StatsBomb
Conference 2023

Clustering women's football players:

Identifying functional patterns for performance optimisation

Maia Trower, Niamh Graham, Natasha Cottrell, Yasmin Hengster

1. Introduction

Men's and women's football are different on both strategic and tactical levels - from player physiology and injury to league structure - to the point that models can easily distinguish between mens' and womens' games from match variables [1]. Men's football analytics is expanding and hugely profitable, and models and insights from men's football are (potentially inappropriately) applied to the women's game. Using the analysis carried out by Soccerment to cluster male players based on their functions as a framework [2], the aim of this paper is to specifically investigate player functions that can be identified in the women's game. We use Statsbomb event data from the 2018/19 to 2022/23 Women's Super League (WSL) to search for correlations between on-ball events and cluster the data based on the function of a player.

We investigate how classifications change from season to season to identify "hybrid players" who change function on the pitch over time, and "constant players", who remain consistent. These hybrid and constant players are analysed and we spotlight player profiles to highlight these identities. Then, we proceed with an analysis of how top WSL teams change their team composition over 4 seasons from 2019/20 to 2022/23, discussing how these choices are reflected in a teams' path to success. We conclude with an investigation of the functional composition of England's Lionesses in recent international competitions to highlight the application of this research beyond domestic football.

1.1 Significance of work

The traditional practice of labelling players based solely on their positions on the pitch can prove limiting, not only in describing a player's capabilities but also in shaping the strategies and formations employed by managers and coaches [3]. By characterising players according to their functional roles on the field, managers may become more creative in their tactics while identifying specific functional deficiencies within their squads. This shift away from traditional formations such as 4-4-2 or 4-3-3 opens up exciting possibilities for the development of innovative playing systems, allowing teams to experiment with new formations that best leverage the diverse functions of their players. Additionally, we envision that this analysis will offer substantial assistance in scouting, as it enables the straightforward replacement of players with analogous roles and facilitates the identification and addressing of gaps within teams based on these newly defined characteristics.

1.2 Application to the women's game

In the women's game, there is still a perception that footballers stay more rigidly in the confines of their specific positions and typically do not explore outside of this. We would like to challenge this perception by investigating whether the functions identified by the Soccerment study are also applicable and whether we can identify novel player functions specific to women's football. For example, we can observe how England captain and Arsenal centre back Leah Williamson plays a unique function with her progressive carries and through balls that often break not just one but two lines of pressure. This style of play is uncommon across both men's and women's football, and this research makes a first step towards identifying other such unique players and functions. Additionally, we anticipate that these results will have tactical as well as scouting applications. This is of significant interest currently as transfer fees are on the rise [4], and the use of data analysis continues to grow [5].

1.3 Data availability in women's football

Women's sports across the board faces a significant lack of data, from basketball to hockey to football [6]. Despite the growing popularity of women's football, there is still a data gap compared with the men's game. Since the rise of data analytics in men's football in the late 90s, data has played a crucial role in the development of men's football [7], serving as a foundation for in-depth analysis and modelling. The comparatively insufficient data available in women's football unfortunately imposes some restrictions on the ability of researchers to conduct comprehensive analysis and modelling to the same degree. In particular, clustering analysis requires a statistically significant number of data points in order to accurately group the data into informative clusters, and for this work we have access to only 5 seasons of WSL data. (Note that WSL seasons typically consist of fewer matches than, for example, a men's Premier League season.) Additionally, while there is 360 data freely available for the women's 2022 Euros and 2023 World Cup, this resource remains absent for WSL matches, which poses limitations on the features we are able to generate. This further restricts analytical capabilities.

2. Background

In 2022, Soccerment published a new framework aimed at making smarter scouting decisions by categorising men's football players as functions on the pitch rather than describing them using traditional positions [2]. The goal of this research was to highlight the skills, playing style, and specific strengths of individual players, which could be valuable when considering the role of a new signing or to improve the range of ability of a squad. The aim of this project is to apply similar methods to women's football players. In what follows we give an outline of the Soccerment paper which is the foundation of our project.

The authors of the Soccerment research used data from Opta's season data feed combined with detailed event data. They selected variables to capture all different types of player contributions, and normalised data to account for variability in total playing time among players. Defensive stats and basic passing metrics were normalised by the total number of on-ball events, known as touches, to account for different teams' opportunity to perform different kinds of actions. Other metrics were normalised per 90 minutes of play time. Advanced metrics such as Expected Goals and Expected Assists were not considered, since these focus mainly on performance, while the goal of the research was to distinguish play style. They looked at almost 2,000 players, each with over 1800 minutes of play time, taking an average of these stats for these players over the last four seasons prior to normalising. The authors chose to exclude goalkeepers from the analysis.

The data were then scaled before applying dimensionality reduction via Uniform Manifold Approximation and Projection. A Gaussian Mixture model was then fitted to the data, to obtain probabilities of each player belonging to each cluster. The use of probabilities allows the flexibility of not restricting players to one specific label, so hybrid players can be identified. It is crucial to note here that Soccerment defined hybrid players as those who had a significant probability of belonging to multiple clusters, as opposed to our definition of a hybrid player who changes clusters over time. Moving forward, we only discuss hybrid players in the latter sense, and we do not investigate players with high probability of membership to distinct clusters.

The outcome of their clustering was five macro-clusters, within which sub-clusters were identified using additional Gaussian mixture models. This led to a total of 13 player functions. Each was given a descriptive name: Ball Stopper, Build-Up Initiator, First Line Breaker, Wide Controller, Wide Creator, Ball Stealer, Build-Up Director, Box-To-Box Raider, Chance Creator, One-To-One Explorer, Mobile Finisher, All-Round Finisher and Target Man. Hybrid players were identified, defined as players who had a maximum classification probability for any cluster of less than two-thirds.

Having identified the functions and categorised each player, the authors analysed the make-up of various teams with regards to these functions. They investigated team composition using players' probabilities of membership to various clusters and then looked at team performance. This led to the finding that a modern top-level team has a distinctly high proportion of clusters characterised by highly technical attributes, such as

Buildup Initiator, Wide Creator, Buildup Director, Chance Creator, One-to-one Explorer and Mobile Finisher. They found that more traditional functions such as Ball Stopper, Wide Controller and Target Man are strongly under-represented in modern top teams. These insights can be used to analyse and improve team performance.

3. Data Description

For this work, we use event data provided by StatsBomb, which records over 3,400 events per match. Our focus was on women's football and the data available was WSL matches for the last 5 seasons (2018-19 season to 2022-23 season). This includes 590 matches across all seasons, and features 561 players. In order to ensure convergence of our model, we decided to consider only players who played for more than 1000 minutes (see Figure 1). This retained over 60% of available players.

