

An investigation of data-driven player positional roles within the Australian Football League Women's competition using technical skill match-play data

Van der Vegt, Braedan; Gepp, Adrian; Keogh, Justin W L; Farley, Jessica B.

International Journal of Sports Science and Coaching

DOI:

[10.1177/17479541231203895](https://doi.org/10.1177/17479541231203895)

E-pub ahead of print: 05/10/2023

Publisher's PDF, also known as Version of record

[Cyswllt i'r cyhoeddiad / Link to publication](#)

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):

Van der Vegt, B., Gepp, A., Keogh, J. W. L., & Farley, J. B. (2023). An investigation of data-driven player positional roles within the Australian Football League Women's competition using technical skill match-play data. *International Journal of Sports Science and Coaching*. Advance online publication. <https://doi.org/10.1177/17479541231203895>

Hawliau Cyffredinol / General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

An investigation of data-driven player positional roles within the Australian Football League Women's competition using technical skill match-play data

International Journal of Sports Science
& Coaching
1–13

© The Author(s) 2023



Article reuse guidelines:

sagepub.com/journals-permissions
DOI: 10.1177/17479541231203895
journals.sagepub.com/home/spo



Braedan van der Vegt¹ , Adrian Gepp^{1,2}, Justin Keogh^{3,4,5} ,
and Jessica B. Farley³

Abstract

Understanding player positional roles are important for match-play tactics, player recruitment, talent identification, and development by providing a greater understanding of what each positional role constitutes. Currently, no analysis of competition technical skill data exists by player position in the Australian Football League Women's (AFLW) competition. The primary aim of the research was to use data-driven techniques to observe what positions and roles characterise AFLW match-play using detailed technical skill action data of players. A secondary aim was to comment on the application of clustering methods to achieve more interpretable, reflective positional clustering. A two-stage, unsupervised clustering approach was applied to meet these aims. Data cleaning resulted in 165 variables across 1296 player seasons in the 2019–2022 AFLW seasons which was used for clustering. First-stage clustering found four positions following a common convention (forwards, midfielders, defenders, and rucks). Second-stage clustering found roles within positions, resulting in a further 13 clusters with three forwards, three midfielders, four defenders, and three ruck positional roles. Key variables across all positions and roles included the field location of actions, number of contested possessions, clearances, interceptions, hitouts, inside 50s, and rebound 50s. Unsupervised clustering allowed the discovery of new roles rather than being constrained to pre-defined existing classifications of previous literature. This research assists coaches and practitioners by identifying key game actions players need to perform in match-play by position, which can assist in player recruitment, player development, and identifying appropriate match-play styles and tactics, while also defining new roles and suggestions of how to best use available data.

Keywords

Field location, performance indicators, sport analytics, team

Introduction

Australian Football (AF) can be described as a 'dynamic invasion team sport'.¹ The national competition for women's AF, the Australian Football League Women's (AFLW) competition, was established in 2017. Research on the AFLW competition has included characterising physical^{2–4} and technical match-play performance.^{5,6} Initial research has discovered key technical performance indicators such as disposals, disposal efficiency, contested and uncontested possessions, marks, marks inside 50, contested marks, and inside 50s are linked to match-play success.^{5,6} The findings of these match-play studies have the potential to support coaches regarding player performance development and match-play tactics, which at the start of the AFLW had been derived from the perspective of the male equivalent Australian Football League (AFL).⁷

Reviewers: Ulf Brefeld (Leuphana University of Lüneburg, Germany)
Sam Robertson (Victoria University, Australia)

¹Centre for Data Analytics, Bond Business School, Bond University, Gold Coast, QLD, Australia

²Bangor Business School, Bangor University, Bangor, Wales, UK

³Faculty of Health Sciences and Medicine, Bond University, Gold Coast, QLD, Australia

⁴Sports Performance Research Centre New Zealand, Auckland University of Technology, Auckland, New Zealand

⁵Kasturba Medical College, Mangalore, Manipal Academy of Higher Education, Manipal, KA, India

Corresponding author:

Braedan van der Vegt, Centre for Data Analytics, Bond Business School, Bond University, 14 University Dr, Gold Coast, QLD 4226, Australia.
Email: bvegt@bond.edu.au

