

A Semi-Supervised Machine Learning Approach To Define Pressing Roles In Football

Andrew Peters^a, Nimai Parmar^a, Michael Davies^a and Nic James^a

^aFaculty of Science & Technology, Middlesex University, NW4 4BT, London, England

Abstract

A player's role for a team can be distinct from their playing position. Positions are generally attributed based on where the players line-up relative to their formation, whereas roles can be defined by frequency of their actions. Hence, the method presented in this research, attributed player roles based on event data. Player role feature selection involved a semi-supervised machine learning approach, that extracted feature importance in the form of Shapley values. These values helped define the KPIs for pressing attacking players. By using the proposed role similarity approach, it is possible for recruitment departments to identify players that occupy similar roles as current players. Furthermore, the evolution of player roles across time can be evaluated, which has applications with performance analysts, as they can interrogate the constituent roles of each player and its influence on overall team performance. Hence, the proposed method can help uncover the optimal KPIs for a given set of roles, while having practitioner applications within elite-level performance analysis and recruitment departments. Future methods should combine physical data sources, such as from tracking data, to enable greater specificity in player role classification.

KEYWORDS: PRESSING, DIMENSIONALITY REDUCTION, PLAYER ROLE, UMAP

Introduction

Pressing As a KPI

Key Performance Indicators (KPIs) have been developed to evaluate different aspects of football performance such as ball retention (pass completion %) or attacking threat (shots on target; Herold, et. 2021). A defensive performance can also be evaluated by KPIs, whereby a defensive tactic can involve restricting passing options and limiting space by applying defensive pressure (or pressing) (Hewitt, Greenham, & Norton, 2016). Pressing is a defensive tactic whereby players collectively exert pressure on the opponent with the ball, and the likely pass recipients, with the objective of winning the ball back, usually in the opponent's half (high press; Merckx, Robberechts, Euvrard, & Davis, 2021). As pressing is a very active yet tiresome action, it cannot be maintained during long spells of opposition possession, and hence its generally performed by teams that maintain established offence through high possession levels (Hewitt, Greenham, & Norton, 2016). Teams that press may have a pre-match strategy to target specific opposition players by enabling passes to them before collectively performing the press in an attempt to regain possession. The measurement of pressing is not straightforward as multiple players are involved and is hence difficult to capture using event data alone. Existing measures of these events include simple counts of pressure actions (as defined by a data provider), and their calculated proportions (such as the proportion of pressure actions in the opposition half), to evaluate a team's performance (Lago-Penas & Dellal, 2010). However, this can lead to inconsistent results as different data providers operationally define pressure events differently as some involve an opposition player entering a certain radial distance from the ball carrier (Closing Down: How Defensive Pressure Impacts Shots, 2022). The frequency of pressure events may, however, be skewed due to the variation in possession duration for different teams. Hence, pressure proxies have been created that normalise the volume of defensive actions (tackles, interceptions & fouls), based on the number of passes the opposition makes, to produce metrics such as passes per defensive action (PPDA; Trainor, 2014). The number of opposition passes is used in this calculation, as this has a significant correlation with possession time (Collet, 2013) and hence is a suitable proxy to measure opposition possession. Data providers such as *Statsbomb* use other methods to calculate possession-adjusted defensive metrics. The data provider proposed a simple adjustment that is proportional to the possession volume of a team, and also a more complex calculation that applies a sigmoid function to adjust their defensive metrics (Introducing Possession-Adjusted Player Stats, 2014).

Team & Individual Playing Styles

Football teams are often described based on their characteristic playing style (Lopez-Valenciano et al., 2021), although as football is a complex interaction of different players and their actions, it can be quite difficult to objectively classify different team playing styles (Low et al., 2020). In order to help classify different playing styles, football has previously been categorised into repeatable phases or moments. Four of these distinct moments (Oliveira, 2004) are: Established Attack, Established Defence, Offensive Transition and Defensive Transition. However, as Set Pieces account for approximately 30 % of all goals in major football competitions (Ensum et al., 2000), it was also considered as a fifth additional moment in subsequent analysis (Hewitt et. al., 2016). By classifying each match into these repeatable moment segments, practitioner performance analysts can associate Key Performance Indicators (KPI's; Herold et al., 2021) with each moment, and subsequently help glean different playing styles (Hewitt et al., 2016). However, there is a dearth of research that determines individual playing style using a data-driven approach. Those that exist explore the concept of player similarity, which determines what player is the most similar to a pre-defined player. For

example, previous research (Mazurek, 2018) used 17 variables to identify the most similar player to *Lionel Messi*. Other research (García-Aliaga, 2021) has analysed passing sequences to identify the successor to Spanish midfielder *Xavi Hernandez*, by using dimensionality reduction and unsupervised clustering techniques. Although these are sound methods, they are supervised in nature, whereby comparisons are made to a pre-defined player. Hence, there is scope to create a more objective method to classify player roles, whilst also using a more comprehensive dataset that considers defensive as well as in-possession variables.

Roles And Positions

There is a distinction to be made about a player's position and their tactical role (which can be approximated by their individual playing style). Previous research has detected roles based on their active positions, as opposed to considering individual playing styles within these roles. For example, different methods have detected dynamic formations (Kim et al., 2022; Bialkowski et al., 2014), while others (Meerhoff et al., 2019; Carrilho et al., 2020) have used tracking data to detect subgroups and hence positional roles within these dynamically changing formations. However, an approach that does consider both playing style and the positional role they occupy was considered by Li et al. (2022). The authors constructed player vectors using Nonnegative Matrix Factorization (NMF), which ultimately classified 18 different individual playing styles. However, the authors used only 10 basic events to describe the player actions to construct their player vectors, with no evidence of more advanced metrics or reasoning for their feature selection. A method that incorporated many more advanced features and adopted a feature selection method was that conducted by Aalbers & Van Haaren (2020). However, the classification of roles in this method is supervised, as the 21 playing roles were created in consultation with domain experts. The aim of the proposed method is to use multiple advanced metrics to classify player roles, as well as utilise a semi-supervised/unsupervised method to eliminate subjectivity or pre-conceived practitioner biases. A barrier for unsupervised approaches that it is difficult to assign the most important event variables to each constituent role in an objective manner. To overcome this, it is possible to implement a semi-supervised machine learning approach to understand the feature importance of variables associated with a specific player role. Hence, these importance (or SHAP; Lundberg & Lee, 2017) values can be inferred by algorithms such as XGBoost and used for feature selection of optimal variables to distinguish different player roles. However, as data providers have vast datasets of complex player event variables, to classify and compare roles amongst many players, it is required to perform a dimensionality reduction.

Dimensionality Reduction

As the volume of data is increasing across many different industries, such as the data available to football clubs, so does its complexity and dimensionality. A way to overcome this complexity issue is by reducing the dimensionality of these datasets, as this may help remove the noise from redundant features whilst also preserving the variance of all the essential features. One of the most common methods of dimensionality reduction is principal component analysis (PCA). However, because PCA projection identifies directions of maximal variance in the data and ignores variation along other directions, it tends to obscure finer-scale patterns within the data (Diaz-Papkovich et al., 2021). Many nonlinear neighbour graph-based dimension reduction algorithms, such as t-SNE (Maaten et al., 2008), have been developed to overcome this limitation. A further method uses Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018) for dimensionality reduction, which has seen widespread use across fields (e.g., single-cell genomics (Becht, et al., 2019)). Rather than trying to preserve large-scale structure, UMAP seeks to preserve local neighbourhoods in a

dataset (Diaz-Papkovich et al., 2021). Hence, this method finds a low-dimensional representation of the data that preserves these neighbourhoods as much as possible, which will serve a purpose in organising similar player roles in close approximation.

