Understanding the Dynamics of Women's Football through Clustering

Yvenn Amara-Ouali [1,2]

[1] *University Paris-Saclay, LMO, Orsay, FRANCE,*
yvenn.amara-ouali@universite-paris-saclay.fr
[2] *EDF R&D, OSIRIS, Palaiseau, FRANCE*

Résumé. La montée en puissance du football féminin ces dernières années nécessite une compréhension plus approfondie de sa dynamique et des caractéristiques des joueuses. En utilisant des techniques statistiques avancées, cet article vise à explorer les subtilités du football féminin grâce à analyse par *clusters*. En analysant les données de performance des joueuses provenant de diverses compétitions, nous cherchons à identifier des types de joueuses distincts en fonction de leurs postes et de leurs styles de jeu. Nous utilisons une combinaison de modèles statistiques, d'algorithmes d'apprentissage automatique et de techniques d'apprentissage profond pour regrouper efficacement les joueuses. Nos résultats préliminaires montrent le potentiel de ces méthodes pour caractériser la diversité et la complexité du football féminin. À travers cette recherche, nous visons à fournir des informations pertinentes pour les entraîneurs et analystes afin d'améliorer le développement des joueuses, la tactique et le niveau global du sport.

Mots-clés. Football féminin, *clustering*, caractérisation statistique, apprentissage automatique, apprentissage profond.

Abstract. The rising prominence of women's football in recent years necessitates a deeper understanding of its dynamics and player characteristics. Leveraging advanced statistical techniques, this paper aims to explore the nuances of women's football through clustering analysis. By analyzing player performance data from various competitions, we seek to uncover distinct player types based on their positions and playing styles. We employ a combination of statistical models, machine learning algorithms, and deep learning techniques to cluster players effectively. Our preliminary results demonstrate the potential of these methods in characterizing the diversity and complexity of women's football. Through this research, we aim to provide valuable insights for coaches, analysts, and stakeholders to enhance player development, team strategies, and overall game performance.

Keywords. Women's football, clustering analysis, player characterization, statistical models, machine learning, deep learning.

1 Women's football on the rise

Women are certainly not new to the game of football, with one of the oldest games recorded taking place in 1881 in Scotland. However, in recent years the number of competitions,

professional female players and spectators has significantly increased. In the 2023 FIFA Women's World Cup, ticket sales targets were smashed as 1,978,274 fans watched the matches played in the tournament across ten stadiums [1]. As the number of competitions and players increase, the emergence of advanced statistics and data can help to better characterise and capture the subtleties of the game. This will be particularly relevant in the run up to and during the upcoming Olympics in Paris 2024.

In 2012, Lydia Nsekera, the President of the Burundi FA, became the first co-opted female member of the FIFA Executive Committee [2]. This comes in contrast with the discussion of the importance of women's football in the wider society by [3], "Through their participation they are destabilising notions about gender roles, as well as notions about the physical and natural difference between men and women, masculinity and femininity." As professional women's football grows in popularity, it is vital that data and statistics are used to improve the sport and to measure and highlight ways to increase and ensure equality between the men's and women's game.

2 Related Work

The field of Women's football research has been evolving significantly over the past two decades [2]. Multiple studies have been led to characterise female football players [4]. For instance, the study presented in [5] examined morphological differences, body composition, and lower extremity explosive strength among 20 female football players in the Serbian Super League, categorizing them by playing positions. Significant disparities in explosive power were noted between midfielders and attackers, but no notable variations in body composition emerged. This might indicate that physical abilities and advanced in-game statistics might be an alternative way to characterise women's football players.

The work we propose builds upon the analysis proposed at the 2023 StatsBomb conference with a work on clustering women's football players [6]. The authors used a PCA over a wide range of event features to then perform clustering using a Bayesian Gaussian Mixture Model (BGMM) including players with more than 1000 minutes game time. However, we've found that the clustering performed may have some limitations as it gathers all players regardless of their position throughout the game (see Figure 1). Our goal with this proposed paper is to correct this inherent bias by performing clustering of players for each player position on the pitch to actually derive features of players that are specific in a certain position played.

3 Methods

In this section we briefly introduce the methods that will be used to perform clustering using statistical models (e.g., Gaussian Mixtures), machine learning (e.g., k-means and DBSCAN) and deep learning (e.g, Autoencoders).

