# Characterizing Soccer Players' Playing Style from Match Event Streams

Aron Geerts, Tom Decroos, and Jesse Davis

KU Leuven, Department of Computer Science

**Abstract.** Transfer fees for soccer players are at an all-time high. To make the most of their budget, soccer clubs need to understand the type of players they have and the type of players that are on the market. Current insights in the playing style of players are mostly based on the opinions of human soccer experts such as trainers and scouts. Unfortunately, their opinions are inherently subjective and thus prone to faults. In this paper, we characterize the playing style of a player in a more rigorous, objective and data-driven manner. We capture the playing style of a player in a so-called 'player vector' that can be interpreted both by human experts and machine learning systems. We demonstrate the validity of our approach by recovering commonly known player types (e.g., left-winger, right-center defender) through unsupervised clustering and by substantiating a number of claims in popular media about soccer players (e.g., "Paolo Dybala is the new Lionel Messi") with our results.

## 1 Introduction

Data analysis is gaining importance quickly in many sports [22]. Sports clubs are analyzing huge amounts of data to gain a competitive advantage on their opposition. Soccer has been a relative late comer in this trend. The classic statistics (e.g., that appear in boxscores or are often reported on television) tend to be raw counts or fractions, such as ball possession percentage, number of shots on target or pass success percentage. While interesting, these statistics do not give the full picture and can sometimes obscure important information. For example, the raw number of shots a player took does not tell us about the relative difficulty or quality of the attempts. Recently, research has focused on approaches that allow for a deeper and more insightful analysis of soccer players [4]. One well-known example is the expected goals statistic [10], which is now discussed on the popular soccer talk show *Match of the Day* on BBC One.

This paper continues in this vein and attempts to characterize a player's style in a 'player vector' in a purely data-driven manner. Characterizing the style of a player in an objective way offers important advantages for soccer clubs in three departments: scouting, player development monitoring, and match preparation.

**Scouting** Soccer clubs can search the market more intelligently if they know the type of player they are looking for and how well prospective targets match that type. Transfers are expensive, and clubs are always looking for bargains and ways to mitigate risks in player recruitment.

**Monitoring player development** The coach can inspect the playing style of a player in a human-interpretable player vector. If the player vector matches the expectations of the coach, then the coach can monitor that this player vector remains stable and unchanged. If the player vector does not match the expectations of the coach, then he can give his player some pointers and afterwards monitor how well the player is implementing the advice.

**Match preparation** Understanding the playing style of your opponent can offer certain tactical advantages. The defenders of a team will wish to know what type of attackers they are up against. Similarly, the attackers of a team will be interested in the playing style of the defenders they need to score against.

## 2 Data

Our data set consists of play-by-play data of 9155 matches of the five major soccer competitions in Europe: the English Premier League, the German Bundesliga, the Spanish Primera Division, the Italian Serie A and the French Ligue Un. Our data spans almost all matches between the 2012/13 and 2016/17 seasons. For all matches, we have access to a *match sheet* and an *event stream*. The match sheet contains information about the players such as their full name, the team they play for, their position, the number of minutes played in the match, etc. The event stream is a sequence of 1750 events (on average) detailing all on-the-ball events such as shots and passes. Each event contains the following attributes:

**Type** The type of event, e.g., shot, tackle, pass.

**Time** A timestamp detailing when the event happened.

**Player** The player involved in the event.

**Position** The position of the player on the pitch, encoded as a $(x, y)$-coordinate. The dimensions of the soccer pitch are normalized such that both $x$ and $y$ are a number between 0 and 100.

## 3 Building Player Vectors

Our approach to characterizing player style involves constructing a vector that describes each player. In order to construct this vector, we must address the following two questions:

1. What information that appears in event stream data is informative about player style?
2. How can we encode this information into a player vector?