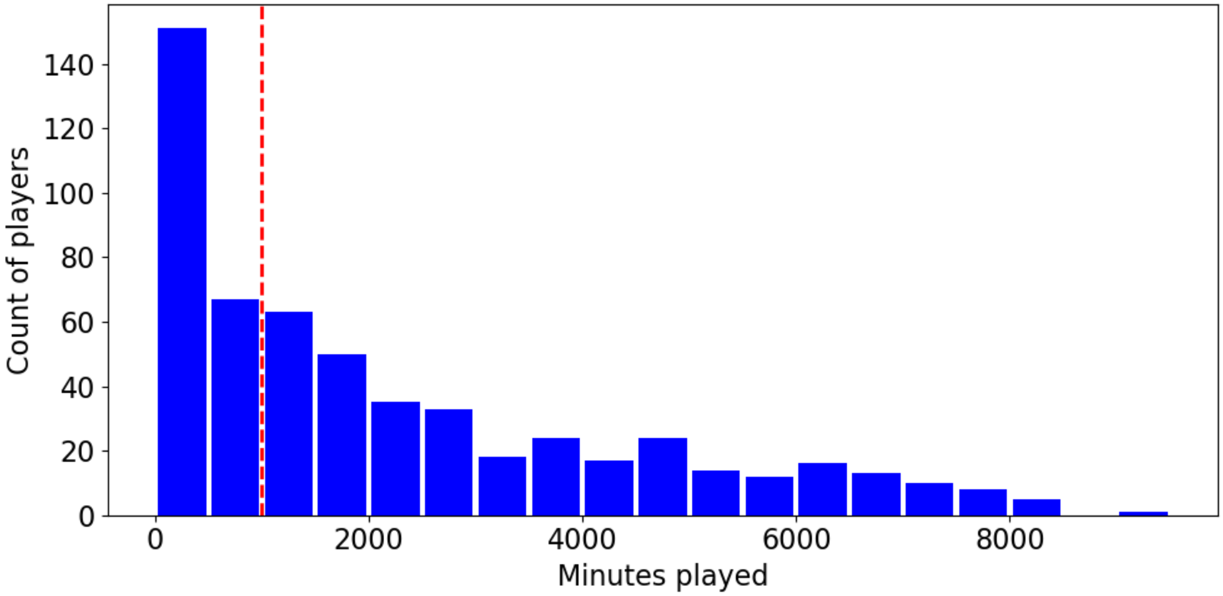

Figure 1: Histogram of minutes played per player. The red line indicates 1000 minutes.

3.1 Event Data

The key component of this analysis is to identify which events players take part in during the match, and based on this, a clustering algorithm can be used to define player functions. We generated a number of features, presented in Table 1. These include cumulative tallies of events, event counts based on the location, the body part used for a shot, as well as the average distance for selected events. The tallied events are normalised to ‘per 90 minutes played’, and the other features are normalised by the global average across all players.

Features			
Number of duels	Number of dribbles	Number of carries	Number of 50-50s
Number of interceptions	Number of clearances	Number of blocks	Number of ball recoveries
Number of fouls committed	Number of shots	Number of dribbles (widest third)	Number of dribbles (attacking third)
Number of passes (defensive third)	Number of passes (midfield third)	Number of passes (attacking third)	
Number of shots (outside box)	Number of shots (inside box)	Number of shots (headers)	Number of shots (non-header)
Number of shots (other)	Number of shots (right foot)	Number of shots (left foot)	
Pass distance	Carry distance	Shot distance	

Table 1: Features generated from WSL event data.

In general, some events are likely to occur significantly more often than others during a football match, and some of the features generated here have differing scales. For example, a Premier League team may attempt 900 passes per match [8] but may only take 13 shots [9]. This means that in order to cluster players using these metrics, we must preprocess the data to ensure all features have the same scale. In this work we use Python’s standard scaler, which ensures that each feature has zero mean and a standard deviation of one.

Prior to scaling the data, we complete the following preprocessing. First, we extract the features from the event data for each WSL season. We store each season individually alongside a dataset containing data for all seasons together, referred to as the averaged WSL dataset. We apply the normalisation outlined above (per minutes played, normalising by global average) to the six generated datasets. As we will see below, we used the concatenation of these six datasets, rescaled using the standard scaler, to train and fit our model.

4. Methodology

In this section we outline key methods and techniques used in our analysis. These include dimensionality reduction and data visualisation techniques, and the clustering model we employ.

4.1 Dimensionality Reduction

In order to reduce the dimension of our dataset before applying our clustering methods, we apply principal component analysis (PCA) to the features. This method combines the original features into principal components which retain the important information in the data while helping to remove noise in the data, which can reduce overfitting. In particular, PCA identifies the dimensions which contain the greatest amount of variation in the data, called principal components (PCs). These PCs are linear combinations of the original predictors, and are ordered by the amount of variation retained, so that the first principal component is the linear combination of the predictors containing the most variance in the data [10].

Whilst PCA is an effective tool for reducing the feature space, we note that the PCs themselves are not readily interpretable. As they are linear combinations of the original features, some work is required to extract meaning from the clusters using the PCs. Therefore, we cluster the transformed data but use the original set of features in order to analyse the functions and players (see Section 5).

Dimensionality reduction techniques can also be used for data visualisation; by projecting the data points into a 2 dimensional or 3 dimensional space, we can show the data easily whilst preserving distances between data points. A particularly effective tool for visualisation is t-distributed stochastic neighbour embedding (t-SNE) which, unlike PCA, preserves pairwise similarity between data points by minimising the divergence between the distributions of the two points [11]. This is useful for capturing nonlinear relationships, but is more computationally expensive than PCA. Additionally, t-SNE depends heavily on hyperparameter selection. A refined analysis would go beyond the scope of this project, and for this reason, we use t-SNE only as a visualisation tool for illustrative purposes.

4.2 Clustering

To identify player functions, we utilise a clustering algorithm to identify groups within the data based on similar playing styles. Since the number of clusters is not known a priori, we explore the performance of the algorithm over a varying number of clusters, using a combination of statistical methods and in-field knowledge to select the appropriate number of functions.

We choose a Bayesian Gaussian mixture model (BGMM) to cluster the data, which is an extension of the Gaussian mixture model (GMM). The latter represents the data as a combination of Gaussian distributions, with each distribution corresponding to a cluster. The parameters of the distributions and the mixing coefficients are estimated using an Expectation-Maximisation algorithm. The importance of each cluster is determined by the mixing coefficients, so it is possible to find the appropriate number of clusters by comparing cluster weights (coefficients) and introducing a cut-off to include only the most important clusters. The GMM then assigns cluster memberships for each player, allowing for the identification of player functions [12].

The BGMM model, unlike GMMs, can infer the number of clusters from the data using Bayesian inference principles. This is an advantage over standard GMMs, especially for problems where the cluster count may vary or be difficult to determine. Selecting the optimal number of clusters is then done by plotting the cluster weights to identify an inflection point on the curve. That is, we look for a point where there is a significant drop in cluster weights, which indicates that additional clusters are not substantially improving the model's ability to explain the data. We find a trade-off between increasing the number of clusters, which can better fit the data but may overfit noise, and simplifying the model, which may underfit the data.

5. Results

5.1 Model

We apply the methodology outlined in Section 4 to analyse the concatenated dataset described in Section 3. Initially, we perform dimensionality reduction using PCA. Figure 2 shows the percentage of variance retained by the projection onto principal components, with each bar showing the variance explained by the individual PCs. We also show in Figure 2 the cumulative variance, and we set a minimum threshold of 95%. This threshold is achieved by 16 PCs, which we regard as sufficient to describe the dataset effectively. We have therefore reduced the dimensionality of the feature space to 16 features, preserving 95% of the information in the dataset, whilst removing noise and highly correlated features.

Before we fit the clustering model we separate out the data into the original 6 datasets. We now have a dataset that contains averaged information from all 5 seasons, scaled and projected onto the first 16 PCs, that we will use to train the clustering model.