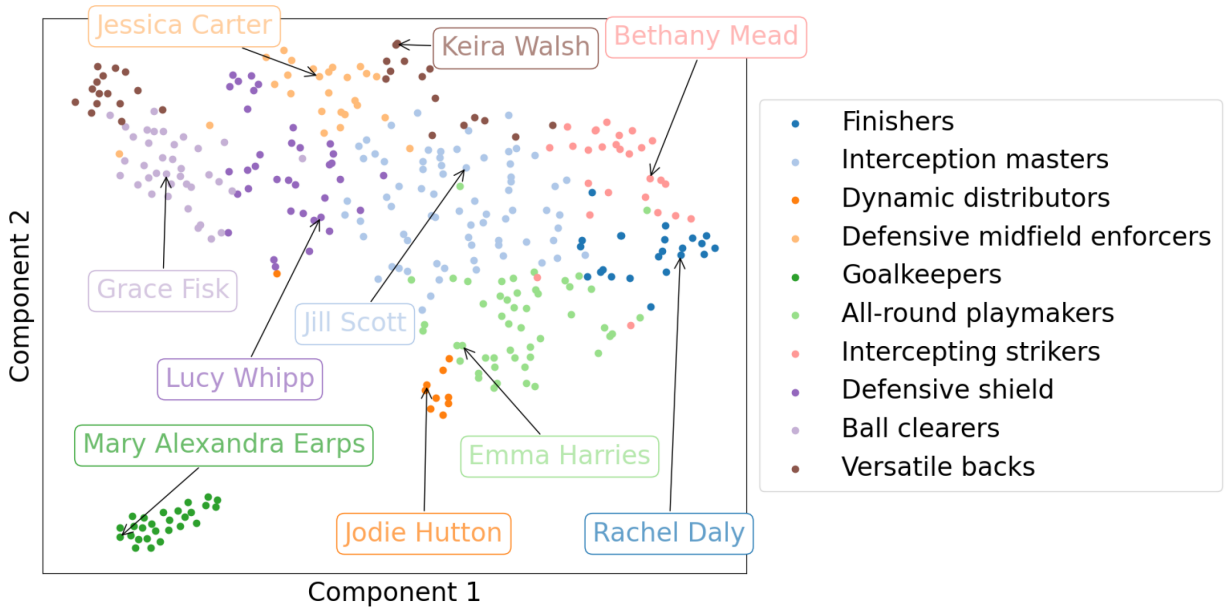

Figure 1: Clusters visualed over the two first components of the PCA in [6]

3.1 Algorithms

K-means K-means clustering is a popular unsupervised learning algorithm used for partitioning a dataset into K clusters. The goal is to minimize the within-cluster variance, where each cluster is represented by its centroid, a point that minimizes the sum of squared distances from all points in the cluster. The algorithm iteratively assigns each data point to the nearest centroid and updates the centroids based on the mean of the points assigned to each cluster. This process continues until convergence, where the assignments and centroids stabilize, or until a predefined number of iterations.

Formally, let $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}\}$ be the dataset with n data points in a d-dimensional space. The algorithm seeks to minimize the objective function:

$$J = \sum_{i=1}^{n} \min_{j} ||\boldsymbol{x_i} - \boldsymbol{\mu_j}||^2$$

where μ_j is the centroid of cluster j, and $|| \cdot ||$ denotes the Euclidean distance.

In this work we iteratively choose initial centroids that are farther away from previously selected ones, aiming to improve the convergence and quality of the final clustering, aslo called K-means++ [7].

Gaussian Mixture Models Gaussian Mixture Models (GMMs) are probabilistic models used for representing the distribution of data as a mixture of several Gaussian distributions. Each Gaussian component represents a cluster in the data, and the model parameters include

the mean, covariance, and mixing coefficients for each component. Given a dataset $X = \{x_1, x_2, ..., x_n\}$, GMM assumes that each data point x_i is generated from one of K Gaussian distributions with probabilities governed by mixing coefficients π_k and parameters $\{\mu_k, \Sigma_k\}$, where μ_k is the mean and Σ_k is the covariance matrix of the k-th Gaussian component.

The probability density function of GMM is expressed as:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the density function of a Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

GMMs are often estimated using the Expectation-Maximization (EM) algorithm, which iteratively updates the parameters to maximize the likelihood of the observed data. In this work, we will be using the R package `mclust` for implementing GMM clustering [8].

