left forward	.402	Neymar	PSG
	.402	M. Pereira	Porto		.402	G. Kashia	Vitesse		.402	M. Depay	O. Lyonnais
	.398	S. Roberto	Barcelona		.398	Réver	Flamengo		.398	L. Insigne	Napoli
	.398	J. Caiçara	Istanbul B.		.398	Tzimopoulos	PAS		.398	L. Sané	Man City
	.397	D. Carvajal	R. Madrid		.397	M. Yumlu	Akhisar		.397	M. Hamšík	Napoli
	.397	L. De Silvestri	Torino		.397	Hilton	Montpellier		.397	M. Dabbur	Salzburg
	.397	R. Pereira	Leicester		.397	Alderweireld	Tottenham		.397	E. Hazard	Chelsea
	.396	D. Caligiuri	Schalke		.396	Bruno Silva	Cruzeiro		.396	P. Coutinho	Barcelona
	.395	N. Skubic	Konyaspor		.395	Y. Ayhan	MKE		.395	I. Perišić	Inter
	.395	S. Widmer	Udinese		.395	J. Schunke	Estudiantes		.395	Isco	R. Madrid
cluster 5 - central back	.393	J. Vestergaard	Southampton	**cluster 6** - right forward	.393	L. Messi	Barcelona				
	.392	Jardel	Benfica		.392	T. Müller	Bayern M.				
	.391	J. Vuković	H. Verona		.391	M. Salah	Liverpool				
	.391	Diego	Antalyaspor		.391	R. Sterling	Man City				
	.390	Raúl Silva	S. Braga		.390	G. Bale	R. Madrid				
	.390	D. Siovas	Leganés		.390	S. Mané	Liverpool				
	.390	M. Hummels	Bayern M.		.390	K. Bellarabi	Bayer L.				
	.389	C. Lema	Belgrano		.389	B. Traoré	O. Lyonnais				
	.389	L. Perrin	S. Étienne		.389	G. Martins	A. Madrid				
	.389	S. Ignashevich	CSKA M.		.389	A. Candreva	Inter				

The club indicated in the table is the one the player played with at the end of 2018.

fraction of pairs for which the scouts' choices are unanimous and PlayeRank agrees with them. We found that $c_{\text{maj}} = 68\%$ and $c_{\text{una}} = 74\%$, indicating that PlayeRank has in general a good agreement with the soccer scouts, compared to the random choice (for which $c_{\text{maj}} = c_{\text{una}} = 50\%$). Figure 11 offers a more detailed view on the results of the survey by specializing c_{maj} and c_{una} on the three ranges of ranking differences: $[1, 10]$, $[11, 20]$, $[21, \infty]$. The bars show a clear and strong correlation between the concordance among scouts' evaluations (per majority or unanimity) and the difference between the positions in the ranking of the checked pairs of players: When the ranking difference is $\leq$10 it is $c_{\text{maj}} = 59\%$ and $c_{\text{una}} = 61\%$; for larger and larger ranking differences, PlayeRank achieves a much higher concordance with experts, which is up to $c_{\text{maj}} = 86\%$ and $c_{\text{una}} = 91\%$ when the ranking difference is $\geq$20. Clearly, the disagreement between PlayeRank with the soccer scouts is less significant when the players are close in the ranking (i.e., their distance <10). Indeed, the comparison between soccer players is a well-known difficult problem,

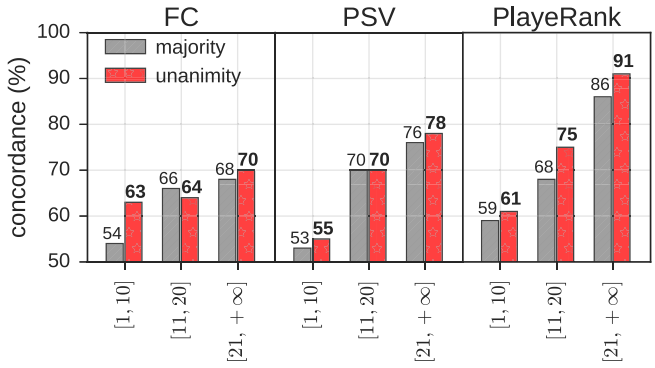


Fig. 11. Majority (grey bars) and unanimity (red) concordance between Flow Centrality and the scouts (left), PSV and the scouts (center), PlayeRank and the scouts (right).

as witnessed by the significant increase in the fraction of unanimous answers by the scouts, which goes from a low 58% in the range $[1, 10]$ to a reasonable 71% in the range $[21, +\infty]$. This *a fortiori* highlights the robustness of PlayeRank: The scouts' disagreement decreases as pairs of players are farther and farther in the ranking provided by PlayeRank.