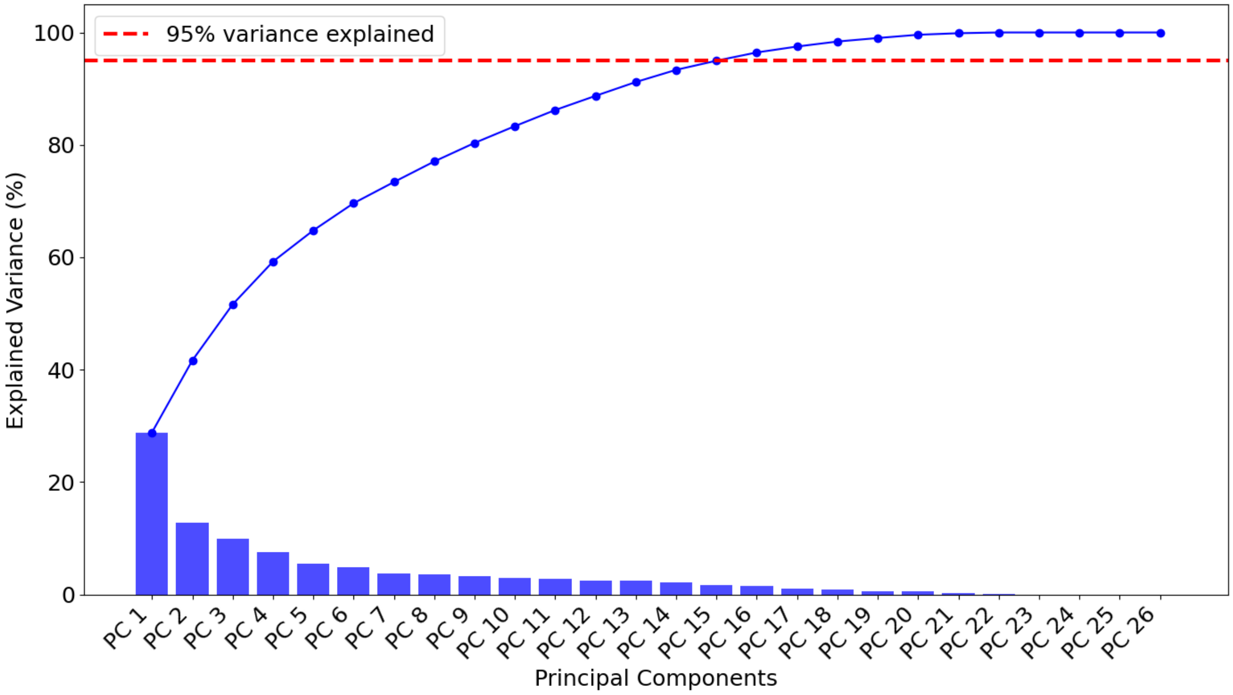

Figure 2: Amount of variance explained by each component using PCA.

We apply a Bayesian Gaussian Mixture Model for clustering of the averaged data. As described in Section 4, each cluster identified by the BGMM has a corresponding weight, or mixing coefficient, which we show in Figure 3. These weights signify the contributions of each cluster to the overall model. Notably, Figure 3 reveals an observable gap between clusters 11 and 12, suggesting the presence of 11 significant clusters. This compares with findings from the Soccerment paper [2], which identified 13 player functions in total. Another smaller gap appears between clusters 5 and 6, implying that clustering the data into 6 groups would also segment the players into similar types. However, this gap is small compared to the gap between clusters 11 and 12, so it is likely that these clusters will over-generalise the data and produce groups that are less informative. We suggest that the 6 general clusters may correspond to the traditional playing positions of strikers, attackers, midfielders, wingers, central defenders, and goalkeepers.

Noting that goalkeepers are excluded from all analysis in [2], the presence of 6 initial clusters in Figure 3 aligns with the finding of 5 initial clusters in [2]. While Soccerment used these 5 clusters to define subclusters within each initial cluster, resulting in 13 groups, we define 11 main clusters. That is, we do not subcluster our 6 initial groups.

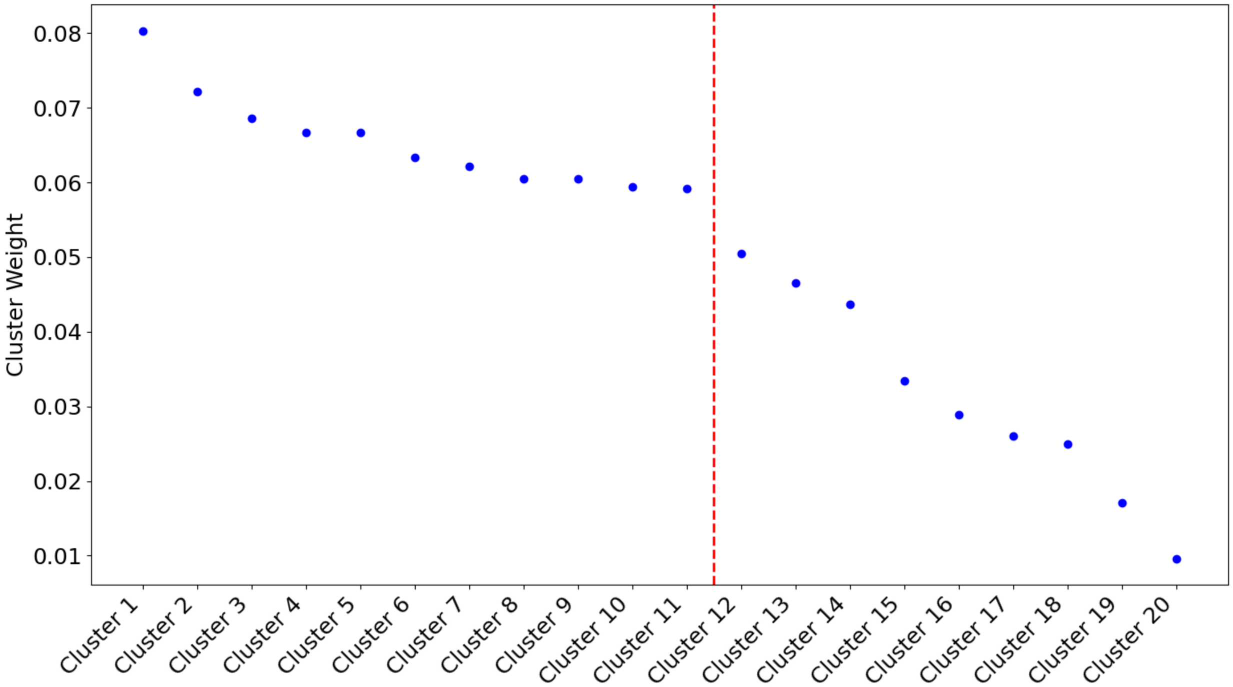

Figure 3: Weights of each Bayesian Gaussian Mixture Model cluster.

5.2 WSL Clustering

Fitting the BGMM to the averaged WSL data, we are now able to investigate the clusters by analysing both the players in each cluster, and the features that correlate strongly with cluster membership. The model produced 11 clusters, which we have named Finishers, Interception masters, Dynamic distributors, Shooting backs, Goalkeepers, All-round playmakers, Wildcards, Goal initiators, Versatile backs, Ball clearers, and Defensive shields. These names are based on analysis of the resulting cluster, and we include further details in the appendix. Note that for much of the proceeding work the seventh cluster, the Wildcards, has been excluded for two reasons. Firstly, although a small number of players are assigned to this cluster in the averaged WSL data, we found that no players were Wildcards when we considered the 5 seasons individually. Secondly, the players in this cluster did not appear to be strongly associated with each other and had no apparent common function. A deeper analysis of these players revealed at least one error in the raw data, and so we do not consider this cluster as a valid function and choose to remove the players from our results.

We can now visualise the groupings of players in the dataset. Figure 4 shows the data projected into two dimensions using t-SNE, colour-coded by their respective clusters. We observe that the clusters occupy largely distinct regions in the projected space, which indicates a separation along the two axes containing the most variation in the dataset. It is crucial here to point out that this visualisation is done using a different dimensionality reduction technique than was used to preprocess the data; we chose t-SNE largely for interpretability. Also, note that the distances between points and the grouping together of clusters is not indicative of geographical proximity in the original data as we have projected onto the first two t-SNE components, and the shape of the clusters seen here is not reflected in the original data. However, we can see from Figure 4 that many clusters separate well. In particular, we note that the cluster containing goalkeepers is the most separated from the other clusters, which suggests that players performing this function are very different across all features contained in components 1 and 2. This confirms the assumption that keepers will perform a very distinct function from all other player types.