DBSCAN DBSCAN [9] is a density-based clustering algorithm that groups together data points that are closely packed, while marking points in low-density regions as outliers or noise. It defines clusters as continuous regions of high density separated by regions of low density. Given a dataset $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}\}$, DBSCAN defines two parameters: ε, the maximum distance between two points to be considered as neighbors, and $minPts$, the minimum number of points required to form a dense region (cluster). DBSCAN operates by iteratively exploring the neighborhood of each data point. Points with a sufficient number of neighbors within distance ε are considered core points and form the core of a cluster. Points that are reachable from core points but do not have enough neighbors to be core themselves are considered border points and are assigned to the nearest core point's cluster. Points that are not core or border points and do not belong to any cluster are considered noise points. DBSCAN is robust to noise and can identify clusters of arbitrary shapes and sizes without prior knowledge of the number of clusters.

Autoencoders Autoencoders are neural network models used for unsupervised learning that aim to learn efficient representations of input data by reconstructing the input from a compressed latent space. An autoencoder consists of an encoder network that maps the input data to a lower-dimensional latent space representation and a decoder network that reconstructs the input data from this representation. The network is trained to minimize the reconstruction error, typically measured as the difference between the input and output data. We usually use the binary cross-entropy loss or the mean squared error as the reconstruction loss. Autoencoders can be applied to clustering by leveraging the latent space representations learned during the training process. Once the autoencoder is trained on the input data, the encoder network can be used to map data points to a lower-dimensional latent space. Clustering algorithms such as K-means or DBSCAN can then be applied to the latent space representations to group similar data points together. This approach enables unsupervised clustering of high-dimensional data by first reducing the dimensionality using autoencoders,

which can capture complex nonlinear relationships in the data. Autoencoders have been shown to be an extension of PCA with non-linear transforms.

3.2 Metrics

In our exploration of clustering techniques, we employed several metrics to evaluate the efficacy and quality of the clustering results. For Gaussian Mixture Models (GMM), we utilized the Bayesian Information Criterion (BIC), defined as:

$$\text{BIC} = -2\log(L) + k\log(n)$$

where L is the likelihood of the data given the model, k is the number of parameters in the model, and n is the number of data points. BIC helps in determining the optimal number of clusters by balancing the goodness of fit with the complexity of the model.

For K-means clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), we employed the elbow method. The elbow method aids in determining the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) against the number of clusters. The point at which the rate of decrease in WCSS slows down, forming an "elbow" shape in the plot, indicates an optimal number of clusters.

In the case of Autoencoder (AE) based clustering, we first assessed the reconstruction error, defined as:

$$\text{Reconstruction Error} = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2$$

where $\boldsymbol{x}_i$ is the input data point, $\hat{\boldsymbol{x}}_i$ is its reconstruction obtained from the autoencoder, and N is the number of data points. Subsequently, for further evaluation of the clustering performance, we utilized clustering validation indices such as the Davies-Bouldin Index (DBI) or the Silhouette score. The DBI is calculated as:

$$\text{DBI} = \frac{1}{K}\sum_{k=1}^{K} \max_{k \neq k'} \left(\frac{\bar{\delta}_k + \bar{\delta}_{k'}}{d(m_k, m_{k'})} \right)$$

where K is the number of clusters, m_k is the centroid of cluster k , $\bar{\delta}_k$ is the average distance from m_k to the points of cluster k and $d(m_k, m_{k'})$ the distance between the centroids of clusters k and k'. The Silhouette score is given by:

$$\text{Silhouette Score} = \frac{1}{N}\sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)}$$

where a_i is the average distance from x_i to other points in the same cluster, and b_i is the smallest average distance from x_i to points in a different cluster.

4 Experiments

In this section we present the StatsBomb data used in this study and the results obtained.

4.1 Data

StatsBomb [10] offers a comprehensive dataset covering various facets of football analysis, including event data, tracking data, advanced metrics, player and team performances and many more. The event data, in particular, provides a granular breakdown of every on-ball action occurring during matches (more than 3000 per game). These actions include passes, shots, tackles, interceptions, fouls, and more (see Table 1). Each event is accompanied by detailed attributes such as the player involved, the location on the pitch, the outcome, and additional contextual information. This rich dataset empowers researchers and analysts to delve deep into player performance analysis, team tactics, and game strategies. This dataset will be used through various Women's football competitions to characterise different player types. This study focuses on 473 matches that involved women's football players. Only players with more than 1000 minutes played across these matches were kept (198 players).