As a final investigation, we compared PlayeRank with the Flow Centrality (FC) [12] and the PSV [6] metrics, which constitute the current state-of-the-art in soccer-player ranking (see Section 2). These metrics are somewhat *mono-dimensional*, because they exploit just passes or shots to derive the final ranking. Figure 11 (right) shows the results obtained by FC and PSV over our set of players' pairs evaluated by the three Wyscout experts. It is evident that FC and PSV achieve significantly lower concordance than PlayeRank with the experts: for PSV, the majority concordance ranges from 53% to 76%, while the unanimity concordance ranges from 55% to 78%; for FC, the majority concordance ranges from 54% to 68%, while the unanimity concordance ranges from 63% to 70%. So, PlayeRank introduces an improvement that is up to 16% (relative) and 13% (absolute) with respect to PSV, and an improvement of 30% (relative) and 21% (absolute) with respect to FC.

6 APPLICATIONS

To demonstrate its usefulness, in this section, we show two examples of analytical services that can be designed using PlayeRank: the retrieval of players in a database of soccer-logs and the computation of the players' versatility.

6.1 Retrieval of Players

One of the most useful applications of PlayeRank is searching players in a soccer-logs database. The search is driven by a query formulated in terms of a suitable *query language* that considers the events occurring during a match and their position on the field. Since we do not want to enter in the formal definition of the *full* query language, which is beyond the scope of this article, we concentrate here only on its specialties that are the most interesting *algorithmically* for the issues we have discussed in this article.

We propose the efficient solution of a *spatial query* over the soccer-field zones that possibly span more roles and have geometric forms that differ from the ones identified by the role detector. We assume a tessellation of the soccer field into h zones of equal size $z_1, \ldots, z_h$. The query is modeled as a vector $Q = [q_1, \ldots, q_h]$ in which q_i expresses *how much relevant* is the presence of the searched player in zone z_i. Similarly, player u is modeled as a vector $V_u = [u_1, \ldots, u_h]$ in which u_i expresses *how much inclined* is player u to play in zone z_i. We can go from binary vectors,

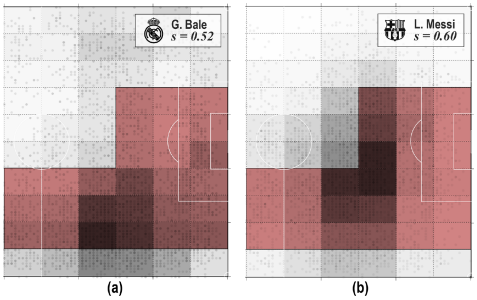


Fig. 12. Visualization of a spatial query Q_1 (red area) and the heatmaps of presence of G. Bale of Real Madrid (a) and L. Messi of Barcelona (b). The darker a zone, the higher a player's propensity to play in it.

Table 6. Top-10 Players According to Their $z(u, M, Q_1)$ with Respect to Query Q_1 in Figure 12, Computed on the Last Four Seasons of the Italian, Spanish, German, and English First Divisions

	player	z	s	$\overline{r}$	club
1	L. Messi	**.28**	.60	**.46**	Barcelona
2	A. Robben	.26	**.61**	.43	Bayern M.
5	M. Salah	.24	.56	.43	Liverpool
3	L. Suárez	.24	.54	.45	Barcelona
4	T. Müller	.24	.56	.43	Bayern M.
6	R. Lukaku	.24	.56	.42	Man. Utd
7	A. Petagna	.23	.55	.42	Atalanta
8	D. Berardi	.22	.54	.41	Sassuolo
9	Aduriz	.22	.55	.40	A. Bilbao
10	G. Bale	.22	.52	.43	R. Madrid