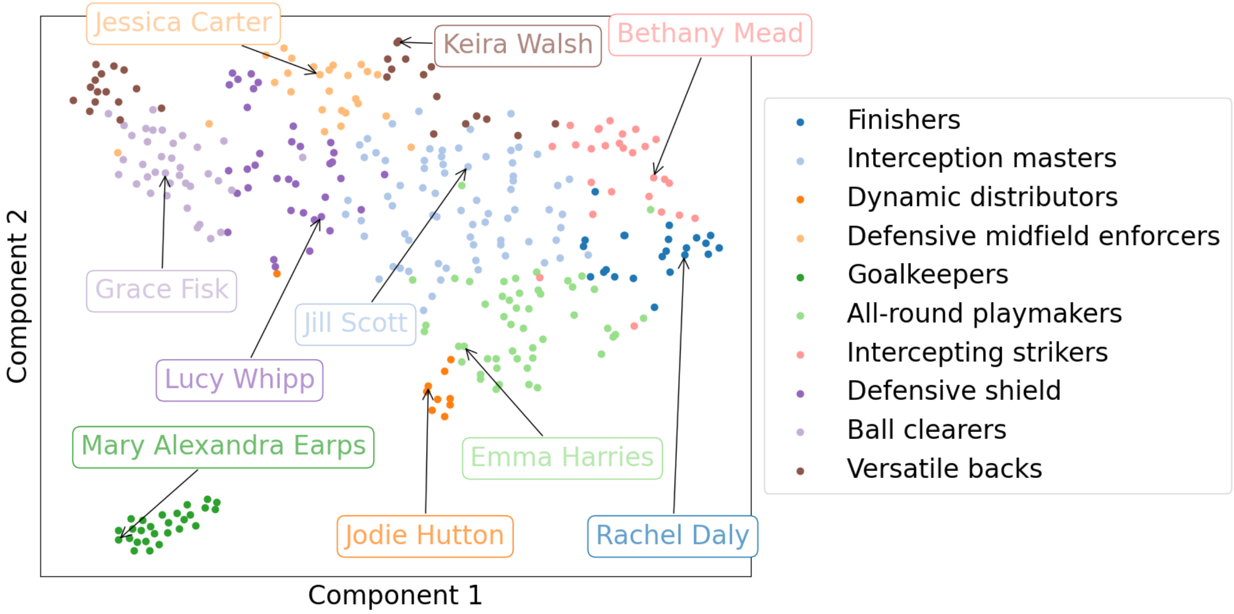

Figure 4: Data visualised in 2 dimensions using t-SNE (coloured by cluster).

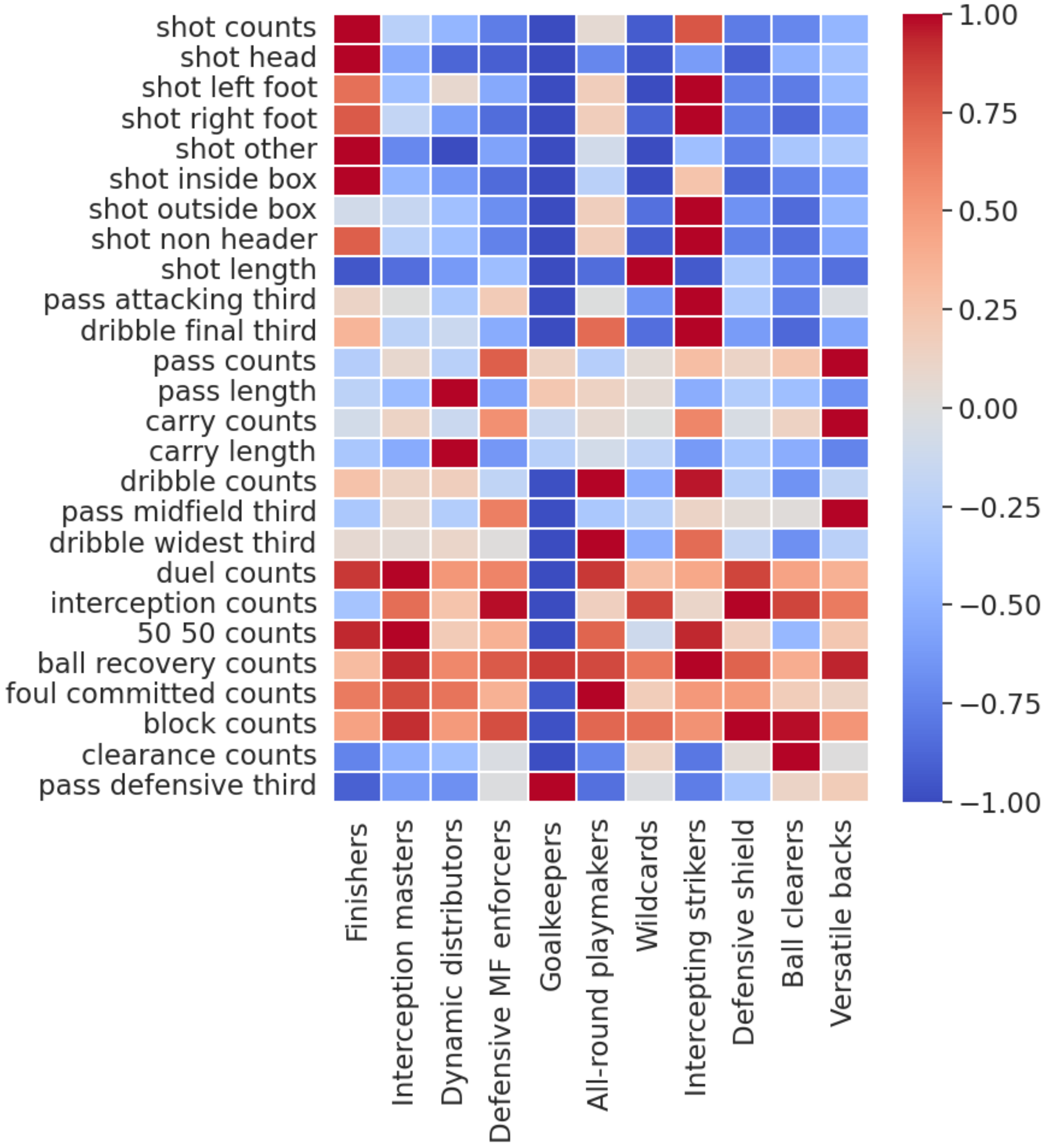

Figure 5: Average values of each statistic per cluster.

In Figure 5, the heatmap illustrates the average value of each variable within each cluster. Features are listed on the y-axis, while clusters are represented on the x-axis. Features with the largest average magnitude are indicated with darker shades; red shades represent a large positive value, with large average negative values shown in blue shades. The averages have been normalised and scaled, so negative values are indicative of below average statistics compared with other players in the dataset. For instance, this plot shows that players in the first cluster complete significantly more shots than players in other clusters. Players such as Rachel Daly, Alessia Russo and Samantha Kerr are within this group, which aligns well with the high number of shots taken. We refer to this cluster as the Finishers for this reason. Also, the fifth cluster, the Goalkeepers, have relatively low correlations with almost all features except passes in the defensive third and number of ball recoveries. This aligns with the function that goalkeepers are known to perform.

We also highlight here the eleventh cluster, which we identify as the Versatile backs. Figure 5 shows that, as we might expect, these players do not perform many shot events and in general are not completing passes in the final third of the pitch. The Versatile backs instead complete a high number of passes and carries, and they are also strongly associated with ball recovery. This cluster contains players such as Leah Williamson and Lucy Bronze, who are known for their skill and strength as defensive players and play predominantly in positions such as right back and centre back.