Variable	**Meaning**
Event (e.g, Pass, Carry)	Average number of occurences per 90 minutes played
pass_lengthvar	Variance of pass distance (in yards)
pass_lengthmean	Mean of pass distance (in yards)
pass_anglevar	Variance of pass angle (in rad)
pass_anglemean	Mean of pass angle (in rad)
durationvar	Variance of carry duration (in seconds)
durationmean	Mean of carry duration (in seconds)
Ground Pass	Proportion of Ground Passes
High Pass	Proportion of High Passes
Low Pass	Proportion of Low Passes

Table 1: Relevant variables used for clustering

4.2 Results

It appears that the various clustering methods returned an optimal number of clusters around 21 (see Figure 2). With this number of clusters used in the kmeans algorithms, the feature pass_lengthvar emerges as a key contributor (see Figure 3), as indicated by its high variance ratio between the within-cluster sum of squares to total sum of squares for each feature. This suggests that the variability in passing lengths plays a significant role in distinguishing between clusters. Despite thorough exploration, the DBSCAN and Autoencoder models used in this study did not demonstrate superior performance or offer additional insights beyond those obtained from K-means and Gaussian Mixture Models (GMM).

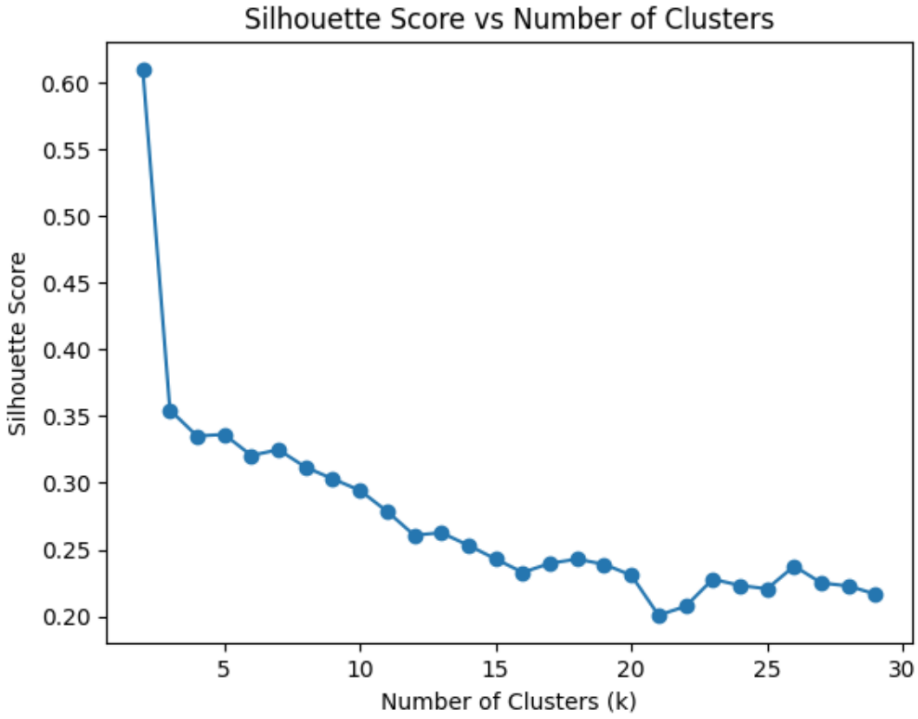

Figure 2: Silhouette scores for $k \in 2 \dots 30$ clusters with kmeans

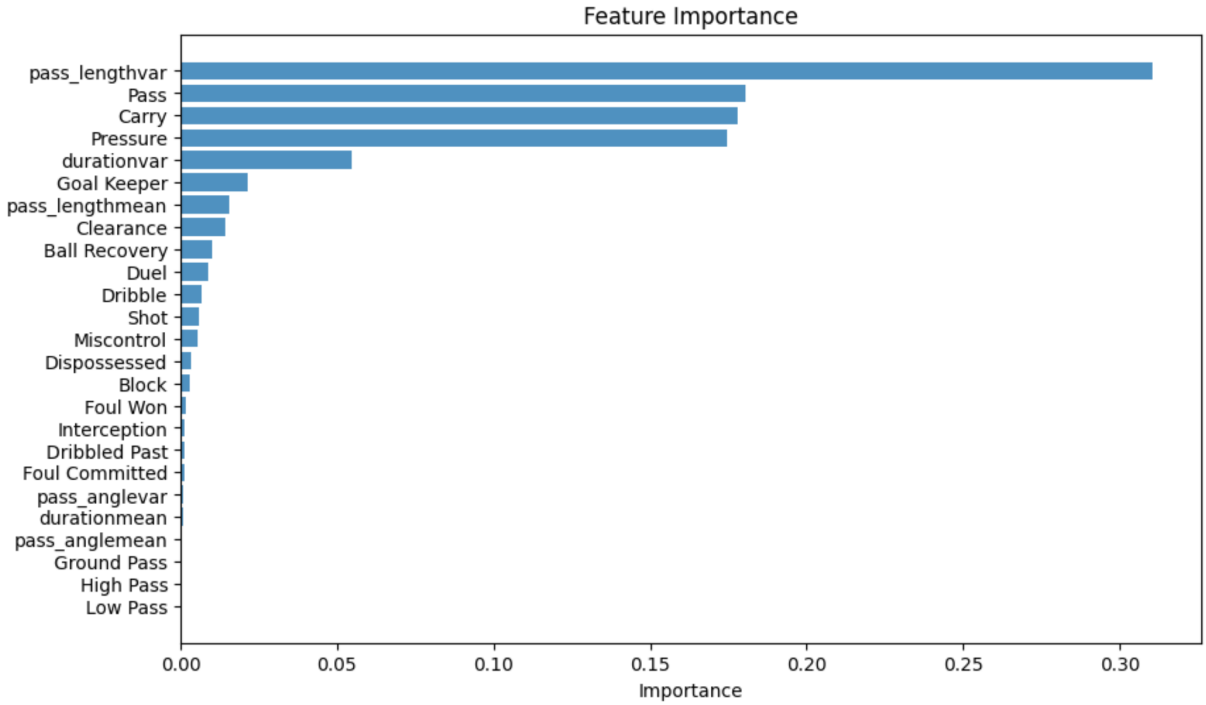

Figure 3: Feature importance calculated using the variance ratio between the within-cluster sum of squares to total sum of squares for each feature

5 Conclusion

In conclusion, this study proposed an initial exploration into characterizing women's football players with clustering methods. While our findings provide valuable insights into clustering methods and prominent features such as pass length variance, it is important to recognize the preliminary nature of this work. Future research endeavors could delve deeper into this, focusing on refining clustering techniques