However, the growing volume of data in football has been exploited in a limited way so far (Cintia et al., 2015). Hence, there is a scarcity of robust literature that underpins using dimensionality reduction techniques in the context of football. Existing methods have applied these techniques to demonstrate match outcome prediction for a reduced set of features (Głowania, Kozak, & Juszczuk, 2023), whereas others have used event data to classify different playing positions amongst players (García-Aliaga, et al., 2021; Lopes & Tenreiro Machado, 2021) and different playing styles amongst teams (García-Aliaga, et al., 2021). However, there's a large scope for development on previous position classifier methods (García-Aliaga, et al., 2021; Lopes & Tenreiro Machado, 2021), as it is possible to further classify playing positions into a series of role subgroups.

Clustering

Clustering algorithms provide an objective way to classify data into a set of groupings (clusters), or in our case, player roles. The K-means algorithm is the best known and most frequently used for clustering, which divides the data set into k clusters by minimizing the sum of all distances to the respective cluster centres (Ramos et al., 2015). This is an example of hard clustering method, which assigns a datapoint to a given cluster, even if the datapoint may intersect two different clusters. Contrasting this, clustering methods such as Gaussian Mixture Modelling (GMM) adopt a soft clustering approach. As the name implies, a Gaussian mixture model involves the mixture of multiple Gaussian distributions. Rather than identifying clusters by nearest centroids, a group of k gaussian distributions are fit to the data. After learning the parameters for each data point, it is possible to attribute a probability of it existing within any of the k clusters. Hence, soft clustering such as GMM, can grant a probability of a given player being assigned to a set of roles (or clusters), which can't be achieved by K-Means. In general, clustering techniques occur after UMAP dimensionality reduction, particularly if the number of variables used is large. This is because there is a reduced amount of data to be classified, with a lot of redundant information removed, whilst the data composition is generally well-preserved (McInnes et al., 2018).

Existing methods in football have applied clustering techniques to classify formations (Narizuka & Yamazaki, 2019) and passing networks (Kawasaki, Sakaue, Matsubara, & Ishizaki, 2019). There is an existing method that clusters players (Lopes & Tenreiro Machado, 2021), however the authors analysed player statistics originating from the video game *FIFA* as their data source. Hence, actual on-ball player event data from game to game may accelerate findings in this field, rather than theoretical ratings of each player attribute on a scale of 1-100.

Overall Aims

Overall, the aim of this study was to create a role classification tool with utility amongst practitioners, with a particular emphasis on classifying pressing player roles (Fig 1). The optimal variables that separate these player roles will be identified using a semi-supervised feature selection method. Using these optimal variables, a UMAP (dimensionality reduction) will be performed, and player roles will be attributed using GMM (Clustering). Hence, it will be possible to identify different players fulfilling similar pressing roles within each cluster. A method to quantify role similarity will then be proposed, that enables an objective way to measure players with similar pressing qualities. This will have implications within practitioner recruitment departments, as players with similar pressing qualities to current players can be identified using this classification system.

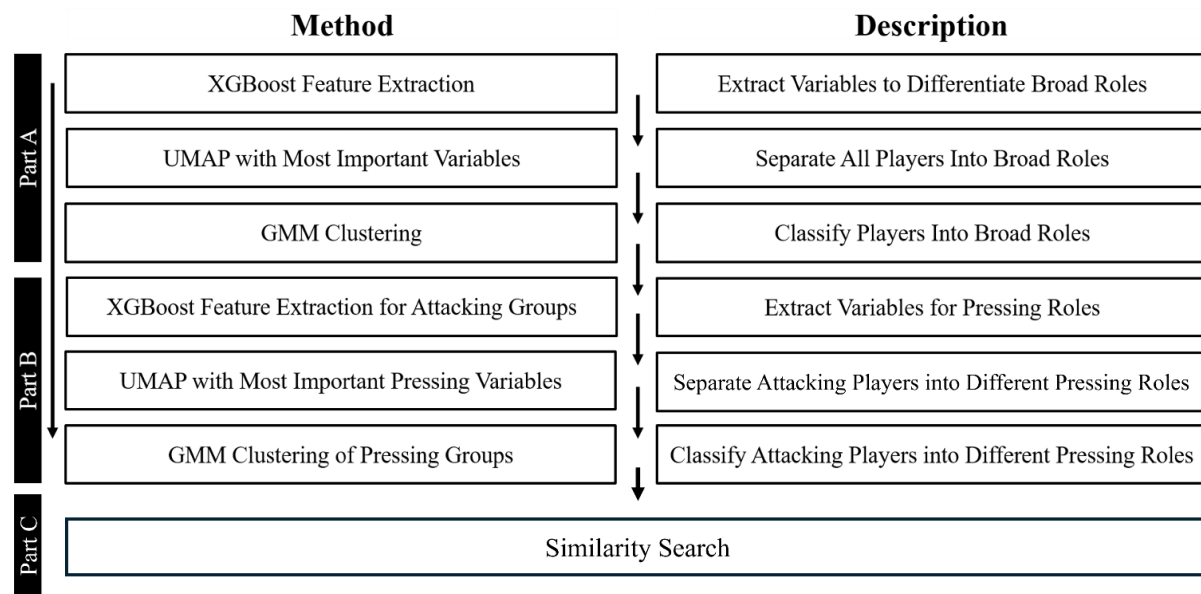


Figure 1: Flow chart of methodology used for current research.

Methods

Sample

Player event data was downloaded using user credentials from *Statsbomb* (Statsbomb 360, 2022). These were 72 player variables available to the Leicester City subscription, that was normalised on a per 90-minute basis. To ensure a large sample to analyse, data was taken across eighteen different competitions from the 2022/2023, 2021/2022, 2020/2021, 2022 & 2021 seasons as labelled by *Statsbomb* (Table 1). Players with less than 540 minutes (or six X 90 minutes) were excluded from analysis due to their limited playing time. Given that *Statsbomb* had a large volume of unique primary position entries, it was required to break these down into smaller labels. Following practitioner consultation, it was agreed to aggregate them into seven positional identifiers: Goalkeeper; Fullback / Wing Back; Centre Back; Defensive Midfielder; Central / Attacking Midfielder; Winger; Forward (Table 2).

