# Pass2vec: Analyzing Soccer Players' Passing Style using Deep Learning

**Hyeonah Cho1, Hyunyoung Ryu12, and Minseok Song123**

**Abstract**

The aim of this research was to analyze the player's pass style with enhanced accuracy using the deep learning technique. We proposed Pass2vec, a passing style descriptor that can characterize each player's passing style by combining detailed information on passes. Pass data was extracted from the ball event data from five European football leagues in the 2017-2018 season, which was divided into training and test set. The information on location, length, and direction of passes was combined using Convolutional Autoencoder. As a result, pass vectors were generated for each player. We verified the method with the player retrieval task, which successfully retrieved 76.5% of all players in the top-20 with the descriptor and the result outperformed previous methods. Also, player similarity analysis confirmed the resemblance of players passes on three representative cases, showing the actual application and practical use of the method. The results prove that this novel method for characterizing player's styles with improved accuracy will enable us to understand passing better for player training and recruitment.

**Keywords:** Sports Analytics, Soccer, Player style, Passing Style, Convolutional Autoencoder

---

1 Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Korea
2 Open Innovation Bigdata Center, Pohang University of Science and Technology
3 Korea Sport Industry Development Institute, Pohang University of Science and Technology

**Corresponding author:**
Minseok Song, Department of Industrial and Management Engineering, Pohang University of Science and Technology, 77 cheongamro Namgu, Pohang, Republic of Korea.
Email: mssong@postech.ac.kr

# Introduction

The recent technological development is transforming game analysis and tactics in professional football. It is now possible to capture more accurate information of the player's moves and team actions using auto-trackers, analyzed video motion, and Global Positioning System (GPS).[1] Additionally, advanced data analytics has not only reduced the chance of misjudgment, but also enabled assessing the performance of teams and players to reinforce scouts and tactics. For example, machine learning and deep learning techniques are increasingly being applied for sports big data analysis. Machine learning, in general, is divided into supervised learning and unsupervised learning, which former includes classification or prediction, and the latter used for clustering. Classification in soccer is used for analyzing past games to identify the important game factors [2,3] or recognize patterns relevant to goal scoring.[4] Models can even predict scores and winning [5–8] and are expected to automatically build strategies against the opponent team's tactic in the future.[9,10] Clustering algorithm can compress large-scale data to describe a team's style such as formation during matches or player role distribution, which can be useful for strategic planning, evaluation, and tactical adjustments.[11–13]

Passing, one of the most important players' actions, has been analyzed with diverse data and methodologies.[14–16] Prior to applying data analytic methods, passes had been described somewhat subjectively by the observers who had experiential knowledge.[17] Studies capture the team pass characteristics from (a) ball movement pattern, (b) within the passing network, (c) and flow motif. Lucey et al.[18] proposed a method to characterize the team's ball movement patterns using occupancy maps. They constructed occupancy maps via play segments, which contain ball movement information over time. Besides pattern recognition for team pass characteristics, some studies focus on the pass itself and measure the effectiveness in relation to involved players. Rein et al.[19] created a Voronoi diagram to create pass effectiveness measures which calculated the pass from the ball movement, the number of bypassing opponent players, and the space created next to the targeted goal. Spearman et al.[20] created a 'pass probability model' based on a mathematical concept that can quantify the value of passes and the skills of players involved in a pass. The metric calculates the likelihood of pass success from time-to-intercept and time-to-control using both tracking and event data. The study defines the pass as a ball control at the player level and only regards the game state when the ball is kicked. The result could be useful for match analysis and player scouting. Another approach is to identify passes within a passing network. Peña & Touchette[21] defined the passing network of a team by setting players as nodes and connecting between two players weighted by the number of successful passes between them. From the result, they identified the team's play patterns, determined hotspots during the play, and located potential weaknesses. Social network analysis (SNA) is often used for analyzing the player connection and team performance.[22] Moreover, flow motifs, an ordered list of players who were involved in a certain pass are used for analysis. Meza[23] classified flow motifs using spatial variables that can better frame unique team passing profiles, however, it focuses on the certain structure of passes and does not differentiate motifs based on the names of the players. Although the approaches vary, these studies describe the passes more objectively with data analysis and provide information to improve team pass tactics or a player's pass skills.

More recently, advanced analytics from ML/AI can capture more complex and dynamic features of the passes compared to the previous statistical analysis. Also, the abundance of video and image data is creating near real-time data for tracking the actions of a players' movement. Positional data and event data are most often used for analyzing passes. Brooks et al.[25] analyzed the team's passing style by building heatmaps of 18 different zones. They counted the number of passes that originated from each zone and

normalized the counts by the total number of passes to visualize. The used k-nearest neighbor (KNN) algorithm to rank offensive players. Goes et al.[26] presented a quantitative model to measure pass effectiveness. Using the tracking data, multiple linear regression and PCA were used for prediction and clustering. The measure of pass effectiveness is derived from line centroids, team spread, and team surface areas. In this study, pass effectiveness was not necessarily relevant to goal scoring. The result can be used to help coaches train their players to increase the effectiveness of passes. As a soccer game has complexed features and dynamic movements regarding both team play and individual players, the pass is also analyzed in this respect. Fernandez and Bornn[27] calculated "probability surfaces" with deep convolutional neural networks. The study estimates pass probability, the likelihood of player's pass selection, and visualized surfaces to show the expected value of the passes. With visual format and detailed information from spatiotemporal data, the coaches can identify optimal passing locations and team-level passing tendencies. Power et al.[28] estimated risks and rewards of all passes using a supervised learning approach. They used logistic regression and K-means clustering for classification. The model can support coaches to choose players according to the opposition in terms of pass risk and reward. Moreover, the analysis of passes is getting more advanced with multiple features involved. Arbue's Sanguesa[29] presented a novel computational model for 'pass feasibility' that can analyze a player's body orientation related to the outcome. The study considers passing features from the point of an individual player, such as the player's direction, distance from a passer, and the status of defenders blocking. To obtain the relevant information on a 2D field, the Open Pose data set and Support Vector Machine model has been applied. Although the purpose of analysis might be different, all these approaches might be different by constructing models and using a large set of data improves the resolution of pass actions. Not only at the team level but also at the individual level assessment is possible.