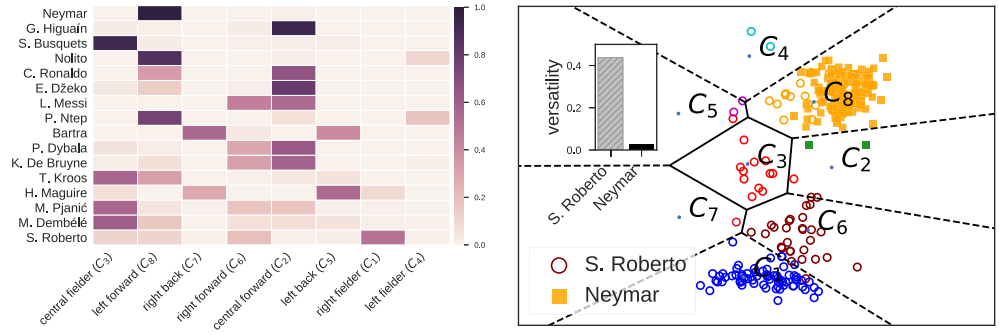
which model interest/no interest for Q and presence/no presence for V_u, to the more sophisticated case in which Q expresses a weighted interest for some specific zones and V_u is finely modeled by counting; for example, the number of events played by u in each zone z_i. Now, given a query Q and the players in the soccer-logs database, the goal is to design an algorithm that evaluates the *propensity* of players to play in the field zones specified by Q. We follow the standard practice of Information Retrieval (IR) and compute for each player u the dot product $s(u, Q) = V_u \cdot Q$. We can efficiently compute this product by means of one of the plethora of solutions known in the IR literature (see, e.g., References [26, 41]). In this respect, we point out that known solutions work efficiently over a million (and more) dimensions, so they *easily* scale to the problem size at hand, because $h \approx 10^6$ if we would assume zones z_i of size 1 cm^2!

Finally, PlayeRank ranks players according to their *rating* over a series of matches and their *propensity* to play in the queried zones by sorting players in decreasing order of the score $z(u, M, Q) = s(u, Q) \times \overline{r}(u, M)$, where $s(u, Q)$ is the dot product between Q and the player vector V_u, and $\overline{r}(u, M)$ is u's player rating over a series of matches. Note that the function $z(u, M, Q)$ could be defined in many other ways; for example, by weighting $s(u, Q)$ and $\overline{r}(u, M)$ differently to better capture the user's needs. Other combinations will be investigated in the future. For the sake of presentation, we consider here a tessellation of the soccer field into 100 equal-sized zones and, thus, define a query Q as a binary vector of 100 components that express the interest of the user about the "presence in a zone" for the searched players. PlayeRank computes $s(u, Q)$ as the dot product between Q and the player vector V_u, and $\overline{r}(u, M)$ as the player rating over all matches of u. Players are ranked in decreasing order of the quantity $z(u, M, Q) = s(u, Q) * \overline{r}(u, M)$ as described above.

Table 6 shows the top-10 players in a portion of our dataset according to their $z(u, M, Q_1)$ for an exemplar query Q_1 showed in Figure 12. Lionel Messi, whose heatmap of positions is drawn in Figure 12(b), has the highest $z(u, M, Q_1)$. In the table, it is interesting to note that, though the vector of Arjen Robben is more similar to Q_1 ($s(Robben, Q_1) = 0.61$) than Messi's vector ($s(Messi, Q_1) = 0.60$), Messi has a higher player rating ($\overline{r}(Messi, M) = 0.46$, $\overline{r}(Robben, M) = 0.43$). As a result, the combination $z(u, M, Q_1)$ of the two quantities makes Messi the player offering the best trade-off between matching with the user-specified zones and performing well in those zones.

6.2 Versatility

The role detector of PlayeRank enables the analysis of an important aspect of a player's behavior: his *versatility*, which we define as a player's propensity to change role from match-to-match. To



(a) (b)

Fig. 13. Versatility of soccer players. (a) Heatmap showing the frequency of top players to play in the 8 roles (the darker a cell, the higher the frequency). The players are sorted from the least versatile (Neymar) to the most versatile (Sergi Roberto). (b) Positions of centers of performance of Sergi Roberto (circles) and Neymar (squares). Each center of performance is colored according to the role assigned by the role-detection algorithm.

investigate this aspect, we define the versatility of a player as the Shannon entropy of his roles in a series of matches M:

$$V(u,M) = -\frac{\sum_{i=1}^{k} p(u,M)_i \log p(u,M)_i}{\log k}, \tag{6}$$

where $k = 8$ and $p(u,M)_i$ is the probability of player u of playing in role i, computed as the ratio of the number of matches in M in which u played in role i.