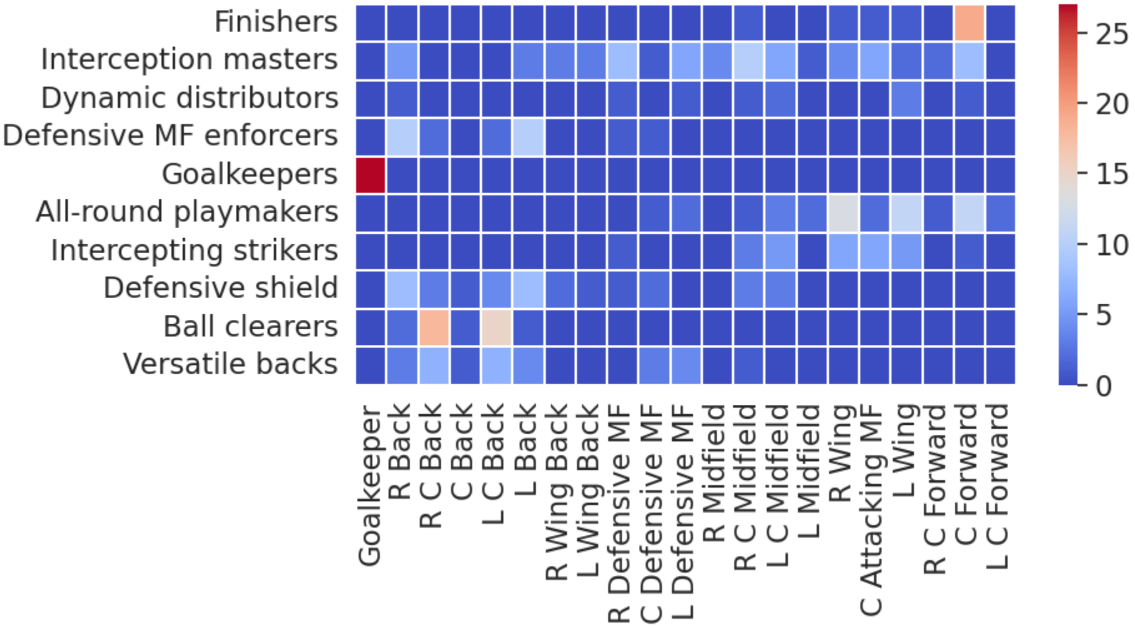

Figure 6: Breakdown of functional role by traditional position (excluding Wildcards).

We now consider the relationship between the newly identified clusters and traditional positions, which is illustrated in Figure 6. The figure clearly shows total overlap between the goalkeeper position and a single cluster. This motivated the naming of this cluster as simply “Goalkeepers”. In general, the goal of this clustering project is to highlight novel functions and so we would like the clusters to separate out positions rather than group them all together. However, since the function of goalkeepers is so distinct from other players, we suggest that this total correlation not only makes sense but is a desired outcome. Based on this finding, for future analyses it would be reasonable to exclude goalkeepers from the dataset from the outset.

Whilst we may not expect the clusters to align completely with a traditional position, an approximate agreement indicates that the model is generating useful clusters. For example, note that a Ball Clearer is very likely to be either a right centre back or a left centre back, which agrees with intuition. However, All-Round Playmakers are usually defined as wingers or centre forwards, which shows that these players can be versatile in their traditional position and that the function of playmaker can be performed in a number of roles on the pitch.

5.3 Hybrid Players

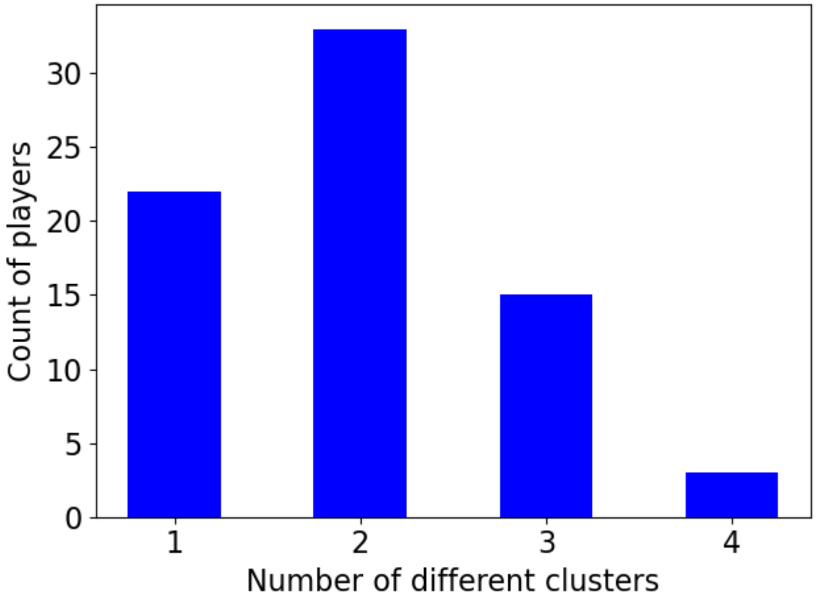

Figure 7: Number of distinct clusters assigned per player over 5 WSL seasons (2018/19-2022/23).

Next, we present the idea of a hybrid player, and compare these to constant players. Hybrid players are, in this context, players who change function on the pitch over time, and constant players instead perform a stable function over the 5 WSL seasons for which data is available. In this analysis we consider only players who played in all 5 seasons in order to meaningfully comment on the stability of their functions, and we use the 5 datasets of individual seasons to track players' clusters over several years (preprocessed as described in Sections 3 and 4).

Examining how football players adapt their function on the pitch over 5 seasons can shed light on their ability to adjust to evolving team tactics, highlighting versatility and adaptability. A hybrid playing style also signifies growth and development on the field. Moreover, changes in player roles may reflect shifts in team dynamics or a team transfer, and analysing this could uncover key insights to contextualise the performance of both the player and the team. Equally, it is important to understand footballers whose function or style remains relatively consistent over the years. These may be players who consistently excel in established roles, and such players often become the backbone of their teams. This can be essential for team cohesion, allowing a successful squad to be built around their dependable function.

Figure 7 shows that, of the 73 players playing in all 5 seasons, the majority of them perform two distinct functions. This indicates relative stability, and suggests that players largely perform similarly from one season to the next. It is relatively uncommon for a player to be grouped in 4 distinct groups over the 5 seasons, and we see that no player is clustered differently for every season. According to Figure 7, we classify as hybrid players those who receive 4 distinct cluster labels across the 5 seasons. Constant players, by contrast, are those who receive the same label for every season.