To address the first question, we build upon the work of Kumar et al. [17], who researched the correlation between 198 different player statistics (e.g., goals scored, number of assists, number of last man tackles, etc.) and expert ratings per player per match found on `WhoScored.com`. Based on this work, we consider the following event types: shots, goals, passes, fouls, yellow or red cards, take-ons, clearances, aerials, interceptions, recoveries, and tackles.

To address the second question, our primary insight is that for some events, the location on the pitch where an event occurs is informative of a player's style whereas for others raw counts are sufficient. However, reasoning about location is somewhat challenging as there are a large number of possible locations on the pitch that we want to summarize with a very small number of features to maintain interpretability. This motivates to explore using non-negative matrix factorization (NMF) [3], which was successfully used to characterize the shooting behavior of basketball players based on shot locations [13]. Next, we describe our approach to use NMF for feature construction in more detail.

### 3.1 Location-Specific Features via NMF

Given a player $p_i$, his events $E_i$ of type $et$ and a playing field $F$, we overlay a grid of size $m \times n$ on the field. For each cell of the grid, we count the number of events in that cell and construct a matrix $X_i \in \mathbb{N}^{m \times n}$ per player $p_i$. As we wish to maintain spatial coherence (i.e., cells that are close to each other should have similar values), we then apply a Gaussian blur to the matrix. This is a standard image processing technique [24] that involves convolving the image with a Gaussian function. Specifically, the value of each cell in $X_i$ is replaced by a weighted average of itself and its neighborhood (in our case, a 5x5 grid), leading to the blurred matrix $X_i' \in \mathbb{R}_+^{m \times n}$ (See Figure 1).

Next, for each player $p_i$, we reshape his matrix $X_i'$ to a 1-dimensional vector $\mathbf{x_i}$ of length $m \cdot n$. We then construct a matrix $M_{et} = [\mathbf{x_0} \mathbf{x_1} \ldots \mathbf{x_l}]$ that encompasses our grid data of all players for a specific event type $et$. Here, $l$ is the total number of players in our data set. We then apply non-negative matrix factorization to $M_{et}$.[1] The result is a set of two matrices, W and H such that:

$$M_{et} \approx WH, \tag{1}$$

where $M_{et} \in \mathbb{R}_+^{m \cdot n \times l}$, $W \in \mathbb{R}_+^{m \cdot n \times k}$ and $H \in \mathbb{R}_+^{k \times l}$. Here, $k$ is a user-defined parameter that refers to the chosen number of principal components. The columns of $W$ are the principal components of the event type that represent different spatial groups of that event type (e.g., different shot areas such as the ones seen in Figure 2). The $i$-th row of $H$ represents the playing style of player $p_i$ for an event type captured in a feature vector of $k$ values. This feature vector can also be visualized by multiplying it with the principal component matrix $W$ (Figure 3).

To illustrate the intuition behind applying NMF, consider the case of shots. Here, we chose to overlay the playing field with a 200x100 grid (length by width). The matrix $M_{shot}$ was decomposed into five principal components as shown in Figure 2. We picked five components based on preliminary experiments, visual

[1] We used the *sklearn.decomposition.NMF* function in the popular Scikit-learn Python package [19].

inspection of the components and expert knowledge. Each of the components shown corresponds to a spatially coherent, and largely non-overlapping area. Figure 3 shows how we are able to accurately characterize the shot behavior of Neymar (FC Barcelona between 2013/14 and 2016/17) by reconstructing his shot behavior using only the five selected components.

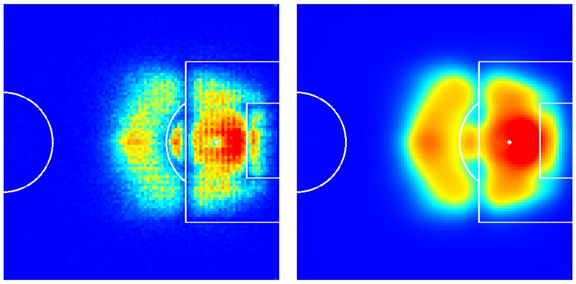

(a) No Gaussian blur (b) With Gaussian blur

Fig. 1: A player shot matrix before and after applying a Gaussian blur.