Table. 1 Competitions Used in Data Sample

League	Territory	Seasons
1. Bundesliga	Germany	2020/21; 2021/22; 2022/23
Bundesliga	Austria	2020/21; 2021/22; 2022/23
Champions League	Europe	2020/21; 2021/22; 2022/23
Championship	England	2020/21; 2021/22; 2022/23
Eredivisie	Netherlands	2020/21; 2021/22; 2022/23
Jupiler Pro League	Belgium	2020/21; 2021/22; 2022/23
La Liga	Spain	2020/21; 2021/22; 2022/23
Liga Nos	Portugal	2020/21; 2021/22; 2022/23

Liga Profesional	Argentina	2021; 2022
Ligue 1	France	2020/21; 2021/22; 2022/23
Premier League	England	2020/21; 2021/22; 2022/23
Serie A	Italy	2020/21; 2021/22; 2022/23
Série A	Brazil	2021; 2022
Super League	Switzerland	2020/21; 2021/22; 2022/23
Superliga	Denmark	2020/21; 2021/22; 2022/23
UEFA Europa Conference League	Europe	2020/21; 2021/22; 2022/23
UEFA Europa League	Europe	2020/21; 2021/22; 2022/23

Table 2: Positional Identifiers

Statsbomb Position Label	Positional Identifier
Centre Attacking Midfielder	Central / Attacking Midfielders
Left Centre Midfielder	Central / Attacking Midfielders
Right Centre Midfielder	Central / Attacking Midfielders
Centre Back	Centre Back
Left Centre Back	Centre Back
Right Centre Back	Centre Back
Centre Defensive Midfielder	Defensive Midfield
Left Defensive Midfielder	Defensive Midfield
Right Defensive Midfielder	Defensive Midfield
Centre Forward	Forward
Left Centre Forward	Forward
Right Centre Forward	Forward
Left Back	Fullback / Wing Back
Left Wing Back	Fullback / Wing Back
Right Back	Fullback / Wing Back
Right Wing Back	Fullback / Wing Back
Goalkeeper	Goalkeeper
Left Attacking Midfielder	Winger
Left Midfielder	Winger
Left Wing	Winger
Right Attacking Midfielder	Winger
Right Midfielder	Winger
Right Wing	Winger

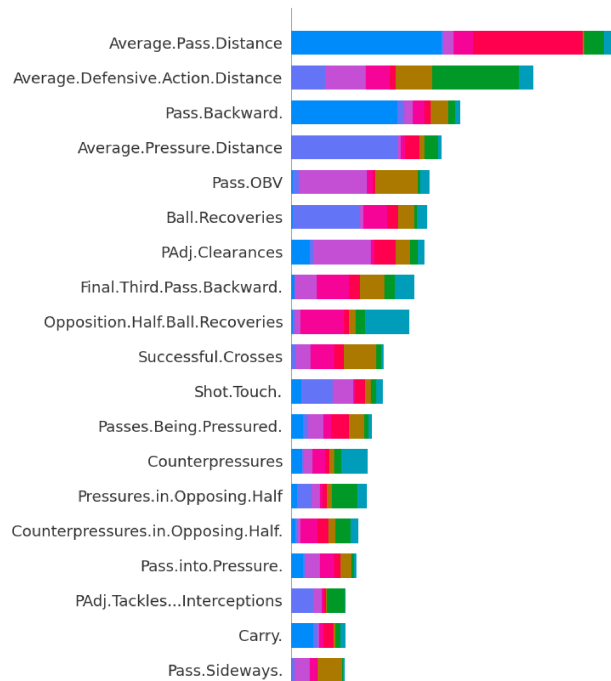


Figure 2. Bar Chart with SHAP values for the most important variables for broad position classification.

Dimensionality Reduction

A dimensionality reduction was performed with Uniform Manifold Approximation and Projection (UMAP), using the nineteen most influential variables as per their SHAP value. The output was visualised and coloured based on the seven positional identifiers defined earlier (Fig. 3).

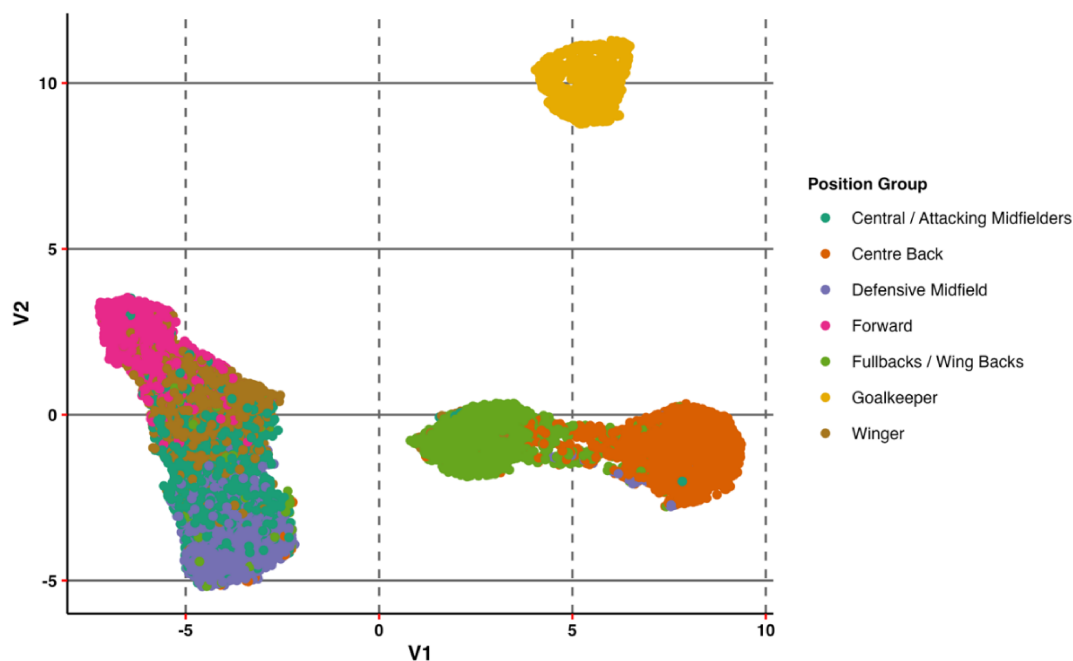


Figure 3. Positional Classification following Uniform Manifold Approximation and Projection (UMAP).

GMM Clustering

To classify these broad player positional roles, Gaussian Mixture Models (GMM) Clustering was performed. Following GMM, the data points were classified into seven different broad positional grouping (Fig. 4). The result is clusters of player datapoints that are derived entirely from the nineteen situational variables, that give broad groupings of their roles. A decision boundary was applied that classified each player to a role if that cluster probability was 50% or greater.