We aim to understand the passes at the player level on the assumption that passes can be characterized more precisely by including multiple features such as location, direction, and length. We also find possibilities from the recent machine learning model to integrate this information and improve the accuracy of pass description. Decroos & Davis[30] characterized the playing style of players with location by using the ball event data. They focused on the moving range of players on the field, the probability of visiting the same location, and the type of movement on each spot to define a player's style. The actions include pass, cross, shot, and dribble. They create a "player vector" by drawing and compressing heatmap for the player's pass, dribble, cross, and shot to characterize a playing style. Non-negative matrix factorization (NMF) was used for clustering and dimensionality reduction. The authors explained the advantages of using vectorization, which is more informative and intuitive than looking at the player's location and movement on each grid cell on a heatmap. A similar approach to this, analyzing an individual player's pass movements can potentially help coaches to devise tactical plans and select players.[31]

In this context, we propose a novel method to analyze soccer players' passing style, Pass2vec, a passing style descriptor combining detailed information on passes. The created vector can represent a player's unique passing style by integrating pass location, length, and direction. This deep learning approach allows us to distinguish players more accurately from other players. Table 1 summarizes the difference between previous studies and this study.

There are three major contributions of this study:

- We propose a novel method to define a soccer player's passing style that distinguishes a player from other players.
- We validate our proposed method on real soccer ball event data.

- We introduce how our method can be applied to support tactical decision-making about player replacement.

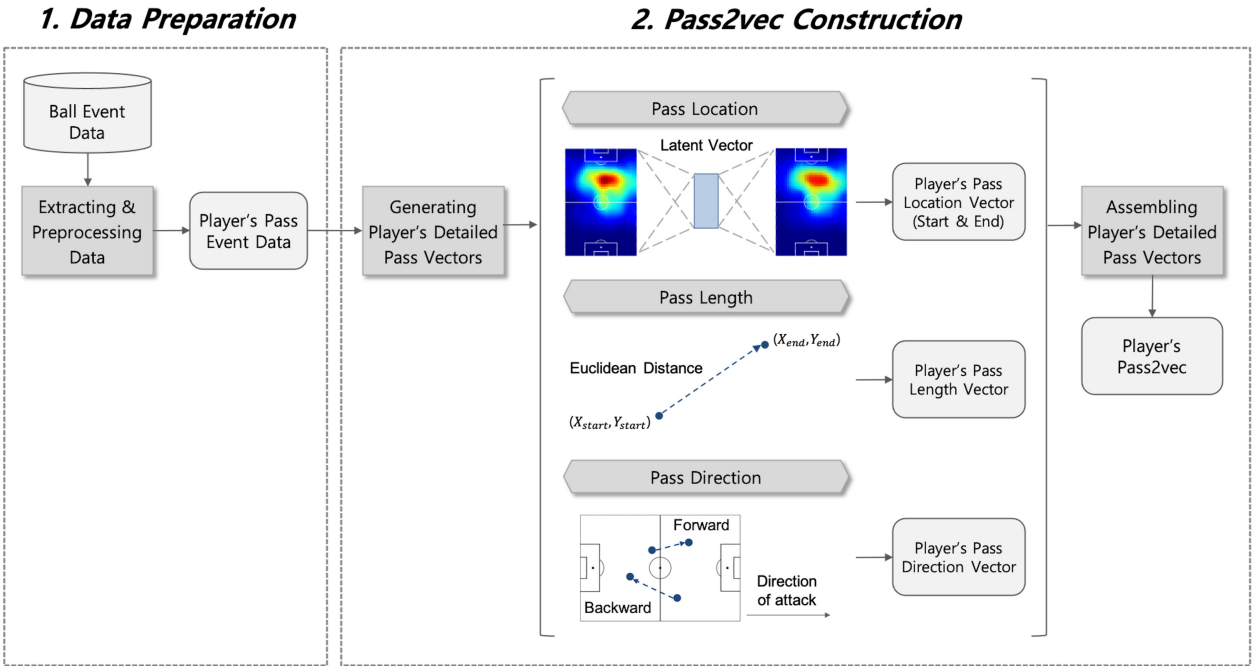
**Table 1.** Summary of previous studies and this study.

| Author | Gyarmati et al. (2016) | Decroos & Davis (2019) | Peña et al. (2015) | Bekkers & Dabadghao (2019) | Goes et al. (2019) | This study |
|---|---|---|---|---|---|---|
| **Objectives** | Characterizing playing style | Characterizing playing style | Characterizing pass style | Measuring the effectiveness of pass style linked to goals | Measuring pass effectiveness apart from goals | Characterizing pass style |
| **Subject of analysis** | player | player | player | player, team | player | player |
| **Actions** | pass, shot, dribble, and tackle | pass, cross, shot, and dribble | pass | pass | pass | pass |
| **Parameters** | time, location, speed, and ball possession | location | relative position within a passing sequence | relative position within a passing and expected goal sequence | time, pass length, pass direction | location, pass length, pass direction |
| **Data** | ball event data | ball event data | ball event data | ball event data | tracking data | ball event data |
| **Analytic measures** | Player's movement characteristics construction using K-means clustering | Player vector construction using NMF (Non-negative Matrix Factorization) | Player's flow motif analysis | Flow motif analysis of player and team | Principal component analysis | Pass2vec construction using CAE (Convolutional Autoencoder) |

# Materials and Methods

Pass2vec is a new method to characterize a soccer player's passing style with detailed information of passes in matches. The development process is in two steps: 1) data preparation and 2) Pass2vec construction. We first extract a player's pass event data from the entire ball event data and preprocess it in the desired format. Then, we draw vectors that represent the location, length, and direction of the player's passes. Finally, combine the player's vectors to construct a player's Pass2vec. As a result, the descriptor carries a large data set of a player's pass styles. Figure 1 describes the overall process of construction.

[insert Figure 1.]