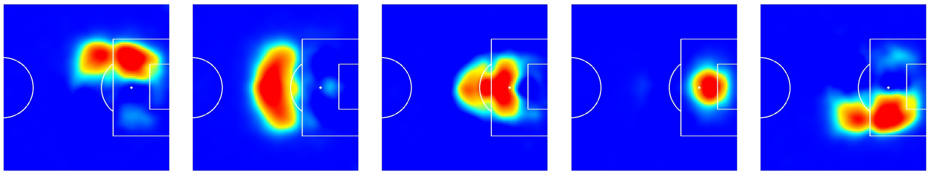
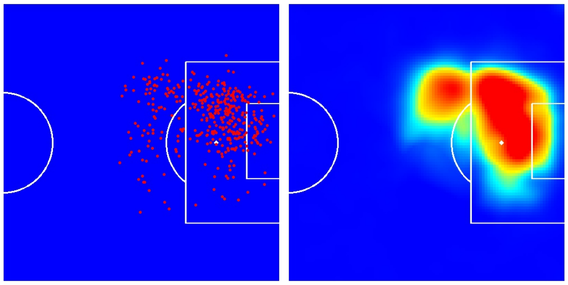

(a) Left (b) Outside box (c) Middle (d) Penalty area (e) Right

Fig. 2: All shots decomposed into five shot components using non-negative matrix factorization.

### 3.2 Building the Vector

We apply the approach outlined in Subsection 3.1 to construct features for all the event types shown in Table 1. However, special care has to be taken when processing goals and passes.

**Goals** Unfortunately we do not have enough data available to get a meaningful decomposition of goals. However, because goals are too important to ignore,



(a) Raw shots (b) Reconstructed shot behavior

Fig. 3: The raw shots of Neymar versus a reconstruction of his shooting behavior using a feature vector containing only 5 values.

we still construct a feature vector for goals through some engineering. More specifically, we piggyback on our decomposition of shots and construct a goal feature vector by assigning every goal a player made to the shot component that was most likely to produce that shot. This produces five counts, one for each shot component. These counts together then form our goal feature vector.

**Passes** Unlike all other events, passes have both a start and end location, which gives them four spatial dimensions $(x_{start}, y_{start}, x_{end}, y_{end})$ instead of two $(x, y)$. We solve this hurdle by slightly modifying the approach in Subsection 3.1 so that it constructs a 4-dimensional grid. We then construct a 4-dimensional pass tensor per player, which can also be converted to a 1-dimensional vector. From there, the approach continues as normal. Visualizing the resulting components is somewhat more challenging, but can still be done.

Finally, clearances, yellow and red cards exhibit little variance in their location. Therefore, instead of using NMF, we simply use the frequency of the event type.

We now construct our player vector that characterizes the playing style of a player by simply concatenating his feature vectors for each event type. Then, we normalize every value in the vector by the total number of minutes each player spent on the field. We do this because we want to compare players invariant to their total playing time, as we believe a per minute ratio is a more fair comparison. The playing time factor was still present in the player vector because NMF usually decomposes data into components and values with real-world meaning [3]. In our case for example, the value in a player vector for shot component 'left' roughly resembles the total number of shots from the left flank a player took.