The AFLW competition has developed greatly in the first years of its existence, with the preliminary investigation of Dwyer et al.⁸ showing an increase in match-play performance metrics over subsequent seasons (2017–2019 seasons).⁹ This has been suggested to be attributable to a variety of factors including the greater opportunities for talent development associated with the establishment of an elite league, new youth pathways, and increased facilities and expertise.^{7,8} Player positional roles are anticipated to also evolve, reflecting the development of match-play performance. Statistically significant differences in physical and technical match-play performance by player position have been reported in both the men's^{10,11} and women's games.^{3,4} However, to what extent the positional roles follow that of the men's equivalent AFL may be unclear.

Important differences exist between elite women's (AFLW) and men's AF competitions (AFL), with the AFLW having shorter quarter lengths and fewer players on the field (16 vs. 18). Fewer players on the field in the AFLW have generally meant one less player in each of the forward and defensive areas on team line-ups, but whether this is reflected in match-play has not been investigated. Departures between the competitions also exist in physiological differences of players and less established development pathways, as well as the level and length of access to high-performance facilities, all with the potential to result in different positional roles and match-play styles.³ These potential differences in positional roles between the AFL and AFLW may be greater than in other sports because these positions are more fluid in AF compared to many other sports.¹² Research following and extending precedents of playing position and role investigation has not been conducted in the AFLW with consideration to the context surrounding the different competition and overall environment surrounding women's AF. A greater understanding of what each positional role constitutes that is reflective of the women's competition may be useful to the AFLW with respect to match-play tactics, player recruitment, talent identification, and development.

The methodology of producing these positional classifications in the AFLW is also of interest, with a look to equivalent research in men's AF as a worthwhile point of comparison. In the men's game, the position of players has been investigated in an *a priori* or supervised manner (i.e. using existing player position assessment to understand what variables are important).^{12–14} These studies in the men's AF literature have used *a priori* methods of dimension reduction by removing correlated variables and choosing variables that have high importance in the prediction of match results or prior knowledge of what differentiates positions.^{12–14} Previous literature precedents of selecting variables by importance to *a priori* positions or to match-play success may bias results as the variables that differentiate between positions are not necessarily identical to those that determine the match result. Previous studies have also had individual issues such as the use of aggregations of multiple variables

obscuring important individual actions¹² and inability to classify position successfully either through not enough variables or different roles between junior and elite male competitions.¹⁴ While Barake et al.¹³ were successful in classifying positions and even pre-defined roles for male players at a sub-elite level, this was achieved under the assumption of the *a priori* positional and roles detail of the AFL being applicable to male, sub-elite competitions. Nevertheless, the results of these studies have had applications for team recruitment, benchmarking player performance, and style of play; along with how the creation of these analyses may be impacted by aspects of the data currently available, or analysis techniques used to date.

Given the mixed use and results of previous applications to determine positional classifications in the AFL, the creation of a reproducible method that can be continually updated, producing interpretable, data-driven insights for practitioners that considers the challenges that the environment specific to the AFLW and wider women's sport present, is warranted. Creating this representation of player position will produce insights into the use of current data, comment on the appropriateness of different methodologies, and suggestions surrounding data cleaning protocols used in practice by data analysts. Additionally, while initial research on women's AF has been identified, current AF literature reflects an issue in broader sports literature with a dearth of research focused on female athletes relative to the men's equivalent which this study can help to begin to alleviate.^{7,15}

Consequently, the primary aim of this analysis was to use data-driven techniques to observe what positions and roles characterise AFLW match-play using detailed technical skill action data of players. A secondary aim was to comment on the application of clustering methods to achieve more interpretable, reflective positional results in the AFLW considering current and potential future data availability that is reproducible and comparable.