Figure 13(a) displays the frequency $p(u,M)_i$ of playing in a role i for a set of top soccer players. We observe that many players have a high versatility, i.e., they play in different roles across different matches. In particular, Sergi Roberto (FC Barcelona) and Neymar (PSG FC) are among the most-versatile and the least-versatile players, respectively. Figure 13(b) visualizes all the centers of performance of Sergi Roberto and Neymar, coloring the centers according to the role assigned by the role detector. We observe that Neymar's centers of performance are concentrated in just one role (C_8, left forward) while Sergi Roberto's centers are scattered around the field, indicating that he plays in all 8 roles, witnessing a high versatility. Numerically, V(Sergi Roberto) $= 0.45$ and V(Neymar) $= 0.016$. The versatility of a player is an important property to take into account when composing a club's roster. PlayeRank embeds versatility within its analytic framework, allowing soccer practitioners and scouts to evaluate the flexibility of a player as well as his playing quality in an automatic way.

7 CONCLUSIONS AND FUTURE WORKS

In this article, we presented PlayeRank, a data-driven framework that offers a multi-dimensional and role-aware evaluation of the performance of soccer players. Our extensive experimental evaluation on a massive database of soccer-logs—18 competitions, 31M events, 21K players—showed that the rankings offered by PlayeRank outperform existing approaches in being significantly more concordant with professional soccer scouts. Moreover, our experiments showed several interesting results, shedding light on novel patterns that characterize the performance of soccer players. Indeed, we found that excellent performances are rare and unevenly distributed, since a few top players produce most of the observed excellent performances. An interesting result is also that top players do not always play excellently, they just achieve excellent performances more frequently

than the other players. Regarding the extraction of feature weights, we found that the difference between the weights extracted from each competition separately is small (i.e., <10%) with the only exception of the Euro Cup and the World Cup for which that difference is slightly higher (i.e., ≈20%), thus highlighting the different nature of competitions for national teams. Last, our role detector found 8 main roles in soccer that we also exploited to investigate the versatility of players, an entropy-based measure that indicates the ability of a player to change role from match-to-match.

PlayeRank is a valuable tool to support professional soccer scouts in evaluating, searching, ranking, and recommending soccer players. We wish to highlight here that, given its modularity, PlayeRank can be extended and customized in several ways. First, more sophisticated algorithms could be designed to detect a player's role during a match or fraction of a match. These algorithms could then be easily embedded in the PlayeRank's architecture, giving the user the possibility to customize role detection according to their needs. Some innovative AI-based solutions to role detection, which we plan to embed into PlayeRank, have been proposed during a Soccer Data Challenge[6] recently organized by Wyscout and the European research infrastructure SoBigData. A similar reasoning applies to the feature-weighting module: As soon as more sophisticated techniques are proposed to weight performance features, they could be embedded in PlayeRank's architecture.

Another direction to improve PlayeRank is to make it able to work with different data sources. In its current version, PlayeRank is based on soccer-logs only, a standard data format describing all ball-touches that occur during a match [15, 35]. Unfortunately, out-of-possession movements are not described in soccer-logs, making it difficult to assess important aspects such as pressing [1] or the ability to create spaces [5]. PlayeRank can be easily extended by making the individual performance extraction module able to extract features from other data sources such as video tracking data [15] and GPS data [42, 43], which provide a detailed description of the spatio-temporal trajectories generated by players during a match.

Finally, it would be interesting to investigate the flexibility of PlayeRank's architecture by plugging it into new performance metrics that will be proposed in the literature; as well as to evaluate its applicability to other team sports, such as basketball, hockey, or rugby, for which data are available in a format that is similar to that of soccer-logs [7, 15, 40, 46].

APPENDICES

A PERFORMANCE FEATURES

Table 7 shows the list of features used in our experiments. Note that PlayeRank is designed to work with any set of features, thus giving to the user a high flexibility about the description of performance. If other features are available from different data sources, describing for example physiological aspects of performance, they can be added into the framework. Section 5 shows that the proposed set of features is powerful enough to make PlayeRank outperform existing approaches in being more concordant with professional soccer scouts.

B EXTRACTION OF WEIGHTS FROM THE LSVC

To classify the outcome of a match given the two teams' performance vectors, we use a Linear Support Vector Classifier (LSVC) [13]. Given a set of instance-label pairs (x_i, y_i), with $i = 1, \ldots, l$, $x_i \in \mathbb{R}^n$, and $y_i \in \{-1, +1\}$, an LSVC solves an unconstrained optimization problem with a loss function ξ:

$$\min_{w} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi(\mathbf{w}; \mathbf{x}_i, y_i)$$

[6] https://sobigdata-soccerchallenge.it/.