Table 1: The number of components and the grid size used when constructing feature vectors for each event type. The number of components used for each event type was chosen based on preliminary experiments, visual inspection of the components and expert knowledge.

| Event type | Components | grid size |
|---|---|---|
| Shot | 5 | 200×100 |
| Goal | 5 | 200×100 |
| Pass | 15 | $(20 \times 10)^2$ |
| Foul | 7 | 100×50 |
| Yellow card | 1 | 1×1 |
| Red card | 1 | 1×1 |
| Take-on | 6 | 100×50 |
| Clearance | 1 | 1×1 |
| Aerial | 9 | 100×50 |
| Interception | 5 | 100×50 |
| Recovery | 4 | 100×50 |
| Tackle | 6 | 100×50 |

### 3.3 Finding Similar Players

The similarity between two players' playing styles is the Euclidean distance between their player vectors after normalizing for feature importance. We studied the importance of every feature to characterize playing style based on the stability of every feature over time.

For each player, we constructed two player vectors: one based on his data in seasons 2012/13 - 2013/14 and one based on his data in seasons 2014/15 - 2015/16. For each feature, we compute the average difference in those two player vectors over all players. The weight by which we normalize each feature is then the inverse of that average difference. While there will be exceptions, we assume that most players have the same playing style in both parts of the data. Changes in a player's style are more likely to emerge over the course of a career (i.e., a longer time period) as opposed to the shorter time span we consider. Hence, features that show large differences are not that relevant to the style of the player, are unstable, and should receive a lower weight. On the other hand, features that show only small differences are stable and should receive a higher weight in the final player similarity function.

The main conclusion of our feature importance study is that most features are roughly equally important and do not need to be normalized, with the exception of a small number of features about rare events (such as fouls) that receive a higher weight in our distance function.

## 4 Experiments

Playing style remains an incredibly subjective concept. There is no ground truth, which poses challenges when evaluating our approach. Therefore, we consider three different use cases to show the potential of our approach. First, we cluster all player vectors to recover commonly known player types. Second, we show how our approach could be used at clubs for scouting by substantiating a number of recent claims in popular media about professional soccer players. Third, we show how our approach can be used for player development monitoring.

### 4.1 Recovering player types through player vector clustering

Using our distance function in Subsection 3.3, we can cluster the player vectors of all players in our data set and identify well known soccer player archetypes. This is similar to Pappalardo et al., who use the average position of players to find roles with clustering, finding eight different prototypical roles [18] Figure 4 shows for each player in our data set his average playing position on the pitch and the cluster he belongs to when we apply $k$-means clustering to the player vectors with $k = 10$. This figure shows a 4-3-3 formation, which is one of the most common team formations in soccer, with a clear distinction among player types such as left-winger, right-back or center-midfielder.

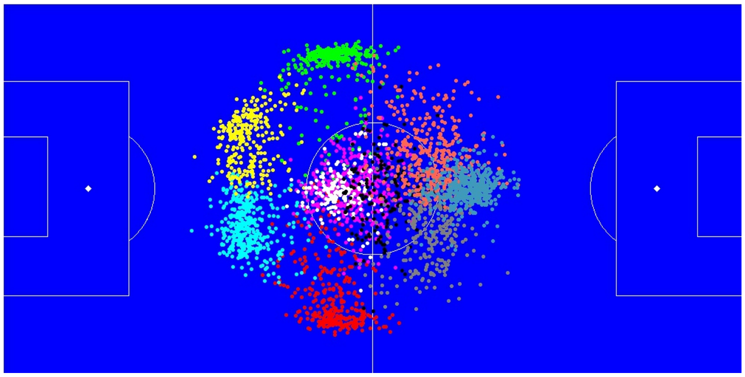


Fig. 4: Hierarchical clustering

### 4.2 Scouting

Lionel Messi is regarded by many as the best soccer player in the world. One player who has been deemed to play similarly to Messi is Paolo Dybala [26, 14]. Dybala (24 years of age) is also an Argentinian attacker. Using our player vectors, we see that Dybala is the $15^{th}$ most similar player (out of 3205) to Lionel Messi.

Idrissa Gueye (28, midfielder at Everton FC) is often hailed as the new N'golo Kante (27, midfielder at Chelsea FC) by many journalists [1, 16, 2]. Gueye is the $13^{th}$ most similar player to Kante in our data set.