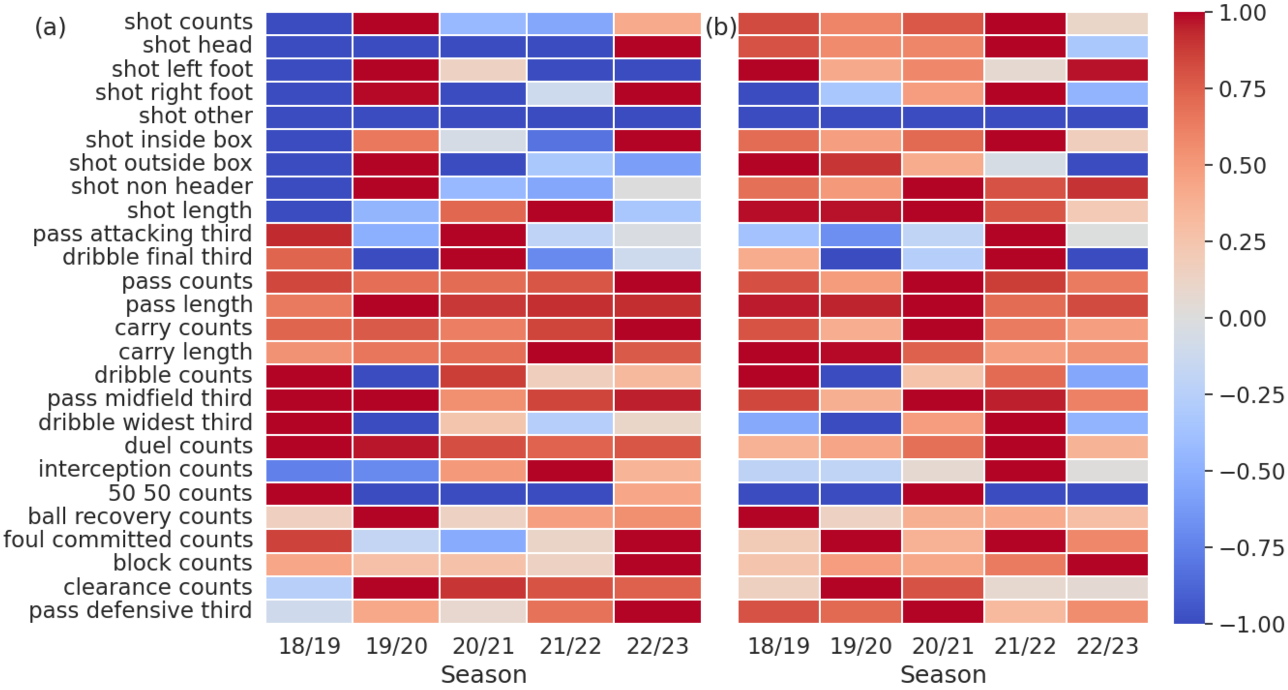

Figure 8: Event statistics for Jess Carter (left, a) and Magdalena Eriksson (right, b) from 2018/19-2022/23 WSL seasons.

To illustrate this, consider players Jess Carter and Magdalena Eriksson. Both Chelsea FC players, Carter is known for her versatility, and is capable of playing as a defender or as a midfielder. Former captain Eriksson, on the other hand, is known primarily for her exceptional skills as a defender. Eriksson is consistently assigned to the Versatile back cluster through every season. Characterised by a high number of passes, carries, ball recoveries, and passes in the midfield third, this group is mostly made up of defenders. Eriksson embodies this cluster with consistently higher values of these statistics than other players, which she maintains consistently over the seasons. Comparing her own performance season-on-season, as shown in the heatmap, we can see consistency (note: the dark blue at -1 on the scale indicates 0 e.g. no shots for that season), particularly in the aforementioned metrics that make up the key identity of this cluster.

In comparison, Jess Carter is a Defensive midfield enforcer. This cluster is characterised by long range shots and attacks from defensive positions. However, season-on-season, she also fits into the Versatile back (2019), Interception master (2020) and Ball clearer (2022) clusters. In 2019, we can see Carter has a lower shot count than is typical for her, whilst still maintaining a relatively high carry count, which is indicative of the Versatile back cluster. Similarly, in 2020 Carter has a lower than average number of shots but higher than average number of duels, which is characteristic of an Interception master. Finally, in 2022, Carter has very high clearance and block counts, which are again key features of the Ball Clearers. This range shows that Carter has the ability to undertake a number of functions on the pitch. This could be dependent on the make-up of the team on a game-by-game or season-by-season basis, or perhaps on the strategy that the team chooses to execute.

A combination of consistent and hybrid players can allow a team to adapt tactics and cope with different opposition teams more easily - utilising the consistent players as anchors and optimising the function of the hybrid players to give the team the greatest chance of success. The versatility of hybrid players removes dependency on specific players and mitigates the impact of players who may be absent, particularly key whilst many players are absent with ACL injuries (a key issue in women's football currently [13]). This also allows for more adaptability to opposition, and in-game adjustments, allowing players to change playing style based on the outcomes of the match as it evolves. In contrast, consistent players are key to providing stability, reliability, and expertise to the team. From these players, the expectation for play is clear.

5.4 Domestic Team Composition

Moving away from individual players, we now investigate the make-up of top WSL teams in terms of the clusters we have defined above. Analysing the composition of these teams provides valuable insights into team performance and tactics. It contains information about the strategic decisions made by clubs, reflecting trends in performance analysis and scouting. Tracking these changes over time aids in understanding how a team manages to maintain and to succeed, or even why a team might fail to perform. We can also pair this with an individual analysis of the players to highlight versatile players capable of fulfilling multiple roles or specialised and consistent players, and how a team may choose to adapt when a player leaves or joins the squad.

We utilise WSL data over the most recent 4 seasons and compare two teams that achieved significant success during the 2022/23 season, namely Chelsea F.C. Women and Manchester United W.F.C. . Note that we use only the most recent 4 seasons because, prior to this, Manchester United did not play in the WSL and we therefore do not have event data from 2018/19. Chelsea have enjoyed a sustained period of success, with a well-established presence in the women's game and have placed top of the WSL table from the 2019/20 season to the 2022/23 season. Having been managed by Emma Hayes since 2012, the team has benefited from an extended period of consistent coaching, resulting in a stable team dynamic. In contrast, Manchester United, whose women's team was only reintroduced in 2018, represents the rapid ascent of a relatively new contender. Placing 4th on the WSL table from seasons 2019/20 to 2022/23, and performing under two different head coaches in this time, the team had to adapt quickly and lacked the stability of their opponents. Despite this, Manchester United achieved a second place position in the WSL 2022/23. We compare the functional composition of the teams of the past 4 years in Figure 9.

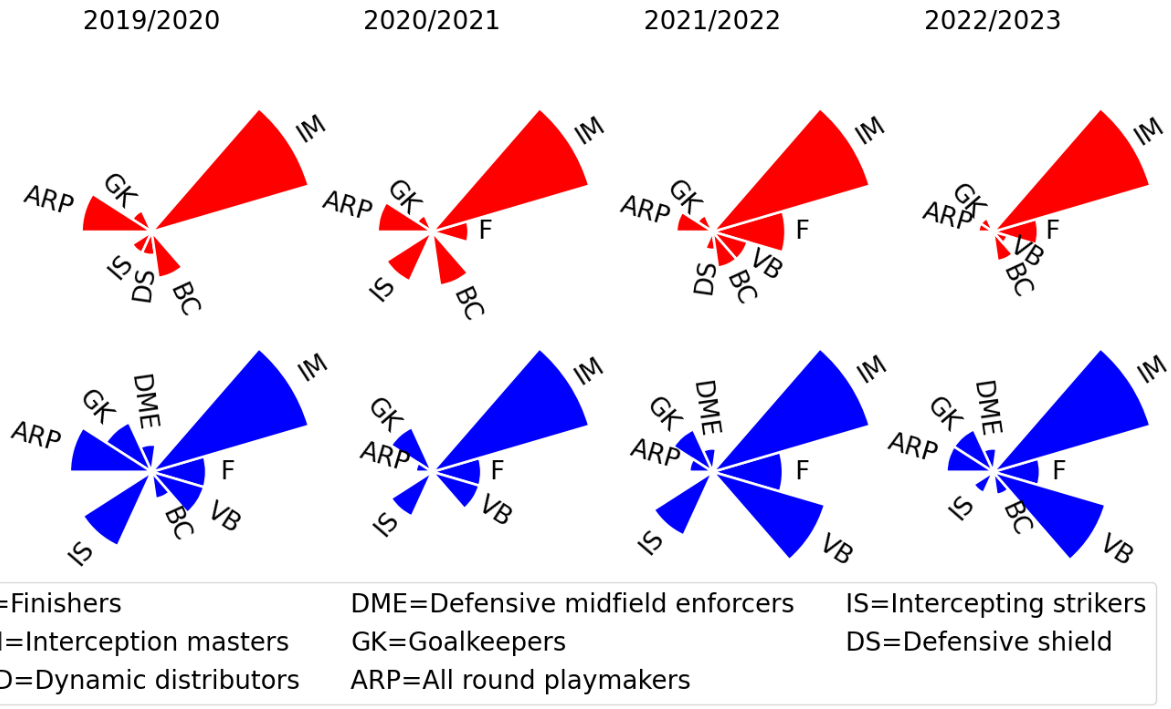

Figure 9: Team composition by functional role of Manchester United (top) and Chelsea (bottom).