*Data Preparation*

As part of efforts to derive insights from a player's passing events in matches, we extract and preprocess data in the initial step. Since we focus on pass events among the various actions taking place in matches, we extract only a player's pass event data from a whole ball event data. A ball event is a record of a player's action when in possession of the ball on the pitch. Some sports analytics companies, such as Wyscout and Opta, have collected ball event data from video records of football matches for analytics. It annotates the timestamp and location in (x, y) coordinates of all player's actions like dribble, pass, and shot during the game.

A ball event data in soccer generally do not include detailed information on passes, such as passing length and direction. Therefore, we derive the length and direction of passes through the location of pass events (Figure 2). However, if the raw data is not in meter but the relative percentage from [0,100], the width and length should be calculated by multiplying the actual size of a soccer field. Then, the length of the pass is calculated with Euclidean Distance between a start location and an end location of the pass in meters. Euclidean Distance is computed by:

$$dist_{Euclidean}(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (2.1)$$

, where $p = (p_1, p_2, \dots, p_n)$ and $q = (q_1, q_2, \dots, q_n)$.

Also, the direction of the pass that played toward the opponent's goal is 'forward' and the direction of the pass that played toward its own goal is 'backward'. In this way, the data for each player's pass event is extracted with detailed pass information.

[insert Figure 2.]

| MatchId | Period | Timestamp | Action | Team | Player | Start Location | End Location |
|---|---|---|---|---|---|---|---|
| 2499968 | 1 | 1381.552 | Pass | Tottenham Hotspur FC | C. Eriksen | [78, 86] | [91, 59] |
| 2499968 | 1 | 1382.761 | Shot | Tottenham Hotspur FC | H. Kane | [91, 59] | [0, 0] |
| 2499968 | 1 | 1405.695 | Pass | Tottenham Hotspur FC | D. Sánchez | [31, 51] | [5, 28] |
| 2499968 | 1 | 1410.69 | Pass | Tottenham Hotspur FC | H. Lloris | [5, 28] | [42, 12] |
| 2499968 | 1 | 1413.09 | Duel | Manchester United FC | A. Valencia | [58, 88] | [62, 93] |
| 2499968 | 1 | 1413.51 | Duel | Tottenham Hotspur FC | Son Heung-Min | [42, 12] | [38, 7] |
| ... | ... | ... | ... | ... | ... | ... | ... |

***Data extracting & preprocessing***

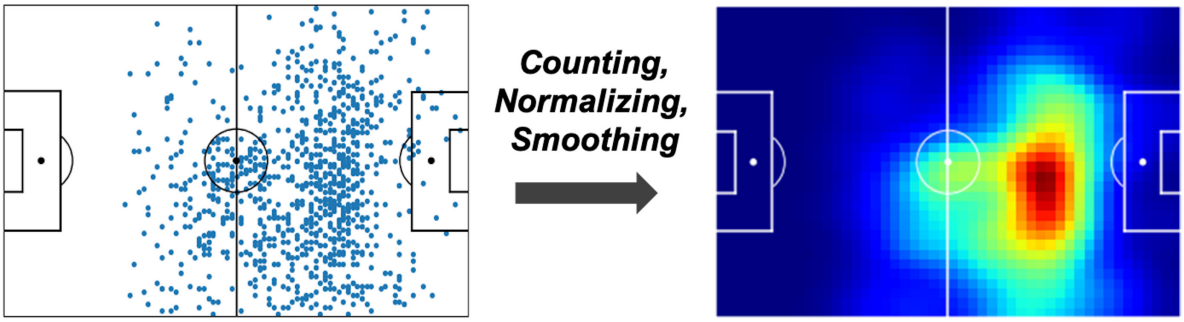
| MatchId | Action | Team | Player | Start Location | End Location | PassLength | PassDirection |
|---|---|---|---|---|---|---|---|
| 2499968 | Pass | Tottenham Hotspur FC | C. Eriksen | [78, 86] | [91, 59] | 29.97 | forward |
| 2499968 | Pass | Tottenham Hotspur FC | C. Eriksen | [61, 95] | [85, 54] | 47.51 | forward |
| 2499968 | Pass | Manchester United FC | C. Eriksen | [60, 67] | [75, 19] | 50.29 | forward |
| ... | ... | ... | ... | ... | ... | | |

*Pass2vec Construction*

***Pass Location Vector.*** To generate a pass location vector of player $p$, we draw a flattened heatmap. The heatmap indicates locations where a player's passes started and ended. We created a vector in the following two steps: constructing and compressing the heatmap.

To construct a pass heatmap, we determined the size of a grid, normalized the pass count in each grid, and smoothed the image (Figure 3). In the process of heatmap construction, we marked the locations where each player's pass started and ended and overlaid on a grid size $m \times n$ over the pitch for each player $p$. [30]. We used a 50 x 50 grid for this study. The number of passes that performed in each cell $X_{ij}$ are counted on a matrix of $X \in \mathbb{N}^{m \times n}$. Then, we divided the counts by total minutes of a player's play and multiplied 90 minutes to normalize the counts within the 90-minute match (i.e., $X' = \frac{90}{Minutes\ player\ played} X$). Gaussian blur, which is an image blurring technique from the Gaussian function in image processing, smoothed the counts to $X'$ which enhances the spatial coherence of the locations where the passes were performed. As we get the blurred matrix $X'' \in \mathbb{R}_{+}^{m \times n}$, we draw separate heatmaps of $X''_{start}$ and $X''_{end}$. we saved these heatmaps in 48 pixels wide and 32 pixels high RGB images, the shape of $48 \times 32 \times 3$ as in 4,608 dimensions.

[insert Figure 3.]




[A scatter plot of a player's pass locations] [A heatmap of a player's pass locations]

We compress and flatten the player's pass heatmap with Convolutional Autoencoder (CAE). CAE has been successful in reducing the dimensionality of images[32]. Therefore, we can draw a compressed representation of the player's pass heatmap by training CAE to reconstruct heatmaps. We outlined three convolutional and deconvolutional layers with Rectified Linear Unit (ReLU) as an activation function for each layer. First, we train CAE to reconstruct the original heatmap. After that, put the heatmap into the trained CAE and extract the latent vector. The process of CAE training is described in detail in Algorithm 1. The structure of the CAE that we used is depicted in Figure 4.