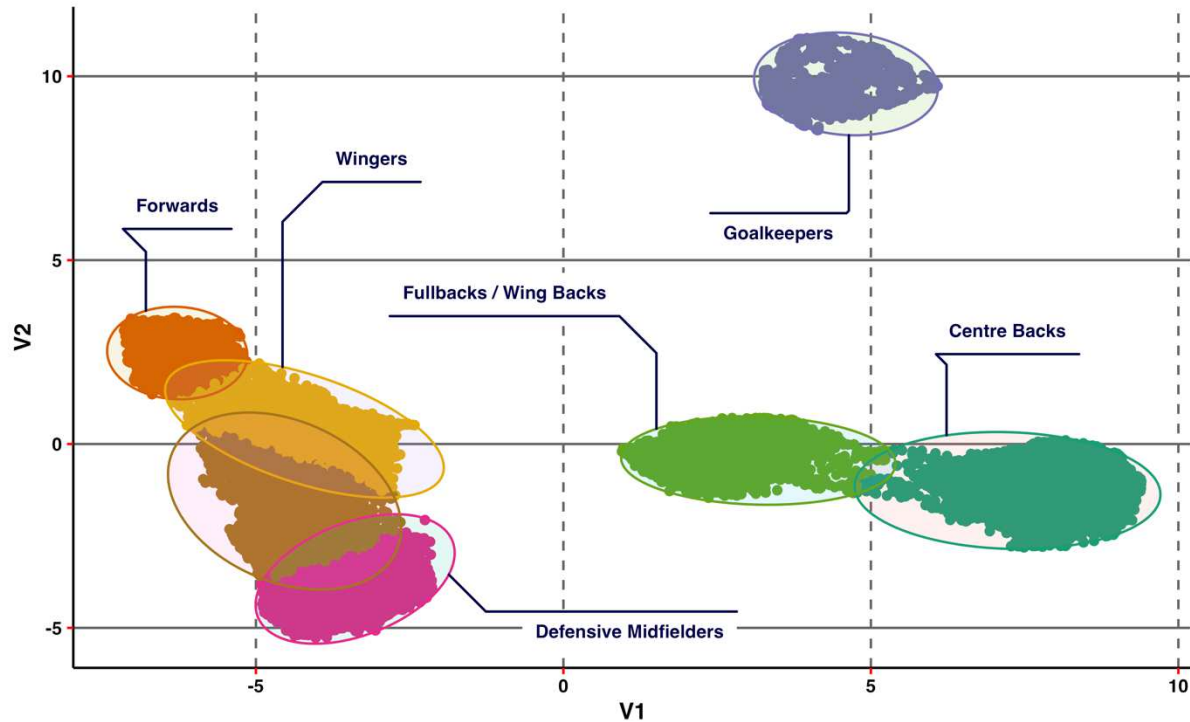


Figure 4. Positional Classification following Uniform Manifold Approximation and Projection (UMAP) and subsequent Gaussian Mixture Model (GMM) Clustering.

Extracting Pressing KPIs

To extract pressing KPIs, teams with a high pressing volume had to be identified. Team data was downloaded from *Statsbomb* using the same eighteen different competitions and seasons discussed earlier. The pressing variable, Passes per Defensive Action (PPDA; Trainor, 2014), was used to approximate high pressing. These values were converted to percentiles, and teams laying within the lower quartile were classified as high pressing teams and the remainder were classified as non-high pressing teams. This team classification information was appended to the initial player data frame. Hence, every player was labelled from a high-pressing or a non-high-pressing team. As pressing behaviour is characterised by players in advanced pitch areas, the *Wingers* and *Forwards* role group (classified earlier), underwent further analysis to identify the most important pressing variables. For this research paper, the method to analyse *Wingers* is detailed henceforth. The important pressing variables were established using the same method as earlier, with the 1) feature matrix being all numeric variables for players classified as *Wingers* following GMM and 2) a binary label that detailed whether players performed in either pressing or non-pressing teams. After fitting of an XGBoost model with default parameters, the variables with the highest feature importance were plotted (Fig. 5). Hence,

these were the top four pressing KPIs that separate *Wingers* from high-pressing teams and non-high pressing teams.

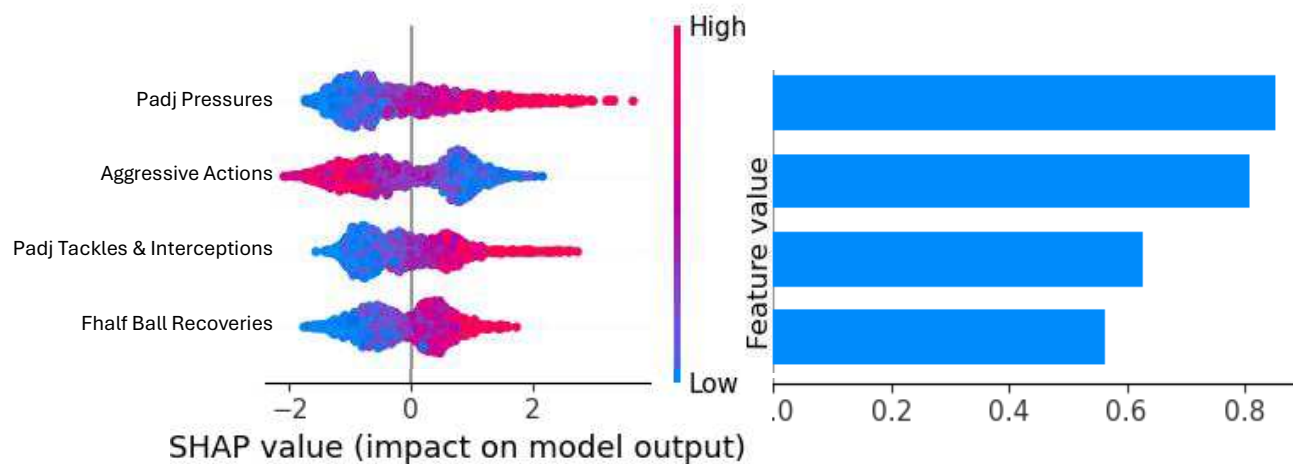


Figure 5. SHAP Value Plot for Winger Classified Group to Identify Optimal Pressing KPIs

Secondary Dimensionality Reduction & GMM Clustering

Using the *Wingers* classified group, the players underwent a UMAP dimensionality reduction using the top four optimal pressing variables: *Padj Pressures*; *Aggressive Actions*; *Padj Tackles & Interceptions*; *Fhalf Ball Recoveries*. The *Aggressive Actions* variable was defined as the number of tackles, pressure and fouls a player makes within 2 seconds of an opposition ball receipt. The *Fhalf* prefix denotes *final half* or actions that take place in the opposition half. Hence the *Fhalf Ball Recoveries* variable denotes the volume of ball recoveries that a player makes in the opposition half. Following UMAP, the *NbClust* package (Charrad et al., 2014) identified four clusters as the optimal number to separate the reduced *Wingers* data into constituent GMM clusters.

Role Similarity

Role similarity was measured by calculating the smallest Euclidean distance between a specific player to all other players present in the dataset that met a pre-defined search criterion. The player (Eberechi Eze) was randomly selected and used as a *proof-of-concept* example for further analysis. A theoretical search criterion was defined as *Wingers* from the most recent 2022/23 season from Europe's Top five leagues (English, French, Spanish, Italian, German). The Euclidean distance was calculated as the geometric distance between Eze's datapoint and all other datapoints on the dimensionally reduced space. The percentage similarity was declared as the proportion of the distance for each player's data point over the distance of the maximum datapoint away from Eze.

Results

Dimensionality Reduction

The positional groupings aggregate well following dimensionality reduction (Fig.3). Small groupings of different colours may indicate players that either have been falsely labelled by the data provider or occupy different roles than is normal given their pre-labelled position. The latter is generally the case as a player's role can intersect different hard-coded positions.

Cluster Analysis

The clusters were labelled a subjective role name, which was informed by the underlying data composition. The *Pressers* group had the largest number of possession-adjusted pressures, tackles and interceptions. The *Passive* group appeared to have the lowest volume for all actions subject to analysis. The *Ball Recoverers* group was very similar to the *Passive* group, with the difference being they scored highly for opposition half ball recoveries. Finally, the *Intermediate Pressers* group appeared to mimic the *Pressers* group but yielded far less possession-adjusted tackles and interceptions. What must be noted, is that half (two) of the variables used are possession-adjusted (*Padj Pressures*; *Padj Tackles & Interceptions*), whereas one other is semi possession-adjusted (*Fhalf Ball Recoveries*) as it is normalised for opportunities in the opposing half. Hence, players with high volume non-adjusted variables, but with low volume adjusted variables, may be an artifact of playing for low possession teams. This is potentially the reason why the *Intermediate Presser* group engages in slightly more aggressive actions than the *Presser* group as indicated by the ridge plot yet perform a lower volume of actions that were subject to possession adjustment.