Methods

Data

Each player's seasonal average of basic performance indicators (Supplemental Material 1) captured by Champion Data match-play statistics from 2019 to the first 2022 AFLW season were used in this study. These basic performance indicators were measured using data collection procedures similar to that currently used in the AFL, which have demonstrated accuracy within the men's competition, although no reliability verification has been conducted in the AFLW.¹⁶

Time spent in each locational zone (Figure 1) by a player throughout a match is used to determine position by Champion Data statistics service in the men's game.¹³ Subsequently, locational detail was sought for this analysis due to its potential to provide greater insight into player

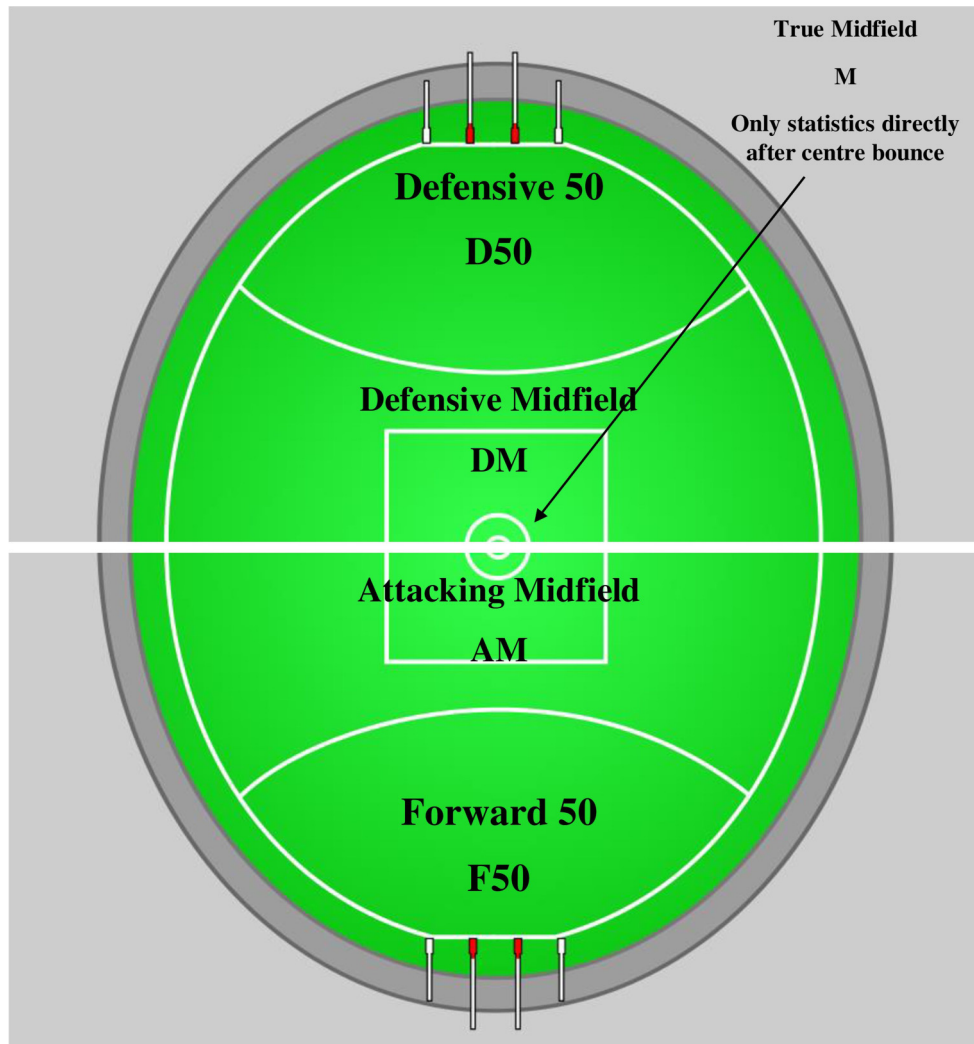


Figure 1. Locational information available in data. This work was built ('remixed') upon the previous open licence work available at: https://commons.wikimedia.org/wiki/File:AFL_stadium.svg. Attribution: Cflm001, CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0/>, via Wikimedia Commons.

roles. Each game-action variable had its associated field location available in data and followed five zones on the field as determined and assessed by Champion Data (attacking midfield [AM], defensive midfield [DM], true midfield [M], forward 50 [F50] and defensive 50 [D50]). These field locations can be seen in Figure 1. As such, the additional context of the field location of each statistic was available to be pivoted onto the dataset. Pivoting the locational component onto each basic variable created new versions of the variable for each zone (e.g. Kicks became Kicks_F50, Kicks_AM, Kicks_D50 and Kicks_DM, Kicks_M), resulting in a total of five variables for each basic statistic.