**Algorithm 1** Convolutional Autoencoder training

**Input:** Heatmap images $x^{(1)}, \ldots, x^{(N)}$

**Output:** encoder $f_{\emptyset}$, decoder $g_{\theta}$

$\emptyset, \theta \leftarrow Initialize\ parameters$

**repeat**

$$E = -\sum_{t=1}^{N} x^t \log\left(g_\theta\left(f_\emptyset(x^t)\right)\right) + (1 - x^t)(1 - \log\left(g_\theta\left(f_\emptyset(x^t)\right)\right))$$

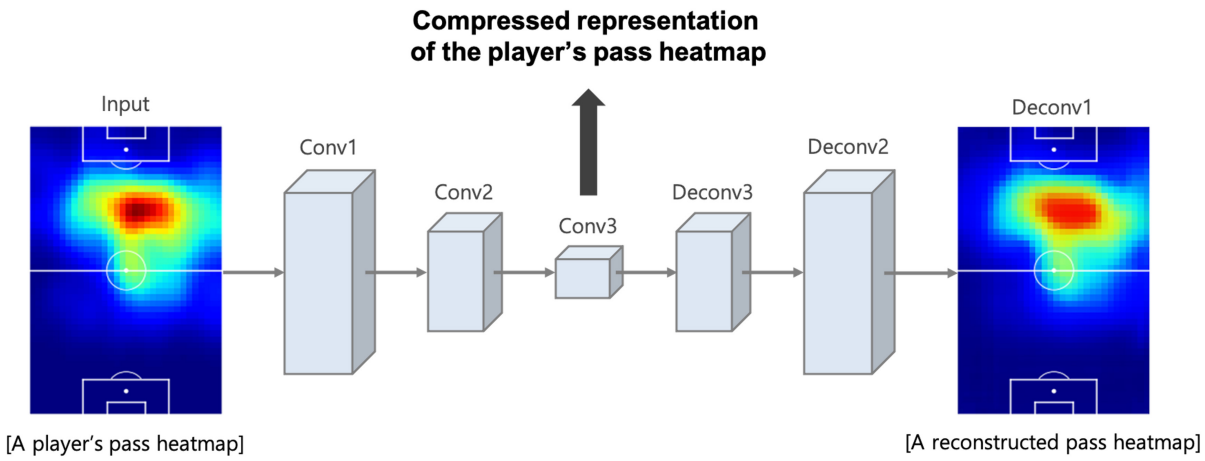
Calculate binary cross-entropy

$\emptyset, \theta \leftarrow$ Update parameters using gradients of $E$

**until** convergence of parameters $\emptyset, \theta$

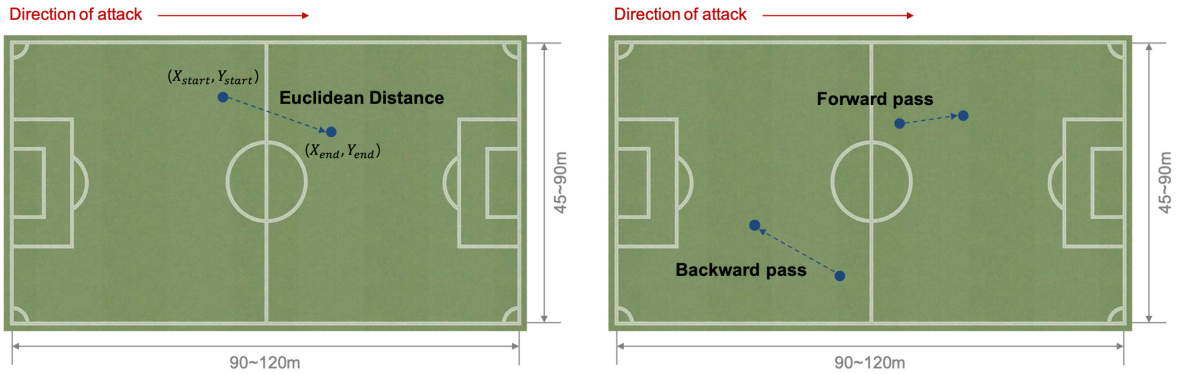
As a result, we compress the matrix $X''_{start}$ and $X''_{end}$ with shape $48 \times 32 \times 3$ to the matrix $X'''_{start}$ and $X'''_{end}$with shape $12 \times 8 \times 4$ (i.e., 384 dimensions). Then, we flatten the compressed heatmaps $X'''_{start}$ and $X'''_{end}$ to 1-dimensional vectors $x_{start}$ and $x_{end}$ of length 384. We eventually complete this step with vectors $x_{locStart}$ and $x_{locEnd}$ that contains the location information of the player's passes. As a result, we created a 384-dimensional vector from the heatmap with start locations and end locations of the player's passes.

[insert Figure 4.]




***Pass Length and Direction Vector.*** Next, we consider the length and direction of the player's passes (Figure 5). By considering the player's pass length vector, we can distinguish players who try to attack through long balls or who make organized attacks through short passes. As we previously calculated the pass length in the data preparation section, we estimated the proportion of long and short passes. A long pass is a pass with 30 meters or more in the distance,[33] on the contrary, a short pass is less than 30 meters. A player's pass length vector $x_{len}$ of length 2 consists of the proportion of long and short passes. We also calculated the direction of the pass. A forward pass is a pass played toward the opponent's goal and a backward pass is a pass played toward own goal. The proportion of a player's forward and backward passes, we generate the pass direction vector $x_{dir}$ of length 2.