Table 7. List of the 76 Features Extracted from the Soccer-logs Database and Used in Our Experiments

type	feature	type	feature
duel	duel-air duel-accurate	others on the ball	others on the ball-acceleration-accurate
	duel-air duel-not accurate		others on the ball-acceleration-not accurate
	duel-ground attacking duel-accurate		others on the ball-clearance-accurate
	duel-ground attacking duel-not accurate		others on the ball-clearance-not accurate
	duel-ground defending duel-accurate		others on the ball-touch-assist
	duel-ground defending duel-not accurate		others on the ball-touch-counter attack
	duel-ground loose ball duel-accurate		others on the ball-touch-dangerous ball lost
	duel-ground loose ball duel-not accurate		others on the ball-touch-feint
foul	foul-hand foul-red card		others on the ball-touch-interception
	foul-hand foul-second yellow card		others on the ball-touch-missed ball
	foul-hand foul-yellow card		others on the ball-touch-opportunity
	foul-late card foul-yellow card	pass	pass-cross pass-accurate
	foul-normal foul-red card		pass-cross pass-assist
	foul-normal foul-second yellow card		pass-cross pass-key pass
	foul-normal foul-yellow card		pass-cross pass-not accurate
	foul-out of game foul-red card		pass-hand pass-accurate
	foul-out of game foul-second yellow card		pass-hand pass-not accurate
	foul-out of game foul-yellow card		pass-head pass-accurate
	foul-protest foul-red card		pass-head pass-assist
	foul-protest foul-second yellow card		pass-head pass-key pass
	foul-protest foul-yellow card		pass-head pass-not accurate
	foul-simulation foul-second yellow card		pass-high pass-accurate
	foul-simulation foul-yellow card		pass-high pass-assist
	foul-violent foul-red card		pass-high pass-key pass
	foul-violent foul-second yellow card		pass-high pass-not accurate
	foul-violent foul-yellow card		pass-launch pass-accurate
free kick	free kick-corner free kick-accurate		pass-launch pass-assist
	free kick-corner free kick-not accurate		pass-launch pass-key pass
	free kick-cross free kick-accurate		pass-launch pass-not accurate
	free kick-cross free kick-not accurate		pass-simple pass-accurate
	free kick-normal free kick-accurate		pass-simple pass-assist
	free kick-normal free kick-not accurate		pass-simple pass-key pass
	free kick-penalty free kick-not accurate		pass-simple pass-not accurate
	free kick-shot free kick-accurate		pass-smart pass-accurate
	free kick-shot free kick-not accurate		pass-smart pass-assist
	free kick-throw in free kick-accurate		pass-smart pass-key pass
	free kick-throw in free kick-not accurate		pass-smart pass-not accurate
		shot	shot-shot-accurate
			shot-shot-not accurate

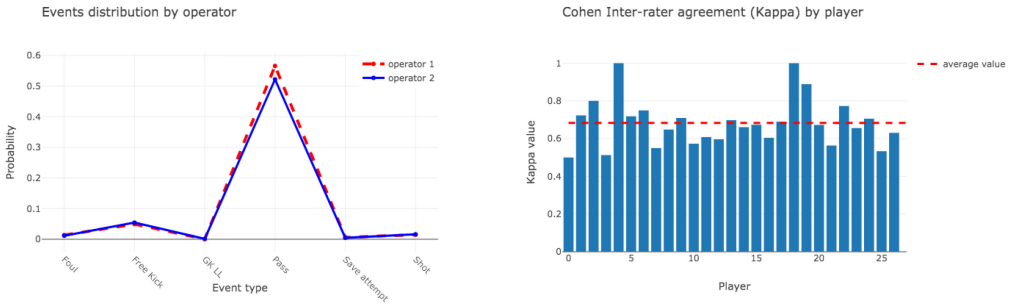
The loss function we use in our experiments is the L2-SVM defined as: $\xi(\mathbf{w}; \mathbf{x}_i, y_i) = \max(1 - y_i \mathbf{w}^T x_i, 0)^2$. Our feature weights are the coefficients $\mathbf{w}$ computed by the LSVC, one per each feature in the vector x_i. In practice, we use the `svm.LinearSVC` object[7] provided by the Python library scikit-learn [37] to train the LSVC and extract the feature weights.