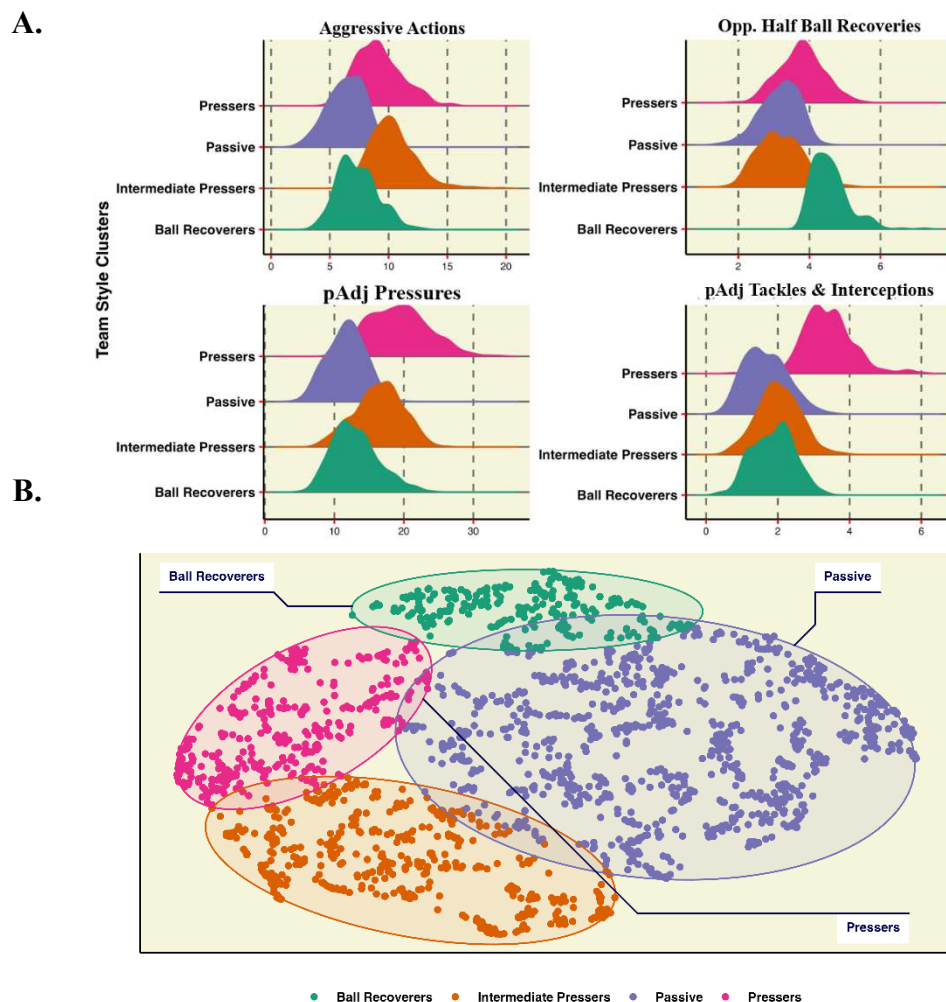


Figure 6 (a) Ridge plot displaying distribution of optimal variables within each cluster. (b) UMAP projection onto 2-dimensional space followed by GMM of optimal pressing variables.

Role Similarity Output

The top five most similar players to Eberechi Eze were tabulated (Table 3) and displayed visually below (Fig. 7). The output is a series of players that are most similar in fulfilling Eze's

Intermediate Pressing role. Fellow Crystal Palace teammate, Wilfried Zaha, had the closest similarity to Eze. This suggests that Crystal Palace have well-defined pressing objectives for their wingers, or that they may recruit players with similar qualities that are capable of fulfilling this pressing role.

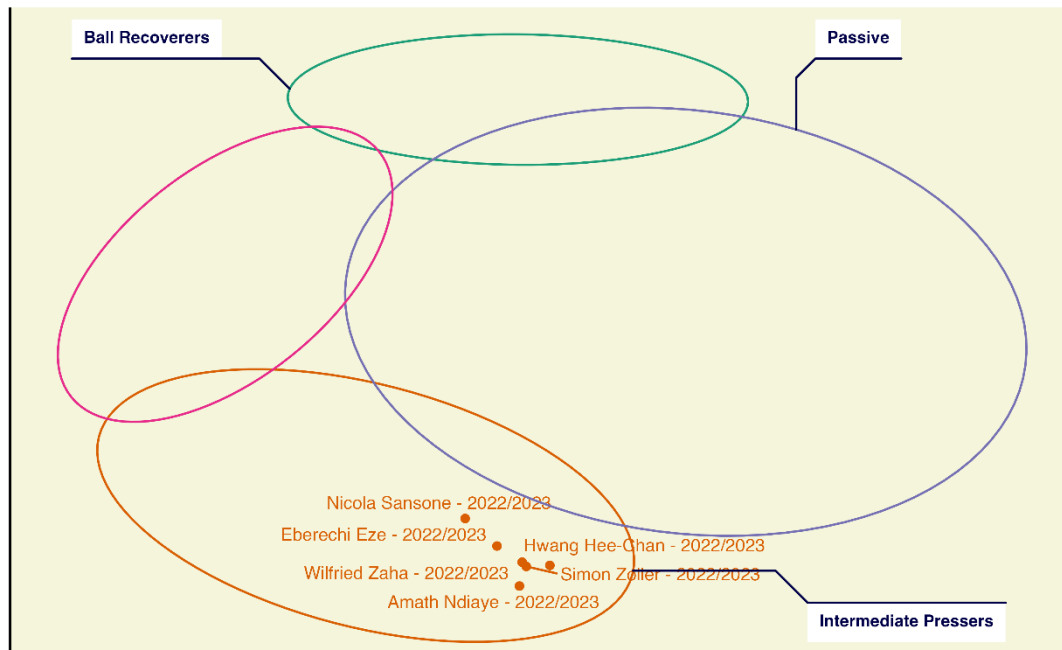


Figure 7: UMAP projection onto 2-dimensional space of players with closest similarity to Eberechi Eze from the 2022/2023 Season.

Table 3: Table of Players Similar to Eberechi Eze following Similarity Search

Player	Competition	Team	% Similarity	Distance
Wilfried Zaha	Premier League	Crystal Palace	94.82851	0.36
Simon Zoller	1. Bundesliga	Bochum	93.78860	0.43
Nicola Sansone	Serie A	Bologna	92.53857	0.51
Amath Ndiaye	La Liga	Mallorca	91.36516	0.59
Hwang Hee-Chan	Premier League	Wolverhampton Wanderers	90.60578	0.64

Evolution Of Roles

The pressing role of a player may evolve over time. There are numerous reasons for this such as: a new manager may bring different pressing tactics to previous managers; a player may suffer a dip in form; or a player may suffer a significant injury. As the soft clustering nature of GMM attributes probabilities based on the clustering classifications, it is possible to investigate how a player is associated with different role clusters over time. This was achieved using data from Eberechi Eze (Fig. 8). Eze started as a passive player in 2020/21 before nearly being classified in the *Presser* group a year later. During the most recent 2022/23 season, Eze fully established himself as an *Intermediate Presser* (94 % classification). This evolution is interesting, particularly after considering the development of Crystal Palace as a pressing side with the accompanying PPDA data established earlier. With percentiles 0-25 classified as a pressing team, Crystal Palace started as extremely low pressing intensity during the 2020/21 season (97th Percentile), before adopting a higher pressing style the following season (43rd Percentile), before reverting back to a non-pressing style (91st Percentile) in 2022/23. Perhaps these differences are due to a managerial change across the seasons, as Patrick Vieira adopted a more proactive pressing style during his only full season as manager (2021/22 season). Therefore, although the team as whole didn't press in seasons 2020/21 and 2022/23, the role of Eze did alter substantially.