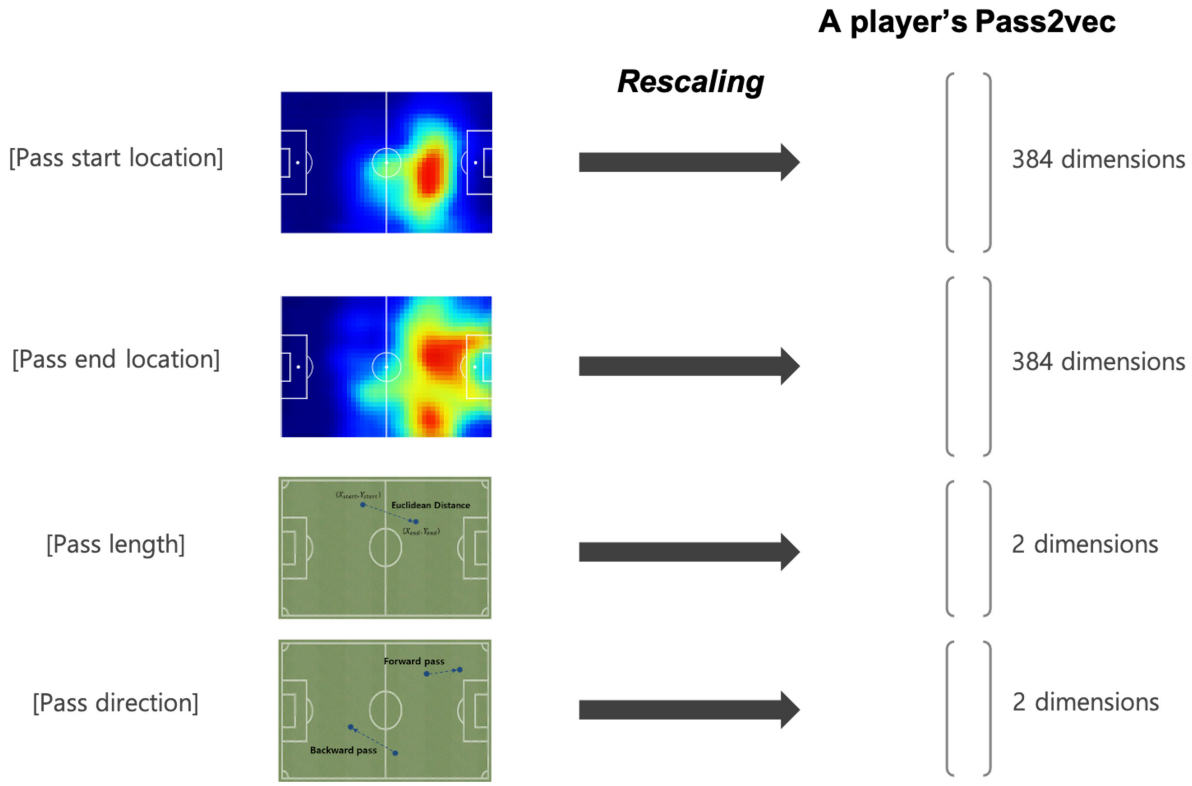
[insert Figure 5.]




***Pass2vec Composition.*** After generating vectors that include the information about the location, length, and direction of the player's passes, we concatenate the player's detailed pass vectors (i.e., $x_{locStart}$, $x_{locEnd}$, $x_{len}$ and $x_{dir}$). However, the dimension of the pass location vectors (i.e., $x_{locStart}$ and $x_{locEnd}$) are higher than the length and direction vectors (i.e., $x_{len}$ and $x_{dir}$) in two dimensions. This is due to compressing the heatmaps using CAE. It naturally creates a high-dimensional vector during the process as a minimum size of latent space for reconstruction is necessary to solve complex problems. Therefore, to treat vectors with fair importance, we adjust the weight of each aspect by rescaling the value of the pass

location vectors $x_{locStart}$ and $x_{locEnd}$. In this way, we can examine the similarity of passing style between two players by calculating the distance between their created Pass2vec. The overall process of Pass2vec is depicted in Figure 6.

[insert Figure 6.]




# Evaluation

To show the effectiveness of our proposed approach, we conducted both player retrieval and a player similarity analysis. The result presents the accuracy of our approach and its application to tactical decision-making on a player replacement for soccer teams.

### *Dataset*

We used soccer match event data from Pappalardo et al.[34] which is the biggest data set open to the public since 2019. The data, initially obtained from Wyscout, contains ball events in all matches in seven major soccer competitions and provides detailed information of involved players, position, time, the outcome, and the type of action. We analyzed ball event data of five national soccer competitions in Europe: first divisions in England, France, Germany, Italy, and Spain from the 2017/2018 season. [34,35] This is regarded as the most adequate data range for our analysis as we wanted the data in coherence with the team compositions and

directing style which might depend on the change of seasons. There are 2,619 players and 1,506,722 passes in the final dataset. A summary of the data is in Table 2 below.

**Table 2.** Summary of real soccer ball event data.

| No. | League | The number of matches | The number of teams | The number of players | The number of passes |
|---|---|---|---|---|---|
| **1** | EPL | 380 | 20 | 514 | 316,406 |
| **2** | Ligue 1 | 380 | 20 | 541 | 307,285 |
| **3** | Bundesliga | 306 | 18 | 473 | 252,431 |
| **4** | Serie A | 380 | 20 | 534 | 324,461 |
| **5** | LaLiga | 380 | 20 | 557 | 306,139 |

Before we extract and preprocess the data, we divided the entire match of each team into training data and test data. For both sets, we proceeded with the same data processing. Afterward, we extracted only pass event data from the whole ball event data to estimate the length and direction of passes. To calculate the actual lengths in meters, we multiplied 1.2 for width and 0.8 for height, which we assumed 120m x 80m as a general size of the professional soccer field.

### *Player Retrieval*

A player retrieval task is carried out through the following step. First, we divided data into two sets: training and test. We divided the sets from match ID and event ID in order, the first half and latter half of the whole season. From 2,619 players and 1,506,722 passes in the total dataset, we divided 751,395 (49.9%) for the training set and 755,327 (50.1%) for the test set. Then, we used the former to label vectors in Pass2vec. After we construct a Pass2vec of a target player using test data, we compare a target player's Pass2vec to a set of labeled Pass2vec and draw a top-k list of the most similar players to that player. We check if the actual player is on the top-k list. For example, we select the vector of player A in the training set and line up similar vectors in the test set. If we find the vector of player A in the Top-3 list, then we count player A in Top-3. We calculate the players in each Top-k list and show the percentage based on the total number of players as 100. Finally, we compared the accuracy of the task with the result from other literature to indicate the efficiency of Pass2vec.