[7]The documentation about the `svm.LinearSVC` class is available at https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html.

C RELIABILITY OF DATA COLLECTION

To assess the reliability of the soccer-logs collection made by Wyscout, we replicate the experiment proposed by Liu et al. [22]: We ask two independent operators to generate the soccer-logs for a match through the Wyscout proprietary tagging software, and then we investigate the agreement between the two sets of soccer-logs. In particular, we compute the so-called Inter-Rater Agreement Rate (IRAR) [22], which indicates the degree of agreement of the way the two operators generate the soccer-logs. The IRAR is computed as $k = 1 - \frac{1-p_o}{1-p_e}$, where p_o is the relative agreement among operators (i.e., the number of the examples over the total that the operators detect the same event) and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each operator randomly seeing each event type [22].

We find an overall IRAR of 0.70 that can be considered a good agreement between the two operators [49]. Figure 14(a) shows, for some event types, the distribution of the number of events per type collected by the two operators. We observe that the two distributions are almost identical, meaning that the two operators collect the data in the same manner. Figure 14(b) shows the IRAR of the two operators computed on the soccer-logs of each player separately. There is agreement between the operators if, when one operator records an event e performed by player p at time t, the other operator records the same event with a timestamp t' close (+ or − 2s) to t. From Figure 14(b), we note that the average IRAR per player is high (around 0.70), denoting a good agreement at the player level, too. In summary, we find a good inter-operator reliability, similar to that guaranteed by other soccer-logs providers.



(a) (b)

Fig. 14. (a) Events distribution for the two operators involved in the experiment. The probability distributions of observing an event of a given type are almost identical. GK LL stays for event "Goalkeeping leaving line." (b) Inter-Rater Agreement Rate (IRAR) for each player. Two operators are in agreement on a player if, for an event e performed by player p observed by one operator at time t, there is an equivalent event e performed by player p observed by the other operator at time $t' \simeq t$.

REFERENCES

[1] Gennady Andrienko, Natalia Andrienko, Guido Budziak, Jason Dykes, Georg Fuchs, Tatiana von Landesberger, and Hendrik Weber. 2017. Visual analysis of pressure in football. *Data Mining Knowl. Disc.* 31, 6 (01 Nov. 2017), 1793–1839. DOI: https://doi.org/10.1007/s10618-017-0513-2

[2] Benjamin Baumer and Andrew Zimbalist. 2014. *The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball.* University of Pennsylvania Press.

[3] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews. 2014. Large-scale analysis of soccer matches using spatiotemporal tracking data. In *Proceedings of the IEEE International Conference on Data Mining*. 725–730. DOI: https://doi.org/10.1109/ICDM.2014.133

[4] Luke Bornn, Dan Cervone, and Javier Fernandez. 2018. Soccer analytics: Unravelling the complexity of the beautiful game. *Significance* 15, 3 (2018), 26–29. DOI: https://doi.org/10.1111/j.1740-9713.2018.01146.x

[5] Luke Bornn and Javier Fernandez. 2018. Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *Proceedings of the MIT Sloan Sports Analytics Conference*.

[6] Joel Brooks, Matthew Kerr, and John Guttag. 2016. Developing a data-driven player ranking in soccer using predictive model weights. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 49–55.

[7] Paolo Cintia, Michele Coscia, and Luca Pappalardo. 2016. The Haka network: Evaluating rugby team performance with dynamic graph analysis. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'16)*. 1095–1102. DOI: https://doi.org/10.1109/ASONAM.2016.7752377

[8] Paolo Cintia, Fosca Giannotti, Luca Pappalardo, Dino Pedreschi, and Marco Malvaldi. 2015. The harsh rule of the goals: Data-driven performance indicators for football teams. In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics*. DOI: https://doi.org/10.1109/DSAA.2015.7344823

[9] Paolo Cintia, Salvatore Rinzivillo, and Luca Pappalardo. 2015. Network-based measures for predicting the outcomes of football games. In *Proceedings of the 2nd Workshop on Machine Learning and Data Mining for Sports Analytics co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'15)*. 46–54. Retrieved from http://ceur-ws.org/Vol-1970/paper-07.pdf.