Aymeric Laporte is a 23-year-old defender playing for Manchester City FC, who was deemed to be the long-term replacement for 32-year-old Real Madrid defender Sergio Ramos[21, 5], who was named best defender in the world in 2017 by UEFA.[2] Laporte is the number one most similar player to Ramos using our player vectors. Looking closer at the data, we see that they perform particularly similar when it comes to aerials and tackles. Both event types are essential elements in defending. This helps explain why they are regarded as similar defenders.

In the summer of 2017, Chelsea FC sold their star striker Diego Costa (29) to Atlético Madrid, replacing him with Alvaro Morata (25) from Real Madrid. Looking at their respective player vectors, Diego Costa is the most similar player to Morata. Looking at their player vectors, we see that they perform particularly similar when it comes to goals, passes and take ons.

Another example is Swansea City FC who sold their striker Wilfried Bony (29) to Manchester City FC for about 32 million Euros in 2015.[3] They replaced him with the cheaper Bafétimbi Gomis (23) on a free transfer from Olympique Lyonnais. Our data shows that Gomis is the $17^{th}$ most related player to Bony. The aspects in which these two players are the most related are: shots, goals, tackles and take-ons, which are typical events for attackers.

### 4.3 Monitoring player development

Journalists agree that Cristiano Ronaldo (Real Madrid) has evolved from his role as a winger to a role as a central striker [25, 23]. Figure 5 shows Ronaldo's shooting behavior from 2012/13 and 2016/17. In addition, we also noticed a clear decrease in his features for the 'left' and 'outside box' shot components.

Most players at Liverpool were coached by Brendan Rodgers for the seasons 2012/13 to 2014/15. Afterwards, they were coached by Jürgen Klopp. A player who has been playing for Liverpool during all five seasons in our data set is Jordan Henderson. The media agrees that in the 2016/17 season, Klopp instructed Henderson to play more defensively, as suggested by these two articles [28, 20]. When looking at Henderson's features for the five shot components, we notice a sharp decrease in all five features starting from 2016/17 onwards. This confirms that Henderson does not shoot as often as he used to, which indicates that he plays more defensively.

## 5 Related Work

Kumar et al. [17] seek to determine which aspects of soccer are relevant in player ratings by experts. Their results were used as guidelines to select the relevant event types from the match stream data to determine the style of a player. Danneels et al. [6] predict a player's position (i.e., attacker, midfielder, defender)

[2] http://www.uefa.com/insideuefa/awards/previous-winners/newsid=2495000.html

[3] https://www.transfermarkt.com/wilfried-bony/transfers/spieler/81808

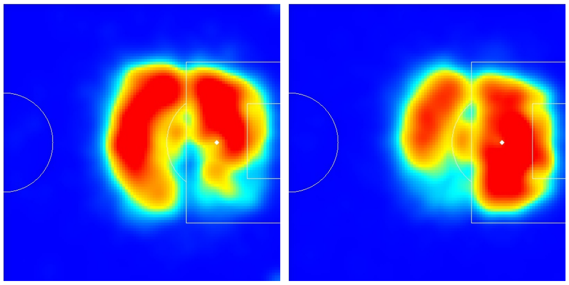

(a) 2012/13 (b) 2015/16

Fig. 5: Shooting behavior of Cristiano Ronaldo

based on their actions. While similar to our research, our goal and approach is more broad and ambitious, as our player vectors are much more detailed than only three distinct labels. Gyarmati et al. [15] construct movement vectors to characterize a player by his movement on the field.

Van Gool et al. [27] analyze the playing style of teams instead of players. While their approach is completely different, their goal is quite similar to ours as they try to capture a subjective concept like playing style in a more objective and data-driven way. Similarly, STATS introduced *STATS Playing Styles* [12], which are eight different styles (e.g., fast tempo, direct play, counter attack) teams use to create shooting opportunities. Fernandez et al. [11] also categorize different styles of play for teams in professional soccer.