First, we performed the player retrieval task by following the procedure mentioned above. To reduce outliers, we excluded players who played less than 810 minutes (i.e., the equivalent of approximately nine matches) in each training set and test set. Then, a total of 809 players were selected for retrieval. Also, we measured the distance between the two player's vectors through Euclidean distance, Cosine distance, and Manhattan Distance. The player retrieval with the Manhattan Distance showed the best performance in all cases, which method is usually preferred when there is high dimensionality in the data. Manhattan Distance is computed by:

$$dist_{Manhattan}(p, q) = \sum_{i=1}^{n} |p_i - q_i| \qquad (3.1)$$

, where $p = (p_1, p_2, \dots, p_n)$ and $q = (q_1, q_2, \dots, q_n)$.

Then, we compare the accuracy of the retrieval task over Decroos & Davis.[30] We first vectorized our data set with the NMF method from the previous study and followed the exact process of player vector generation and compared the top-k results of the player retrieval task with the result of Pass2vec. It also shows the comparison of the player retrieval task in which Pass2vec outperforms the previous approach. We also proved the best performance in the combination of multiple pass vectors.

**Table 3.** Top-k results of player retrieval task.

| | **Top-1** | **Top-3** | **Top-5** | **Top-10** | **Top-15** | **Top-20** |
|---|---|---|---|---|---|---|
| Player vectors (Decroos & Davis[30]) | 22.5% | 41.0% | 50.9% | 63.7% | 71.1% | 75.3% |
| Pass location vector | 21.9% | 34.6% | 43.0% | 54.1% | 60.4% | 66.5% |
| Pass length vector | 1.4% | 3.2% | 5.6% | 9.5% | 12.5% | 16.6% |
| Pass direction vector | 1.0% | 3.2% | 4.9% | 8.7% | 11.6% | 14.8% |
| **Pass2vec** | **29.4%** | **45.6%** | **54.3%** | **66.6%** | **72.4%** | **76.5%** |

To measure the effect from additional pass features such as length and direction, we performed the same experiments four more times with different divisions of the dataset for cross-validation. We created training and test sets from the random order of the whole dataset. The average percentage of task accuracy for Pass2vec is shown in Table 4.

**Table 4.** Top-k results of player retrieval task: cross-validation on length and direction

| | **Top-1** | **Top-3** | **Top-5** | **Top-10** | **Top-15** | **Top-20** |
|---|---|---|---|---|---|---|
| **Pass2vec** | **30.8%** | **47.9%** | **56.3%** | **67.6%** | **73.9%** | **78.0%** |
| **1** | 29.4% | 45.6% | 54.3% | 66.6% | 72.4% | 76.5% |
| **2** | 32.6% | 49.8% | 57.6% | 69.3% | 75.7% | 79.0% |
| **3** | 29.5% | 49.6% | 56.9% | 67.7% | 73.2% | 77.4% |
| **4** | 30.7% | 45.5% | 56.5% | 66.8% | 73.1% | 78.1% |

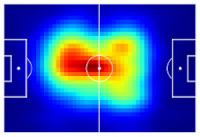
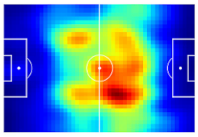
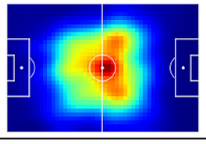
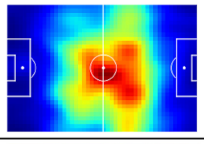
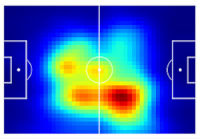
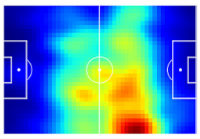
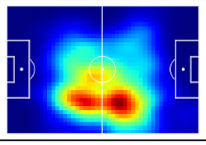
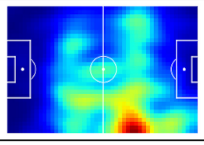
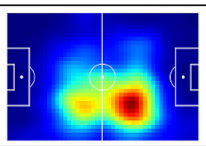
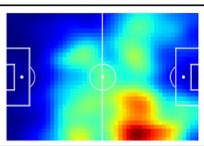
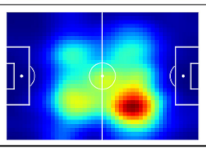
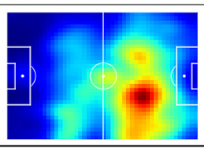
| | | | | | | |
|---|---|---|---|---|---|---|
| **5** | 31.8% | 48.8% | 56.0% | 67.6% | 75.3% | 79.0% |

*Player Similarity Analysis*

Soccer teams often search for adjustable players to replace a vacancy caused by a player's transfer of retirement. We suggest Pass2vec support decision-making by the team's coaches or professional scouts.

To create a pool of data, we used the same data used in the previous section. However, different from the previous analysis, we used the entire data to obtain the Pass2vec of players without dividing the training set and test set. We selected major players running 810 minutes and over during the whole season from the entire dataset. As a result, we finalized 1,701 players and compared the players' similarities. We present three examples of similar players using the player's Pass2vec. The selected are popular players well known in the media. A comparison of the two players is shown in Figure 7.

[insert Figure 7.]

| Comparison of Players | | Pass Location | | Pass Length | | Pass Direction | |
|---|---|---|---|---|---|---|---|
| | | Start | End | Long | Short | Forward | Backward |
| 1 | **Sergio Busquets** |  |  | 16.0% | 84.0% | 70.4% | 29.6% |
| | **Lucas Torreira** |  |  | 12.2% | 87.8% | 69.1% | 30.9% |
| 2 | **Ivan Rakitić** |  |  | 15.8% | 84.2% | 66.2% | 33.8% |
| | **Youri Tielemans** |  |  | 19.9% | 82.4% | 66.2% | 33.8% |
| 3 | **Luka Modrić** |  |  | 17.8% | 82.2% | 70.0% | 30.0% |
| | **Cesc Fàbregas** |  |  | 18.5% | 81.5% | 70.2% | 29.8% |

# Discussion

This paper presented a novel approach that aims to construct a player's passing style descriptor by taking the detailed information of the player's passes into account. One of the main contributions of this work is showing how descriptors can help the decision-making of scouts and training with individual pass characteristics. From the evaluation, we find this approach valid by comparing vectors of the players in both quantitative and qualitative ways.