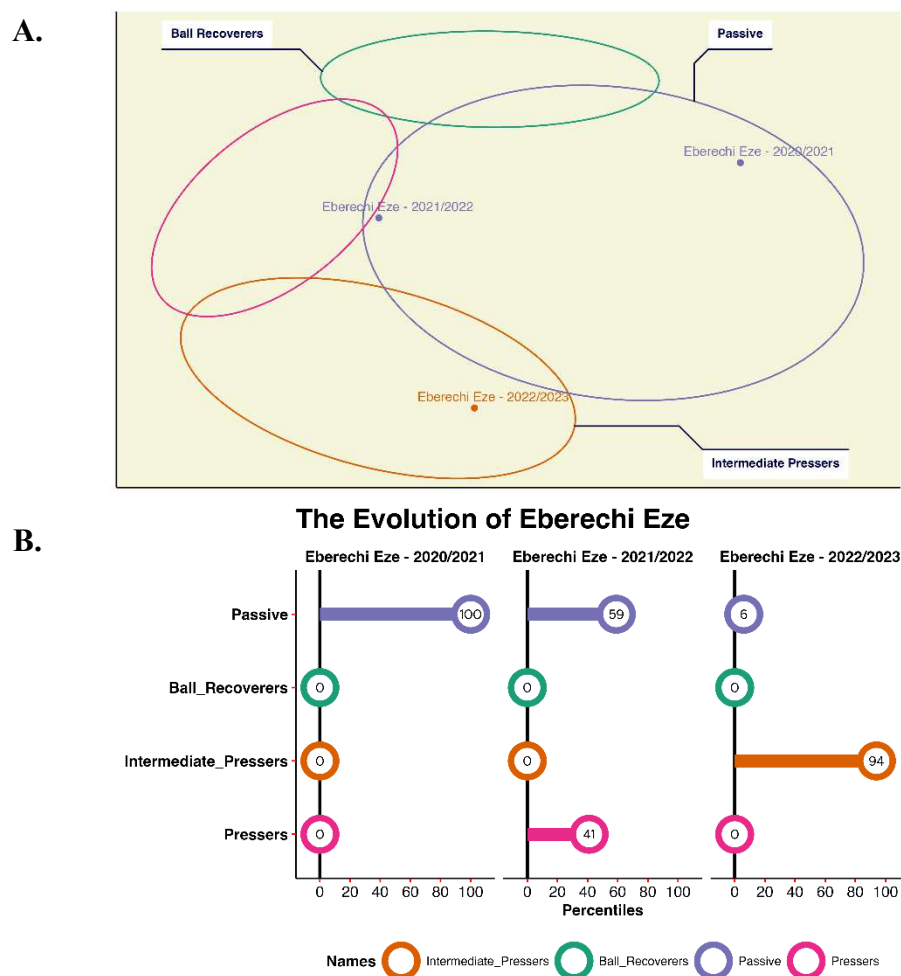


Figure 8: (a) Dimensionality reduction & (b) lollipop chart displaying the evolving pressing role for Eberechi Eze over a 3-year season span.

The Relationship Between Possession & Pressing

Possession dominant teams were labelled as those with ball possession across the season greater than the 75th percentile. A chi-square test was performed for players from possession dominant teams and pressing teams, which established a significant association between the two variables ($p < 0.001$). The relationship between the optimal pressing variables and the players labelled within possession teams was also investigated. The variables: *Padj Pressures* ($p < 0.001$), *Padj Tackles & Interceptions* ($p < 0.001$), *Fhalf Ball Recoveries* ($p < 0.01$) & *Aggressive Actions* ($p < 0.001$) all reached significance following a two-sample t-test. These variables all had higher values within the possession team group, with the exception of aggressive actions which had a greater volume amongst non-possession dominant sides. This suggests the importance of possession adjusted variables, as high aggressive output may be an artifact of playing in a non-possession dominant side. Interestingly, the same trend was observed between *Aggressive Actions* and pressing teams ($p < 0.001$). This suggests that a higher volume of aggressive actions is indicative of players that are not part of pressing teams (Fig. 9).

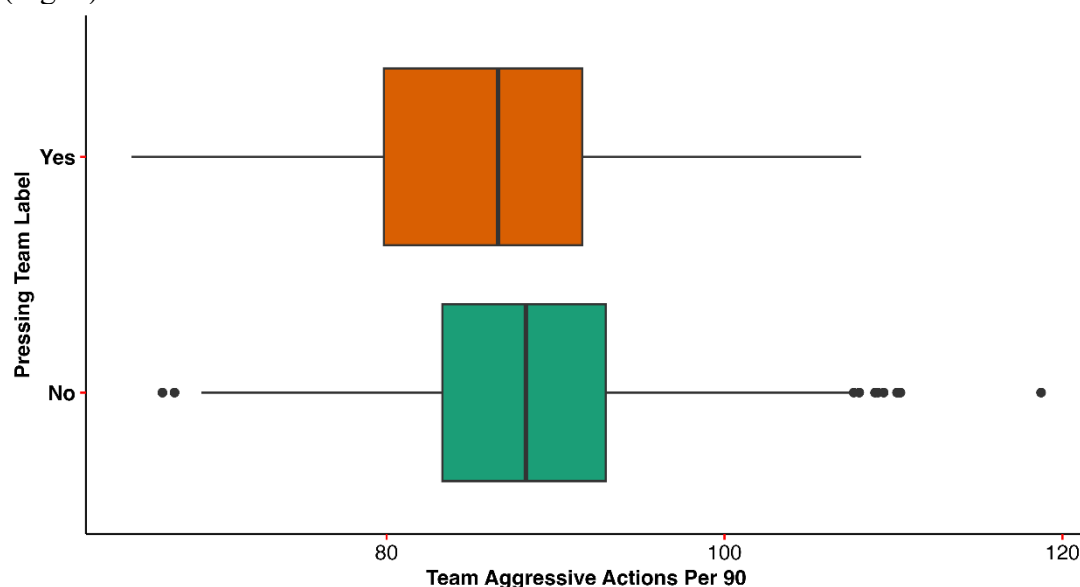


Figure 9: Boxplot showing the significant difference between pressing teams and aggressive actions.