Pappalardo et al. [18] introduce *PlayeRank*, a framework that ranks soccer players according to the quality of their performance in a series of games. They also construct player roles by clustering the average position of players, which is similar to our experiment in Section 4. The biggest difference between their work and ours is that Pappalardo et al. analyze player data with the goal of evaluating a player's quality of performance, while we aim to characterize his playing style, with less emphasis on their quality of play. *STARSS* [9] is a framework that ranks soccer players based on their contribution to their team's offensive output. The approach for measuring this contribution is loosely based on the *POGBA* algorithm [7]. One way we could improve our approach is to expand our player vector with features that capture the tactics a player is involved in (e.g. [8]).

In other sports, Franks et al. [13] used spatial information to categorize shots in professional basketball. In this work, data from the NBA was collected and analyzed using non-negative matrix factorization (NMF). This paper was a huge influence on our work, as our approach on soccer event data is largely inspired by their approach on basketball event data.

## 6 Conclusions

Objectively characterizing the playing style of professional soccer players has important applications in scouting, player development monitoring, and match preparation. We showed how to construct player vectors by transforming sets of events from match stream event data to fixed-size feature vectors using non-negative matrix factorization. These player vectors offer a complete view of a player's playing style (within the limits of the data source), are constructed in a purely data-driven manner, are human-interpretable and can be used in machine learning systems such as clustering and nearest neighbor analysis. We presented a number of use cases in player type discovery, scouting, and player development to show the potential of our approach.

Future directions for this work could be to include additional information about players, such as (a) height, weight, and age, (b) skill judgments by human soccer experts at `SoFifa.com`, or (c) tracking data, which can give more insight into how a player moves around the field. Taking age into account could also expand our use case to predicting the development of young players based on the development of players who had similar player vectors in the past.

## Acknowledgements

Tom Decroos is supported by the Research Foundation-Flanders (FWO-Vlaanderen). Jesse Davis is partially supported by KU Leuven Research Fund (C14/17/070 and C22/15/015), FWO-Vlaanderen (SBO-150033) and Interreg V A project NANO4Sports.

## References

1. Adewoye, G.: Everton boss Sam Allardyce compares Idrissa Gueye to N'Golo Kante, http://www.goal.com/en/news/everton-boss-sam-allardyce-compares-idrissa-gueye-to-ngolo/gddgazktcl3b1ayeadrva1o18
2. Callaghan, S.: Everton boss was spot-on with Idrissa Gueye - N'Golo Kante comparison (2018), http://www.hitc.com/en-gb/2018/04/12/everton-boss-was-spot-on-with-idrissa-gueye-ngolo-kante-comparis/
3. Cichocki, A., Phan, A.H.: Fast local algorithms for large scale nonnegative matrix and tensor factorizations. IEICE transactions on fundamentals of electronics, communications and computer sciences **92**(3), 708–721 (2009)
4. Coles, J.: The Rise of Data Analytics in Football: Expected Goals, Statistics and dam (2016), http://outsideoftheboot.com/2016/07/21/rise-of-data-analytics-in-football/
5. Collins, T.: 4 Possible Replacements Should Real Madrid Sell Sergio Ramos (2015), http://bleacherreport.com/articles/2509541-4-possible-replacements-should-real-madrid-sell-sergio-ramos#slide3
6. Danneels, G., Van Haaren, J., Op De Beck, T., Davis, J.: Identifying playing styles in professional football. KU Leuven (2014)
7. Decroos, T., Dzyuba, V., Van Haaren, J., Davis, J.: Predicting soccer highlights from spatio-temporal match event streams. In: AAAI. pp. 1302–1308 (2017)