[10] Varuna De Silva, Mike Caine, James Skinner, Safak Dogan, Ahmet Kondoz, Tilson Peter, Elliott Axtell, Matt Birnie, and Ben Smith. 2018. Player tracking data analytics as a tool for physical performance management in football: A case study from Chelsea football club academy. *Sports* 6, 4 (2018). DOI: https://doi.org/10.3390/sports6040130

[11] Tom Decroos, Jan Van Haaren, and Jesse Davis. 2018. Automatic discovery of tactics in spatio-temporal soccer match data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'18)*. ACM, New York, NY, 223–232. DOI: https://doi.org/10.1145/3219819.3219832

[12] Jordi Duch, Joshua S. Waitzman, and Luís A. Nunes Amaral. 2010. Quantifying the performance of individual players in a team activity. *PLOS ONE* 5, 6 (2010), 1–7. DOI: https://doi.org/10.1371/journal.pone.0010937

[13] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.* 9 (June 2008), 1871–1874. Retrieved from http://dl.acm.org/citation.cfm?id=1390681.1442794.

[14] Bill Gerrard. 2017. *Analytics, Technology and High Performance Sport*. Routledge, London, UK.

[15] Joachim Gudmundsson and Michael Horton. 2017. Spatio-temporal analysis of team sports. *Comput. Surveys* 50, 2 (2017), 22:1–22:34. DOI: https://doi.org/10.1145/3054132

[16] László Gyarmati and Mohamed Hefeeda. 2016. Analyzing in-game movements of soccer players at scale. In *Proceedings of the MIT SLOAN Sports Analytics Conference*.

[17] László Gyarmati, Haewoon Kwak, and Pablo Rodriguez. 2014. Searching for a unique style in soccer. *CoRR* abs/1409.0308 (2014).

[18] J. A. Hartigan and M. A. Wong. 1979. A k-means clustering algorithm. *JSTOR: Appl. Stat.* 28, 1 (1979), 100–108.

[19] Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill^{TM}: A Bayesian skill rating system. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*. 569–576.

[20] A. Jayal, A. McRobert, G. Oatley, and P. O'Donoghue. 2018. *Sports Analytics*. Routledge, London, UK.

[21] Jan Lasek, Zoltán Szlávik, and Sandjai Bhulai. 2013. The predictive power of ranking systems in association football. *J. Appl. Pattern Recog.* 1, 1 (2013). DOI: https://doi.org/10.1504/IJAPR.2013.052339

[22] Hongyou Liu, Will Hopkins, A. Miguel Gómez, and S. Javier Molinuevo. 2013. Inter-operator reliability of live football match statistics from OPTA sportsdata. *Int. J. Perf. Anal. Sport* 13, 3 (2013), 803–821. DOI: https://doi.org/10.1080/24748668.2013.11868690

[23] J. López Peña and H. Touchette. 2012. A network theory analysis of football strategies. *ArXiv e-print* arxiv:math.CO/1206.6904.

[24] Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, and Iain Matthews. 2014. quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In *Proceedings of the 8th MIT Sloan Sports Analytics Conference*. 1–9.

[25] Patrick Lucey, Dean Oliver, Peter Carr, Joe Roth, and Iain Matthews. 2013. Assessing team strategy using spatiotemporal data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1366–1374. DOI: https://doi.org/10.1145/2487575.2488191

[26] Christofer D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

[27] Daniel Memmert, Bischof Jurgen, Stefan Endler, Andreas Grunz, Schmid Markus, Schmidt Andrea, and Perl Jurgen. 2011. *World-Level Analysis in Top Level Football Analysis and Simulation of Football Specific Group Tactics by Means of Adaptive Neural Networks*. IntechOpen, London, UK. DOI: https://doi.org/10.5772/14919

[28] Daniel Memmert, Koen A. P. M. Lemmink, and Jaime Sampaio. 2017. Current approaches to tactical performance analyses in soccer using position data. *Sports Med.* 47, 1 (01 Jan. 2017), 1–10. DOI: https://doi.org/10.1007/s40279-016-0562-5