The player retrieval task suggests that our proposed method outperforms the existing approach by successfully summarizing the player's passing style into a vector. We successfully retrieved 76.5% of all players in the top-20. The previous study resulted 75.3% from the same task. The performance gap is larger in Top-1 list which we showed higher percentage of accuracy of 29.4% over 22.5%. This implies that our Pass2vec effectively captures the player's unique characteristics. The major difference between the previous approach[36] and our method is in player style definition. The previous study characterized a player's playing style with locations of diverse actions performed by the player including passes, dribbles, crosses, and shots. We focused on passes and created a player's vectors with location, length, and direction of passes. It is shown that the various aspects of passes can represent the characteristics of the player's style and the result is more accurate when combining diverse pass features. On the other hand, location alone shows the highest contribution to the task compared to the other features. We could also recognize the impact from the length and direction of pass features which the previous study has not considered. We have validated the results from further experiments. As a result, the accuracy is consistent during the process in the range of 30.8% for Top-1 and 78.0% for Top-20. The average became higher than the initial dataset.

In this study, we also presented a player similarity analysis showing how a soccer team can use Pass2vec to search for players to replace a specific player. We provided three examples of the player similarity test. First, the most similar player to Sergio Busquets from our approach is Lucas Torreira. He is comparable to Busquets, an admirable passer with outstanding ball distribution capabilities.[37,38] The pass position, length, and directional descriptor of the two players show that they are very similar players. Also, Youri Tielemans is referred to be an alternative player of Ivan Rakitić,[39] which shows the same results in our method. Finally, the most similar player of Luka Modrić is Cesc Fàbregas. He is a creative playmaker with great vision, and this style exactly resembles Modrić.[40] These examples suggest that our approach is effective and meaningful. As we verified the practical use of Pass2vec with real data and players, we showed how it can support a soccer team's decision-making.

However, the limitations of the proposed method exist. First, we did not consider a team's tactic that a player is involved in. A style of a player can be seen as a combination of the player's skills but also the preferences and team tactics. The actions taken by the player in a match can be greatly influenced by team tactics. Recent work of Decroos et al.[41] uses an approach to capture both playing styles of teams and players and showed improved accuracy of a mixture model including the direction of actions.

Second, the problem of relativeness of Euclidean distance exists when estimating the length and location. As the dataset provided raw data in the percentage unit for positions, x and y coordinates in the range [0,100], the pass lengths and location might differ from actual meters. The actual width and length of each soccer field affect the value. To prevent this, we considered the one sized field (120m x 80m) in our study, but actual soccer fields could be considered for every match for improved accuracy of distance measurement.

Lastly, the high-leveled form of vector prohibits the immediate interpretation of the information. It is complex to extract individual pass factors from the created vector. For example, recognizing a tendency

of a player's passing location by a Pass2vec alone is difficult as it is compressed and flattened in a heatmap. Also, the length of the vector is relatively long, which makes it difficult for simple observation or comparison. This can be the challenge of most computer science based studies that the individual level of features is not interpretable compared to the traditional statistical data analysis.[42] This would require the collaborative involvement of sports scientists to relate the results to the actual performance and practical impact on players' behavior.

# Conclusion

We presented a novel approach to characterize a soccer player's passing style by using soccer data and applying the deep learning technique. As a player's passing style can be summarized with various aspects of a player's passes such as location, length, and direction, we created a player's passing style descriptor that reflects the detailed information of the player's passes. In our proposed method, incorporating various pass elements described the passing style more precisely, which extends the range of data-based analytic methods in characterizing a player's pass style. To consider the spatial information of a player's passes, we constructed a heatmap and compressed it using Convolutional Autoencoder (CAE) that shows its competency to preserve the information with a compressed format. We also generated a player's detailed pass vector about length and direction by computing the proportion of each kind of pass. Our two-step method-the preparation of a player's pass event data and construction of the Pass2vec-resulted in the form of vectors from which we can distinguish a player's passing style.

As a result, we enhanced the accuracy of the pass style description. By using the real soccer ball event data, we successfully identified a player from the player's anonymized match data. Pass2vec outperformed other approaches from the player retrieval task and we also showed its application on extracting similar players using our vectors. This suggests that our proposed method has the competency to support the soccer team's decision-making process.

As future work, we plan to extend the approach by characterizing a player's comprehensive playing style with detailed information of the player's various actions in matches. As we documented the passing style of players with Pass2vec, we expect to further analyze diverse actions in a soccer match to understand the player better and enhance the accuracy of analytics with advanced vectorization. We also plan to consider a tactic of the team to which a player belongs. This allows us to take a step closer to capture the player's style from the wide-ranging action in soccer matches.

# References


1. Herold M, Goes F, Nopp S, et al. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *Int J Sports Sci Coach* 2019; 14: 798–817.

2. Joseph A, Fenton NE, Neil M. Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Syst* 2006; 19: 544–553.

3. Pappalardo L, Cintia P. Quantifying the relation between performance and success in soccer. *Adv Complex Syst* 2018; 21: 1–29.

4. Rathke A. An examination of expected goals and shot efficiency in soccer. *J Hum Sport Exerc* 2017; 12: 22–23.

5. Haghighat M, Rastegari H, Nourafza N, et al. A review of data mining techniques for result prediction in sports. *Adv Comput Sci an Int J* 2013; 2: 7–12.

6. Hucaljuk J, Rakipović A. Predicting football scores using machine learning techniques. In: *Proceedings of the 34th International Convention MIPRO*. 2011, pp. 1623–1627.

7. Constantinou AC, Fenton NE, Neil M. pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Syst* 2012; 36: 322–339.

8. Schumaker RP, Jarmoszko AT, Labedz Jr CS. Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decis Support Syst* 2016; 88: 76–84.

9. Lucey P, Bialkowski A, Carr P, et al. Characterizing Multi-Agent Team Behavior from Partial Team Tracings : Evidence from the English Premier League. In: *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. 2012, pp. 1387–1393.