8. Decroos, T., Van Haaren, J., Davis, J.: Automatic discovery of tactics in spatio-temporal soccer match data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 223–232. ACM (2018)
9. Decroos, T., Van Haaren, J., Dzyuba, V., Davis, J.: Starss: A spatio-temporal action rating system for soccer. In: Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2017 workshop (2017)
10. Eggels, H.: Expected Goals in Soccer: Explaining Match Results using Predictive Analytics (2016), eindhoven University of Technology
11. Fernandez-Navarro, J., Fradua, L., Zubillaga, A., Ford, P.R., McRobert, A.P.: Attacking and defensive styles of play in soccer: analysis of spanish and english elite teams. Journal of sports sciences **34**(24), 2195–2204 (2016)
12. Flynn, M.: STATS Playing Styles An Introduction (2016), `https://www.stats.com/industry-analysis-articles/stats-playing-styles-introduction`
13. Franks, A., Miller, A., Bornn, L., Goldsberry, K.: Characterizing the spatial structure of defensive skill in professional basketball. Annals of Applied Statistics 2015, Vol. 9, No. 1 pp. 94–121 (2015), arXiv:1405.0231
14. Goal.com: Messi admits difficulties in Dybala partnership: He plays like me at Juve, http://www.goal.com/en/news/messi-admits-difficulties-in-dybala-partnership-he-plays-like-me-/1uq96ju5zageb1s1vez93omsi3
15. Gyarmati, L., Hefeeda, M.: Analyzing in-game movements of soccer players at scale. arXiv preprint arXiv:1603.05583 (2016)
16. Kleebauer, A.: Everton's Idrissa Gueye is the new N'Golo Kante - and here are the stats to prove it (2017), https://www.liverpoolecho.co.uk/sport/football/football-news/evertons-idrissa-gueye-new-ngolo-12965076
17. Kumar, G., Alfarone, D., Van Haaren, J., Davis, J.: Machine learning for soccer analytics. KU Leuven (2013)
18. Pappalardo, L., Cintia, P., Ferragina, P., Pedreschi, D., Giannotti, F.: Playerank: Multi-dimensional and role-aware rating of soccer player performance. arXiv preprint arXiv:1802.04987 (2018)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
20. Pierce, J.: Henderson: I'm learning fast in the new midfield role Klopp's given me (2016), https://www.liverpoolecho.co.uk/sport/football/football-news/henderson-im-learning-fast-new-11862193
21. Prenderville, L.: Sergio Ramos 'identifies Aymeric Laporte and Matthijs de Ligt as his long-term replacements' at Real Madrid (2017), https://www.mirror.co.uk/sport/football/transfer-news/sergio-ramos-identifies-aymeric-laporte-11710624
22. Pritchard, S.: Marginal gains: the rise of data analytics in sport (2015), https://www.theguardian.com/sport/2015/jan/22/marginal-gains-the-rise-of-data-analytics-in-sport
23. Romero, A.: Cristiano Ronaldo: the change to a 'number 9' (2016), https://en.as.com/en/2016/12/19/opinion/1482164003_264275.html
24. Shapiro, L., Stockman, G.C.: Computer vision. 2001. Ed: Prentice Hall (2001)
25. Sharma, R.: How Cristiano Ronaldo has been transformed from a winger into a deadly No 9... and why he could really play for Real Madrid into his 40s (2017), http://www.dailymail.co.uk/sport/football/article-4469198/How-Ronaldo-transformed-winger-deadly-No9.html

26. Smith, R.: Is Paulo Dybala the Next Lionel Messi? “He Can Go as High as He Likes“ (2017), https://www.nytimes.com/2017/04/10/sports/soccer/paulo-dybala-juventus-lionel-messi-barcelona.html
27. Van Gool, J., Van Haaren, J., Davis, J.: The automatic analysis of the playing style of soccer teams. KU Leuven (2015)
28. Williams, G.: Jordan Henderson is relishing his new role in the Liverpool midfield (2016), https://www.liverpoolecho.co.uk/sport/football/football-news/liverpool-jordan-henderson-jurgen-klopp-12123785