[29] D. Memmert and D. Raabe. 2018. *Data Analytics in Football: Positional Data Collection, Modelling and Analysis*. Taylor & Francis. Retrieved from https://books.google.it/books?id=O9tdDwAAQBAJ.
[30] Elia Morgulev, Ofer H. Azar, and Ronnie Lidor. 2018. Sports analytics and the big-data era. *Int. J. Data Sci. Anal.* 5, 4 (01 June 2018), 213–222. DOI: https://doi.org/10.1007/s41060-017-0093-7
[31] Oliver Müller, Alexander Simons, and Markus Weinmann. 2017. Beyond crowd judgments: Data-driven estimation of market value in association football. *Euro. J. Op. Res.* 263, 2 (2017), 611–624. DOI: https://doi.org/10.1016/j.ejor.2017.05.005
[32] Maria Nibali. 2016. *The Data Game: Analyzing Our Way to Better Sport Performance*. Routledge, London, UK.
[33] Edward Nsolo, Patrick Lambrix, and Niklas Carlsson. 2018. Player valuation in European football. In *Proceedings of the Machine Learning and Data Mining for Sports Analytics workshop (MLSA'18)*.
[34] P. O'Donoghue and L. Holmes. 2015. *Data Analysis in Sport*. Routledge, London, UK.
[35] Luca Pappalardo and Paolo Cintia. 2017. Quantifying the relation between performance and success in soccer. *Adv. Complex Syst.* 20, 4 (2017). DOI: https://doi.org/10.1142/S021952591750014X
[36] Luca Pappalardo, Paolo Cintia, Dino Pedreschi, Fosca Giannotti, and Albert-Laszlo Barabasi. 2017. Human perception of performance. *ArXiv preprint arXiv:1712.02224* (2017).
[37] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (Nov. 2011), 2825–2830. Retrieved from http://dl.acm.org/citation.cfm?id=1953048.2078195.
[38] S. Pettigrew. 2015. Assessing the offensive productivity of NHL players using in-game win probabilities. In *Proceedings of the MIT Sloan Sports Analytics Conference*.
[39] Paul Power, Héctor Ruiz, Xinyu Wei, and Patrick Lucey. 2017. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1605–1613. DOI: https://doi.org/10.1145/3097983.3098051
[40] Robert Rein and Daniel Memmert. 2016. Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus* 5, 1 (2016), 1410. DOI: https://doi.org/10.1186/s40064-016-3108-2
[41] Berthier Ribeiro-Neto and Ricardo Baeza-Yates. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA.
[42] Alessio Rossi, Luca Pappalardo, Paolo Cintia, Javier Fernández, Marcello Fedon Iaia, and Daniel Medina. 2017. Who is going to get hurt? Predicting injuries in professional soccer. In *Proceedings of the MLSA@PKDD/ECML*.
[43] Alessio Rossi, Luca Pappalardo, Paolo Cintia, F. Marcello Iaia, Javier Fernández, and Daniel Medina. 2018. Effective injury forecasting in soccer with GPS training data and machine learning. *PLOS One* 13, 7 (07 2018), 1–15. DOI: https://doi.org/10.1371/journal.pone.0201264
[44] Oliver Shulte and Zeyu Zhao. 2017. Apples-to-apples: Clustering and ranking NHL players using location information and scoring impact. In *Proceedings of the MIT Sloan Sports Analytics Conference*.
[45] R. Stanojevic and L. Gyarmati. 2016. Towards data-driven football player assessment. In *Proceedings of the IEEE 16th International Conference on Data Mining*. 167–172.
[46] Manuel Stein, Halldór Janetzko, Daniel Seebacher, Alexander Jäger, Manuel Nagel, Jürgen Hölsch, Sven Kosub, Tobias Schreck, Daniel A. Keim, and Michael Grossniklaus. 2017. How to make sense of team sport data: From acquisition to data modeling and research aspects. *Data* 2, 1 (2017). DOI: https://doi.org/10.3390/data2010002
[47] Benno Torgler and Sascha L. Schmidt. 2007. What shapes player performance in soccer? Empirical findings from a panel analysis. *Appl. Econ.* 39, 18 (2007), 2355–2369. DOI: https://doi.org/10.1080/00036840600660739
[48] Bruno Travassos, Keith Davids, Duarte Araújo, and T. Pedro Esteves. 2013. Performance analysis in team sports: Advances from an ecological dynamics approach. *Int. J. Perf. Anal. Sport* 13, 1 (2013), 83–95. DOI: https://doi.org/10.1080/24748668.2013.11868633
[49] Anthony J. Viera and Joanne Mills Garrett. 2005. Understanding interobserver agreement: The kappa statistic. *Family Med.* 37, 5 (2005), 360–3.
[50] Qing Wang, Hengshu Zhu, Wei Hu, Zhiyong Shen, and Yuan Yao. 2015. Discerning tactical patterns for professional soccer teams: An enhanced topic model with applications. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2197–2206. DOI: https://doi.org/10.1145/2783258.2788577

Received January 2019; revised May 2019; accepted July 2019