Discussion

The aim of this study was to produce a role identification method and role similarity tool that has utility for analyst and recruitment staff at the practitioner level. There is some evidence within existing literature of methods that classify player positions using technical-tactical variables (Mazurek, 2018; García-Aliaga et al., 2021). However, there is large scope to improve current role identification methods. Existing methods have used limited initial datasets (Lopes & Tenreiro Machado, 2021); supervised classification to label roles which may include inherent biases (Aalbers & Van Haaren, 2020) and use basic metrics for role classification, without detail of their feature selection methods (Li et., 2022). Hence, the proposed method develops on previous dimensionality reduction and clustering practices by classifying specific player roles using *Statsbomb* player event data. The novelty in this method, is that by using a semi-supervised approach, the most important variables attributed to player pressing roles were identified in a semi-objective manner. To our knowledge, this is the first method of its kind that uses SHAP feature importance values as a feature selection method for the optimal variables to discern different player pressing roles in football.

Using the *Wingers* classified group, the optimal variables that discriminated between players at pressing and non-pressing teams were identified. These were determined optimal KPIs for wingers and can subsequently be sought after by player recruitment strategies at the elite-level, that look to recruit players that fit a pressing role. Interestingly, the defensive possession-adjusted variables appear to have high importance when identifying pressing related players. This is underlined by the fact that high aggressive volume has a significant association with non-pressing teams ($p < 0.001$), suggesting that non-adjusted defensive metrics may be artifacts of playing in a non-possession dominant side. This conveys the importance of adjusting variables by possession whilst also suggesting that pressing teams tend to dominate possession as they win the possession quickly following possession loss (due to pressing). Going forward, different possession adjustments could be trialled in order to optimally approximate the volume of different player defensive actions.

A method with further utility at the practitioner level is the role similarity tool, that determines similar players within pressing roles. Using this tool, it is possible to get similar pressing profile for a) an optimal pressing player as identified by a recruitment network or b) current players that may leave. Recruitment is particularly difficult when high quality players leave, and hence by using similarity searches such as this, it will help identify players that fulfil similar roles to current ones. This, in theory, should ease the adaptation period when a new player enters a team, as they perform a similar role to previous players. In addition, filters can be applied such as specific leagues or price tag within the search, so the optimal pressing players for specific markets and financial constraints can be investigated.

A further utility of the classification method is that player roles can be assessed over time. This can be important for analysts in identifying how players' tactical roles can alter within a given timeframe. Additionally, it can help identify a timepoint of stylistic or role change. Ultimately, many factors may contribute to a player evolving roles such as aging, a persistent injury or a managerial change that adopts a different style. However, although the team-level pressing intensity may change, the pressing intensity of every player may stay constant or alter in an inverse direction to the team. Hence, this stresses the importance of analysis of roles at the player-level to understand their constituent roles in relation to team setup.

Limitations and Future Work

This method adopted a semi-supervised approach, which used SHAP values of importance to identify optimal variables of importance. However, default parameters when implementing the XGBoost algorithm were used and hence these hyperparameters can be iterated until the optimal ones are identified and implemented. Additionally, a lot of variables were used in the feature matrix as input for the XGBoost algorithm. Perhaps this should be subject to a smaller set of features, such as defensive related features when establishing pressing roles, to prevent overfitting of the model by inclusion of unnecessary variables.

A further limitation of the analysis is the nature of the dataset. To our knowledge, there hasn't been published work regarding the validation accuracy of *Statsbomb* data, although previous research has constructed highly accurate models using their datasets (Peters et al., 2024), suggesting the validity of the data source. In addition, although *Statsbomb* do provide datasets (Statsbomb 360, 2022) with positional information, this study was conducted on an event dataset alone. Future methods should incorporate other data sources such as from positional/tracking data to establish more niche roles. For example, incorporating physical metrics from tracking data will help refine pressing roles, as they can be characterised by off-the-ball running intensity. In addition, by combining event and tracking data, it is possible to create more advanced metrics to approximate player roles. For example, *Rest Defence* (Peters et al., 2025) variables and output from sophisticated pressing models (Andrienko et al., 2017) should more accurately characterise these player roles.

Another limitation is due to the nature of the clustering. The unsupervised approach performs the clustering based on the signal within the data, but it is up to the data scientist to produce labels that match the language used within each club. Many scouts produce similar but different words to review players, and hence the terminology of role definition must be consistent across the club to enable comprehension. This is imperative for data science to practitioner translation, so that the exact nature of the role doesn't get lost in translation.

Overall, roles will be specific to a manager or a style of play, and hence should be described by senior representatives at football clubs. From here it should be possible to separate and classify all potential player roles using different metrics as required for the club. This will have obvious utility in the transfer market, but also for performance analysts as they can evaluate if a player is successfully performing their required role over time.

Conclusion

A player's role for a team can be distinct from their playing position. Positions are generally attributed based on where the players line-up relative to their formation, whereas roles are defined by frequency of actions. By using a large dataset comprising of 72 event variables across 17 different leagues, the proposed method attributed roles based on event data. Feature selection involved training a machine learning model, that extracted feature importance in the form of Shapley values. These values helped define the KPIs for pressing *Wingers*, which can be used in future recruitment strategies. In addition, using a role similarity search, it is possible for recruitment departments to identify targets that occupy similar roles as current players. The method also has utilities amongst performance analysts as the evolution of different roles over time can be presented and discussed. Hence the proposed method can uncover the optimal KPIs for a given set of roles, while having practical applications within elite-level performance analysis and recruitment departments.

Disclosure Statement

The authors report that there are no competing interests to declare.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

References

- Aalbers, B., & Van Haaren, J. (2019). Distinguishing between roles of footballplayers in play-by-play match event data. In U. Brefeld, J. Davis, J. V. Haaren, & A. Zimmermann (Eds.), *Machine learning and data mining for sports analytics MLSA 2018. Lecture notes in computer science* (Vol. 11330, pp. 31–41). Springer. https://link.springer.com/chapter/10.1007/978-3-030-17274-9_3
- Andrienko, G L, Andrienko, N V, Budziak, G, Dykes, J, Fuchs, G, von Landesberger, T, & Weber, H (2017). Visual analysis of pressure in football. *Data Mining and Knowledge Discovery*, 31, 1793-1839, <https://doi.org/10.1007/s10618-017-0513-2>
- Bialkowski, A, Lucey, P, Carr, P, Yue, Y, Sridharan, S & Matthews, I (2014). "Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data," 2014 IEEE International Conference on Data Mining, Shenzhen, China, 2014, pp. 725-730, <https://doi.org/10.1109/ICDM.2014.133>
- Becht, E, McInnes, L & Healy, J (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 37, 38–44 (2019). <https://doi.org/10.1038/nbt.4314>