It is clear that for both sides, Interception masters form a very important part of each team's strategies, with a large proportion of players falling into that group.
Looking at Manchester United's team composition over the 4 years it is evident that the team has trialled different strategies. Beginning with a more defensive format in 2019/20,

the team has slowly introduced more Finishers, ending with a more attacking squad in 2022/23 with Finisher Alessia Russo scoring 10 goals making her the teams top goal scorer that year. Manchester United also reinforced their Interception masters over the 4 years, growing from 7 in 2019/20 to 11 in 2022/23. Manchester United's team only consists of 6 of the functional roles over the 4 years (7 in 2021/22 season) however each year, Chelsea's team consists of 6-8 of these roles.

Chelsea benefits from a wider range of functions on the pitch and favours a more versatile team consisting of dual-function players such as Intercepting strikers, Versatile backs and Defensive midfield enforcers. Over these years, the team has been consistent in their range of functions, with strong defence and a strong Finisher in Sam Kerr, who was the league's top goal scorer in the 2020/21 and 2021/22 seasons.

This comparison highlights the different approaches taken, where Chelsea embodies consistency and a long-term commitment, and Manchester United exemplifies the determined pursuit of rapid improvement and evolution. It also highlights the success in using a more diverse range of functions on the pitch.

5.5 International Team Composition

5.5.1 WSL vs International Composition

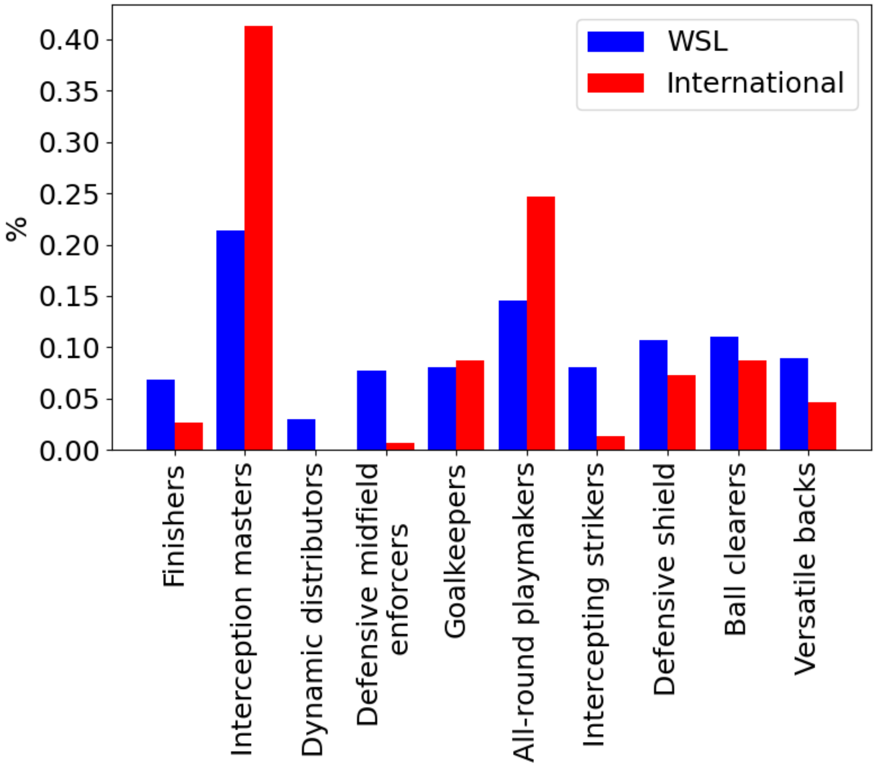

Figure 10: Distribution of functional roles across all seasons of the WSL and three international tournaments (Euros 2022, World Cup 2019 and World Cup 2023).

Using the clusters derived from the WSL data, we analyse the functional roles of players in international tournaments. For this, we have looked at the 2022 Euro's and the 2019 and 2023 World Cup data. Figure 10 shows a differing composition of teams by function when comparing domestic and international data. The international data show a significantly higher proportion of Interception masters and All-round playmakers and fewer Finishers, Dynamic distributors, Defensive midfield enforcers and Intercepting strikers than the domestic data.

Approach to play between international and WSL matches varies. At the simplest level, the standard of competition at international tournaments is higher. Additionally, teams in the WSL have the advantage of playing together regularly throughout the season, allowing for better team cohesion, whereas national teams often have limited time to prepare for tournaments. In the WSL, playing styles can vary significantly between clubs based on the preferences of their respective managers, whereas for national teams, the tactical approach is typically more standardised. The international tournaments also feature knockout games so every match is crucial, which could lead to more cautious and risk-averse play. In contrast, the WSL follows a league format, which can promote more open, attacking and riskier tactics.

Bearing this in mind, we expect some key differences in team composition between international and domestic tournaments. For the international tournaments, we observe fewer Finishers, which we suggest is due to more cautious playing styles. Similarly, the lower number of Dynamic distributors could be due to the high pressure of international matches, meaning pass and carry length is likely to be shorter as this reduces chances of mistakes. This same logic can be applied to the Defensive midfield enforcers. On the other hand, we find that international tournaments have a greater number of Interception masters. We suggest that this is due to the greater emphasis on possession, increasing the number of blocks, duels and 50/50s as players battle for the ball or defend their goal. Interception masters also have high ball recovery counts, which aligns with this playing style. Additionally, teams often focus on strong defensive performances in international tournaments during knock-out stages, which can lead to a higher number of blocks as defenders and midfielders work hard to obstruct shots and prevent goals. Lastly, we also find an increase in All-round playmakers for the international tournaments, typified by a higher than average number of dribbles, particularly in the widest third. This may be due to tighter marking which can force attacking players to use dribbles to progress up the field - particularly in wide areas where they have more space to operate.

5.5.2 Lionesses Euros vs World Cup Composition

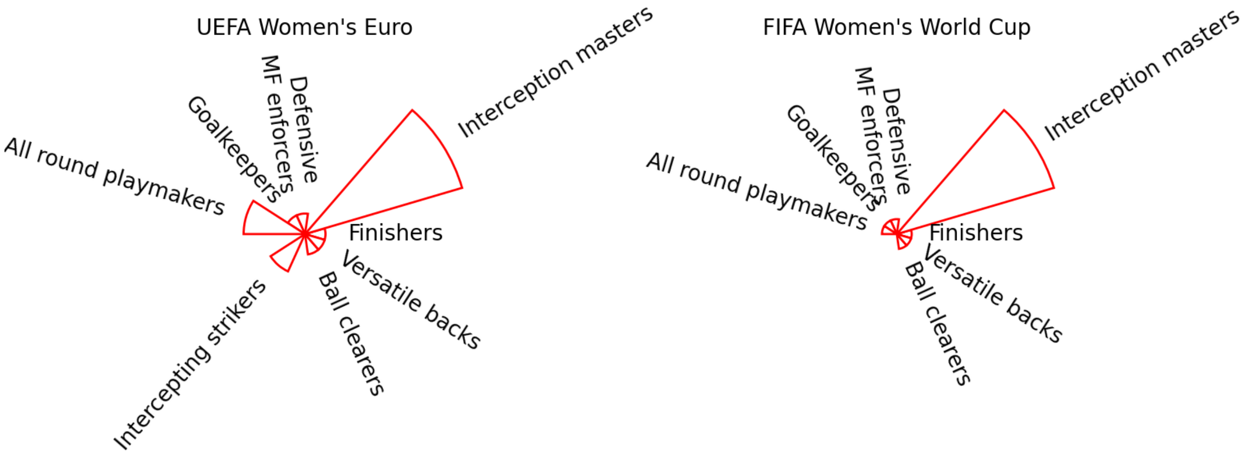

Figure 11: England Lionesses team composition.