10. Le HM, Carr P, Yue Y, et al. Data-Driven Ghosting using Deep Imitation Learning. 2017, pp. 1–15.

11. Bialkowski A, Lucey P, Carr P, et al. Discovering Team Structures in Soccer from Spatiotemporal Data. *IEEE Trans Knowl Data Eng* 2016; 28: 2596–2605.

12. Fernando T, Wei X, Fookes C, et al. Discovering Methods of Scoring in Soccer Using Tracking Data. *KDD Work Large-Scale Sport Anal* 2015; 1–4.

13. Wei X, Sha L, Lucey P, et al. Large-scale analysis of formations in soccer. *2013 Int Conf Digit Image Comput Tech Appl (DICTA), 2013* 2013; 1–8.

14. Gyarmati L, Hefeeda M. *Analyzing In-Game Movements of Soccer Players at Scale*, http://arxiv.org/abs/1603.05583 (11 March 2016, accessed 3 January 2020).

15. Peña JL, Navarro RS. *Who can replace Xavi? A passing motif analysis of football players*, https://arxiv.org/abs/1506.07768 (2015, accessed 8 March 2020).

16. Bekkers J, Dabadghao S. Flow Motifs in Soccer: What can passing behavior tell us? *J Sport Anal* 2019; 5: 299–311.

17. Horton M, Gudmundsson J, Chawla S, et al. Automated classification of passing in football. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2015; 9078: 319–330.

18. Lucey P, Oliver D, Carr P, et al. Assessing team strategy using spatiotemporal data. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. Association for Computing Machinery, 2013, pp. 1366–1374.

19. Rein R, Raabe D, Memmert D. Human Movement Science " Which pass is better ? " Novel

approaches to assess passing e ff ectiveness in elite soccer. *Hum Mov Sci* 2017; 55: 172–181.

20. Spearman W, Basye A, Dick G, et al. Physics-Based Modeling of Pass Probabilities in Soccer. *MIT Sloan Sport Anal Conf Bost* 2017; 1–14.

21. Peña JL, Touchette H. *A network theory analysis of football strategies*, https://arxiv.org/abs/1206.6904 (2012, accessed 17 January 2020).

22. Yu Q, Gai Y, Gong B, et al. Using passing network measures to determine the performance difference between foreign and domestic outfielder players in Chinese Football Super League. *Int J Sport Sci Coach* 2020; 15: 398–404.

23. Meza DAP. Flow Network Motifs Applied to Soccer Passing Data. In: *Proceedings of MathSport International 2017 Conference*. 2017, pp. 305–319.

24. Goes FR, Kempe M, van Norel J, et al. Modelling team performance in soccer using tactical features derived from position tracking data. *IMA J Manag Math* 2021; dpab006: https://doi.org/10.1093/imaman/dpab006.

25. Brooks J, Kerr M, Guttag J. Using machine learning to draw inferences from pass location data in soccer. *Stat Anal Data Min* 2016; 9: 338–349.

26. Goes FR, Kempe M, Meerhoff LA, et al. Not Every Pass Can Be an Assist: A Data-Driven Model to Measure Pass Effectiveness in Professional Soccer Matches. *Big Data* 2019; 7: 57–70.

27. Fernández J, Bornn L. SoccerMap: A Deep Learning Architecture for Visually-Interpretable Analysis in Soccer. 2020; 12461: 491–506.

28. Power P, Ruiz H, Wei X, et al. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 1605–1613.

29. Arbues-Sanguesa A, Martin A, Fernandez J, et al. Using player's body-orientation to model pass feasibility in soccer. *IEEE Comput Soc Conf Comput Vis Pattern Recognit Work* 2020; 2020-June: 3875–3884.

30. Decroos T, Davis J. Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 569–584.

31. Lindström P, Jacobsson L, Carlsson N, et al. Predicting Player Trajectories in Shot Situations in Soccer. In: Brefeld U, Davis J, Van Haaren J, et al. (eds) *Machine Learning and Data Mining for Sports Analytics: 7th International Workshop*. 2020, pp. 62–75.

32. Masci J, Meier U, Cireşan D, et al. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In: *International conference on artificial neural networks*, pp. 52–59.

33. Tenga A, Holme I, Ronglan LT, et al. Effect of playing tactics on achieving score-box possessions in a random series of team possessions from Norwegian professional soccer matches. *J Sports Sci*

2010; 28: 245–255.

34. Pappalardo L, Cintia P, Rossi A, et al. A public data set of spatio-temporal match events in soccer competitions. *Sci data* 2019; 6: 1–15.

35. Pappalardo L, Cintia P, Ferragina P, et al. PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Trans Intell Syst Technol* 2019; 10: 1–27.

36. Decroos T, Davis J. Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams. In: Brefeld U, Fromont E, Hotho A, et al. (eds) *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Lecture Notes in Computer Science, vol 11908*. Springer, Cham, 2020, pp. 569–584.

37. Christian D. Why Lucas Torreira will be the perfect long-term replacement for Sergio Busquets. *Sportskeeda*, https://www.sportskeeda.com/football/why-lucas-torreira-will-be-the-perfect-long-term-replacement-for-sergio-busquets (2019, accessed 7 June 2020).

38. Khadilkar A. 5 Potential replacements for Sergio Buquests at Barcelona. *FanSided*, https://everythingbarca.com/2019/11/18/5-replacements-sergio-busquets-barcelona/2/ (2019, accessed 7 June 2020).

39. Lazar S. Forget Rakitic – 3 top quality midfield stars who Man United should target instead including this Belgian sensation. *Soccersouls*, https://www.soccersouls.com/manchester-united-should-not-be-looking-at-this-ageing-barcelona-star-here-are-3-alternatives-for-the-red-devils/ (2019, accessed 8 June 2020).

40. Hanagudu A. 6 players who can replace Luka Modric at Real Madrid. *Sportskeeda*, https://www.sportskeeda.com/football/6-player-replace-luka-modric-real-madrid/3 (2017, accessed 14 June 2020).

41. Decroos T, Van Roy M, Davis J. SoccerMix: Representing Soccer Actions with Mixture Models. In: *Proceedings of the 2020 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. 2020.

42. Goes FR, Meerhoff LA, Bueno MJO, et al. Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *Eur J Sport Sci* 2020; 21: 481–496.