- Carrilho, D, Santos Couceiro, M, Brito, J, Figueiredo, P, Lopes, R J & Araújo, D (2020). Using Optical Tracking System Data to Measure Team Synergic Behavior: Synchronization of Player-Ball-Goal Angles in a Football Match. *Sensors*. 20(17):4990. <https://doi.org/10.3390/s20174990>
- Charrad, M, Ghazzali, N, Boiteau, V & Niknafs, A (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6), 1-36. <http://www.jstatsoft.org/v61/i06>
- Cintia, P., Giannotti, F., Pappalardo, L., et al. (2015) The harsh rule of the goals: data-driven performance indicators for football teams. In: *Proceedings of the 2015 IEEE international conference on data science and advanced analytics, DSAA, Paris, France, 19--21 October 2015*.
- Closing down: How defensive pressure impacts shots. (2022). Available at: <https://statsbomb.com/2018/09/closing-down-how-defensive-pressure-impacts-shots/> (Accessed: 21 April 2022).
- Collet, C. (2013). The possession game? A comparative analysis of ball retention and team success in European and international football, 2007–2010. *Journal of Sports Sciences*, 31(2), 123–136. <https://doi.org/10.1080/02640414.2012.727455>
- Diaz-Papkovich, A, Anderson-Trocmé, L & Gravel, S (2021). A review of UMAP in population genetics. *J Hum Genet* 66, 85–91. <https://doi.org/10.1038/s10038-020-00851-4>
- Ensum, J, Williams, A, & Grant, A (2000). Analysis of attacking set plays in euro 2000. *Masters, Universidade do Porto*, 4, 36-39.
- García-Aliaga, A., Marquina, M., Coterón, J., Rodríguez-González, A., & Luengo-Sánchez, S. (2021). In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International Journal of Sports Science & Coaching*, 16(1), 148-157. <https://doi.org/10.1177/1747954120959762>
- García-Aliaga, A., Marquina Nieto, M., Coterón, J., Rodríguez-González, A., Gil Ares, J., & Refoyo Román, I. (2023). A Longitudinal Study on the Evolution of the Four Main Football Leagues Using Artificial Intelligence: Analysis of the Differences in English Premier League Teams. *Research Quarterly for Exercise and Sport*, 94(2), 529–537. <https://doi.org/10.1080/02701367.2021.2019661>
- Głowania, S., Kozak, J., Juszczuk, P. (2023) Dimensionality reduction for real sports data from the German Bundesliga and English Premier League. *Procedia Computer Science*, 225 (4334-4343). <https://doi.org/10.1016/j.procs.2023.10.430>
- Herold, M, Kempe, M, Bauer, P, & Meyer, T (2021). Attacking key performance indicators in soccer: Current practice and perceptions from the elite to youth academy level. *Sports Science and Medicine*, 158-169, <https://doi.org/10.52082/jssm.2021.158>
- Hewitt, A, Greenham, G, & Norton, K (2016). Game style in soccer: what is it and can we quantify it? *International Journal of Performance Analysis in Sport*, 16 (1), 355-372. Available at: <https://doi.org/10.1080/24748668.2016.11868892>
- Imburgio, M & Goldberg, S (2020). Introducing DAVIES: A Framework for Identifying Talent Across the Globe. Available at: *Introducing DAVIES: A framework for Identifying Talent Across the Globe — American Soccer Analysis* (Accessed: 19 June 2023)
- Introducing Possession-Adjusted Stats (2014). Available at: <https://statsbomb.com/articles/soccer/introducing-possession-adjusted-player-stats/>. (Accessed: 02 May 2024)
- Kawasaki, T., Sakaue, K., Matsubara, R., & Ishizaki, S. (2019). Football pass network based on the measurement of player position by using network theory and clustering. *International Journal of Performance Analysis in Sport*, 19(3), 381–392. <https://doi.org/10.1080/24748668.2019.1611292>

- Kim, H, Kim, B, Chung, D, Yoon, J, & Ko, SK (2022). SoccerCPD: Formation and Role Change-Point Detection in Soccer Matches Using Spatiotemporal Tracking Data. arXiv preprint arXiv:2206.10926, arxiv.org, <https://arxiv.org/abs/2206.10926>
- Lago-Penas, C, & Dellal, A (2010). Ball possession strategies in elite soccer according to the evolution of the match-score. *Journal of human kinetics*, 25, 93-100, <https://doi.org/10.2478/v10078-010-0036-z>
- Li, Y, Zong, S, Shen, Y, Pu, Z, Gómez, M Á, & Cui, Y. (2022). Characterizing player's playing styles based on player vectors for each playing position in the Chinese Football Super League. *Journal of Sports Sciences*, 40(14), 1629–1640. <https://doi.org/10.1080/02640414.2022.2096771>
- Lopes, A M, Tenreiro Machado, J A. (2021). Uniform Manifold Approximation and Projection Analysis of Soccer Players. *Entropy* 23, no. 7: 793. <https://doi.org/10.3390/e23070793>
- Lopez-Valenciano, A, Garcia-Gómez, J. A, López-Del Campo, R., Resta, R., Moreno-Perez, V, Blanco-Pita, H, ... Del Coso, J. (2021). Association between offensive and defensive playing style variables and ranking position in a national football league. *Journal of Sports Sciences*, 40(1), 50–58. <https://doi.org/10.1080/02640414.2021.1976488>
- Low, B, Coutinho, D, Gonçalves, B. et al. A Systematic Review of Collective Tactical Behaviours in Football Using Positional Data. (2020) *Sports Med* 50, 343–385. <https://doi.org/10.1007/s40279-019-01194-7>
- Lundberg, S & Lee, S (2017). “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems* 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., pp. 4765–4774, https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43df_d28b67767-Paper.pdf
- Maaten Lvd & Hinton G (2008). Visualizing Data using t-SNE. *J Mach Learn Res*, 9 (86), pp. 2579–2605. <http://jmlr.org/papers/v9/vandermaten08a.html>
- Mazurek, J. (2018). Which football player bears most resemblance to Messi? A statistical analysis. arXiv preprint. <https://arxiv.org/abs/1802.00967>
- McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv 2018. <http://arxiv.org/abs/1802.03426>
- Meerhoff, L A, Goes, F R, De Leeuw, A & Knobbe, A (2019). Exploring Successful Team Tactics in Soccer Tracking Data. In: Cellier, P., Driessens, K. (eds) *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science*, vol 1168. Springer, Cham. https://doi.org/10.1007/978-3-030-43887-6_18
- Merckx, S, Robberechts, P, Euvrard, Y, & Davis, J (2021). Measuring the effectiveness of pressing in soccer. *Proceedings of the Workshop on Machine Learning and Data Mining for Sports Analytics, Virtual*. Vol. 13.
- Narizuka, T, Yamazaki, Y. Clustering algorithm for formations in football games. *Sci Rep* 9, 13172 (2019). <https://doi.org/10.1038/s41598-019-48623-1>
- Oliveira, J. G. (2004). Specific knowledge in football: contributions to the definition of a dynamic array of the teaching -learning training game. Masters, Universidade do Porto.
- Peters, A J, Parmar, N, Davies, M, Reeves, M, Sormaz, M, & James, N. (2024). Expected Pass Turnovers (xPT) - a model to analyse turnovers from passing events in football. *Journal of Sports Sciences*, 42(11), 1002–1010. <https://doi.org/10.1080/02640414.2024.2379697>
- Peters, A, Parmar, N, Davies, M, & James, N. (2025). A rule-based approach to classify counterpressing – analysis of its risks and relationship with rest defence. *International*

- Journal of Performance Analysis in Sport, 1–17.
<https://doi.org/10.1080/24748668.2025.2473799>
- Ramos, S, Duarte, J, Duarte & Vale, F (2015). A data-mining based methodology to support MV electricity customers' characterization, *Energy Build.* 91 16–25.
- Statsbomb 360 - see how we're changing the game. (2022). Available at: <https://statsbomb.com/360-data/> (Accessed: 21 April 2022).
- Trainor, C (2014). Defensive Metrics: Measuring the Intensity of a High Press, <http://statsbomb.com/2014/07/defensive-metrics-measuring-the-intensity-of-a-high-press/>. (Accessed: 20 February 2022).