We now compare the difference in the Lioness team composition between the 2022 Euros and the 2023 World Cup. The key trend here is a reduction in All-round playmakers and Intercepting strikers in the World Cup compared to the Euros. These clusters were represented by Jill Scott and Beth Mead respectively in the Euros, but neither Scott nor Mead played in the 2023 World Cup (due to Scott's retirement and Mead's ACL injury), which left the Lionesses without these key players.

Overall, we see a strong number of Interception masters for both Euros and World Cup line-ups. In the Lionesses' line-up, this cluster is predominantly made up of players who are not typically in this cluster in domestic play, but are able to adapt their play style for these games and transition between clusters. For example, Lucy Bronze and Chloe Kelly are both Interception masters when playing internationally, but in the WSL, they are identified as a Versatile back and an Intercepting striker respectively.

As the international tournaments feature only a small number of matches, there is limited data available for this analysis. Therefore, due to limitations in data availability, a comprehensive comparison of domestic and international team compositions is outside the scope of this paper.

6. Conclusions

The aim of this work was to explore functional roles in women's football, highlighting a move away from traditional positions that has been exploited in men's football in [2]. Our analysis focused on 5 seasons of domestic WSL data, as well as data from the Lionesses squad at three international tournaments spanning 4 years.

We have shown that utilising standard data processing and dimensionality techniques allows for the extraction of interpretable and explainable features from event level data, which can then be used to group players into functional types. We identify each cluster with a descriptive name based on the features that correlate strongly with each cluster in combination with analysing the WSL players assigned to each cluster. We were then able to define hybrid and consistent players, that is, players who change clusters at least 4 times across seasons and players who change clusters not more than once, respectively. For each of these player types, we provide a case study of such a player.

Finally, we conclude by showing that functional types can be used to analyse team composition both domestically and internationally. We highlight differing trajectories for top WSL teams based on the make-up of their squads, the composition of Chelsea, the team winning the last 4 seasons of the WSL, is consistent with only small changes across seasons. Manchester United, a team that was only reintroduced in 2018, shows a trend from more defensive tactics towards a team with a higher focus on attacking players. We also extended this analysis to the team composition of England's Lionesses. Changing from a domestic league to an international tournament, we find that the team is more defensive in comparison to the two analysed teams in the WSL. We concluded by noting how the Euros and World Cup squads have evolved over the past 2 years. Interestingly, we were able to find a connection between the change of the team composition and players not attending the World Cup, such as Beth Mead or Jill Scott.

Ultimately, we believe that adopting approaches similar to the one we've employed here could enhance and streamline scouting in women's football. Our clusters not only reveal the team's current configuration but also pinpoint areas where certain skills might be deficient or overly abundant. This gives the ability to efficiently identify players from other teams who can address these skill imbalances, which goes beyond the basic task of substituting players in a like-for-like manner by position.

This project underscores the need to approach women's football with the same analytical rigour and curiosity as the men's game. We emphasise the relative scarcity of data and analysis in this space, and note that our work necessarily focused exclusively on English football as a result of this. As women's football continues to grow and evolve, this research contributes to a better understanding of the game, how it compares with and differs from the men's game, and highlights the potential for further development of data analytics in this area.

References

[1] Pappalardo, L., Rossi, A., Natilli, M., & Cintia, P. (2021). Explaining the difference between men's and women's football. *PLOS ONE **16**:8.*
[2] Gagliardi, A. (2022). The Clustering Project: A New Framework for Smarter Scouting. *Soccerment Research.*
[3] Imburgio, M. (2020). Defining Roles: How Every Player Contributes to Goals — American Soccer Analysis. *American Soccer Analysis*.
https://www.americansocceranalysis.com/home/2020/8/3/defining-roles-how-every-player-contributes-to-goals
[4] Marsh, M. and Toloui, A. (2023). How the January window changed WSL transfers forever: Alessia Russo bid and Bethany England's British record move help spark new era. *Sky Sports*.
https://www.skysports.com/football/news/11095/12800148/how-the-january-window-changed-wsl-transfers-forever-alessia-russo-bid-and-bethany-englands-british-record-move-help-spark-new-era
[5] Clark, M. (2023). Future of Football: Alex Greenwood shows benefits data analysis can have on women's game - when will technology catch up? *Sky Sports*.
https://www.skysports.com/football/news/11095/12923617/future-of-football-alex-greenwood-shows-benefits-data-analysis-can-have-on-womens-game-now-time-technology-caught-up
[6] Kleen, B. (2022). Without numbers, you can't tell the story: Understanding the gender stats gap in sports. *Global Sport Matters*.
https://globalsportmatters.com/business/2022/07/13/lack-womens-sports-data-hurts-product/

[7] XFB Analytics (2022).The history of football analytics - Part 2*. XFB Analytics.*
http://www.xfbanalytics.hu/blog/blog-post/32

[8] Manuel, J. (2022). What you didn't know about Premier League passing. *The Analyst*.
https://theanalyst.com/eu/2022/10/what-you-didnt-know-about-premier-league-passing/#:~:text=One%20thing%20we've%20seen,cross%20delivering%2C%20you%20name%20it

[9] Pugsley, D. (2013). A Deeper Look At Shots on Target. *Bitter and Blue*.
https://bitterandblue.sbnation.com/2013/1/17/3880454/a-look-at-shots-on-target-epl#:~:text=After%20all%2C%20in%20any%20given,average%20will%20be%20on%20target

[10] Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A. **374**: 20150202*

[11] van der Maaten, L. and Hinton, G.,(2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, **9**: 2579-2605*

[12] Bishop, C. M. (2006). Pattern recognition and machine learning. **4**: 4. New York: *Springer*.

[13] Sky Sports(2023). Inside the WSL: Why are ACL injuries so common in women's football? *Sky News*
https://www.skysports.com/football/news/35730/12748748/inside-the-wsl-why-are-acl-injuries-so-common-in-womens-football

Appendix

Cluster Name	Positions	Key Players	Description
Finishers	19 forwards, 4 midfields	Rachel Daly, Samantha May Kerr, Alessia Russo	Highly correlated with shots of all kinds low correlation with defence
Interception Masters	14 defenders, 48 midfielders, 10 forwards	Aileen Wheelan, Kate Longhurst, Lucy Hope	Highly correlated with duels, 50/50s and blocks
Dynamic Distributors	1 defenders, 8 midfielders, 1 forwards	Freya Gregory, Emily Simpkins, Libby Bance	High correlation with carry length and pass length
Defensive Midfield Enforcers	24 defenders, 2 midfielders	Stephanie Elise Catley, Jessica Carter, Demi Stokes	Contribute to long range shots and attacks from defensive positions
Goalkeeps	Goalkeepers	Mary Earps	N/A
All-round Playmakers	35 midfielders, 14 forwards	Lauren James, Hayley Emma Raso, Katie Robinson	Show similar correlation with all event types
Wildcards	1 goalkeeper, 3 defenders, 2 midfielders	Manuela Zinsberger, Anna Filbey, Fuka Nagano	Very highly correlated with long shots
Intercepting Strikers	26 midfielders, 1 forwards	Vivianne Miedema, Bethany Mead, Francesca Kirby	A higher correlation with shots but involved in all events
Defensive Shield	27 defenders, 9 midfielders	Esmee de Graaf, Jorja Fox, Natasha Harding	High interception, block and dribble counts
Ball Clearers	37 defenders	Gilly Louise Scarlett Flaherty, Maya Le Tissier, Niamh Fahey	High correlation with clearances and blocks
Versatile Backs	22 defenders, 8 midfielders	Leah Williamson, Lucy Bronze, Keira Walsh	Very high correlation with pass and carry counts and passes in midfield third while also holding down defence