and incorporating additional features to enhance player characterization. Notably, our observations suggest that pass length variance serves as a robust indicator of the diversity in players' styles, with strikers demonstrating preference for playing in confined spaces, midfielders exhibiting a mix of short and long passes, and defenders favoring longer passes. Further investigations into these nuances could yield valuable implications for player development and tactical strategies in women's football.

References

[1] SportsPro, "Women's world cup 2023: Attendance figures, viewership & social media," 2023.

[2] J. Williams and R. Hess, "Women, football and history: international perspectives," *The international journal of the history of sport*, vol. 32, no. 18, pp. 2115–2122, 2015.

[3] M. H. Engh, "Tackling femininity: The heterosexual paradigm and women's soccer in south africa," in *Sport Past and Present in South Africa*, pp. 136–151, Routledge, 2013.

[4] V. Martinez-Lagunas, M. Niessen, and U. Hartmann, "Women's football: Player characteristics and demands of the game," *Journal of Sport and Health Science*, vol. 3, no. 4, pp. 258–272, 2014.

[5] K. Goranovic, A. Lilić, S. Karišik, N. Eler, M. Anelić, and M. Joksimović, "Morphological characteristics, body composition and explosive power in female football professional players.," *Journal of Physical Education & Sport*, vol. 21, no. 1, 2021.

[6] M. Trower, N. Graham, N. Cottrell, and Y. Hengster, "Clustering women's football players," *StatsBomb Conference*, 2023.

[7] D. Arthur, S. Vassilvitskii, *et al.*, "k-means++: The advantages of careful seeding," in *Soda*, vol. 7, pp. 1027–1035, 2007.

[8] C. Fraley, A. E. Raftery, L. Scrucca, and M. L. Berrendero, *mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation*, 2022. R package version 5.4.14.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231, 1996.

[10] Benjamin Robinson, *statsbombR: R Client for the 'StatsBomb' API*, 2022. R package version 0.5.1.