A summary of data cleaning steps and justifications can be seen in Table 1. All data and methods have received ethical approval by the Bond University Human Research Ethics Committee (BV00011).

Statistical analysis

Due to the number of variables, many of which being highly correlated with one another, data was normalised, before principal components analysis (PCA) was applied, helping to enable better clustering performance. Additionally, the clustering large applications (CLARA) using the partition around medoid (PAM) technique from the cluster R package¹⁷ was used due to the advantages surrounding sampling methods used to assist in ensuring the most relevant features are utilised. The CLARA algorithm performs iterative clustering meaning that further variable reduction does not need to be performed as variable selection is conducted through the clustering process. All stages of data cleaning and modelling were conducted in free, open-source software, R.¹⁸

Clustering was applied in a two-step process to improve the ability to cluster on more position-specific differences in

Table 1. Data cleaning steps used.

Data Cleaning Action	Rationale
1. Locational component pivoted onto existing variables	Allows for incorporation of where an action occurs which can have an important effect on clustering for a player's position.
2. Removal of player seasons with two or fewer matches played	To ensure that only player-seasons that are representative are included as often players with fewer matches are injured. Two matches were selected as a threshold due to the short season lengths of the Australian Football League Women (AFLW).
3. Variables with > 99.5% of data points equal to zero removed	Threshold selected to remove primarily data with no clustering power. A more typical threshold of 95% was deemed inappropriate as given that there are 21 players on a team, an individual player represents ~ 5% of a team. Therefore, a smaller threshold ensures that variables that are specific to one player on a team consistently remain captured.
4. Data normalised relative to the dataset	Conducted in both stages of modelling. The second stage was conducted relative to its first-stage cluster subset to better capture intra-position variance compared to when scaled against all player data.
5. Principal components analysis (PCA) applied	While the chosen clustering algorithm can handle a large number of variables, highly correlated variables can result in over-weightings of some correlated variables. As a result, variable reduction through PCA was used to enable a fairer relative representation of each variable. Conducted in both stages of modelling.

the second stage. This is similar to Barake et al.,¹³ who used a secondary step to split players within a position. This was performed by subsetting the data based on their first-stage cluster and then steps 3 to 5 described in Table 1 were repeated on each subset before second-stage clustering was conducted. This is important because re-normalising data relative to its own cluster allows for variations specific to the cluster that are not as obvious when comparing across all players.

The CLARA algorithm in both stages was applied with a prespecified number of clusters between two and 20. The appropriate number of clusters was then determined by firstly finding that which minimised the average cluster widths or 'silhouette' measure as in Wedding et al.¹⁹ to result in the tightest, most distinct clusters through a metric that gives a relative figure of 'tightness' by the number of clusters that can be compared to other cluster amounts. It should be noted that this metric cannot be compared between datasets or other techniques due to differences in variation meaning its use should only be relative to other cluster amounts on the same data. Following this, individual clusters were manually checked to ensure clusters without data points were not present which would skew the silhouette metric through the creation of meaningless clusters.

As per our aim to observe what positions and roles within these positions emerge through data-driven techniques, it is not our purview to comment on the performance of players within positions. For this reason, clustering that can be seen to result in differences due to performance was to be reconsidered by merging these clusters to better distinguish between positions and roles. It was also part of our research ethical approval that we do not comment on the performance of individual players.

Variable importance interpretation

While the application of PCA multiple times provides benefits to identifying representative clusters, it does make ascertaining individual variable importance more difficult.²⁰ As a result, a method was used to find the underlying variables most important for distinguishing between both first- and second-stage clusters. This method of variable importance interpretation was achieved by applying random forest models using the original, non-PCA data to predict the cluster result in each stage, with results showing the most important variables to determine the cluster. Observing key variables via the mean decrease in Gini index (MDGI) provided a metric of relative variable importance, allowing for comparison between clusters using variables that best distinguish each grouping.²¹ Mean decrease in accuracy attributable to each variable was also assessed as a measure of importance and was found to give very similar results. Differences in the importance of variables between clusters were best represented by descriptive tables comparing cluster characteristics.

Results

There were a total of 33 variables across 1425 player seasons across the 2019–2022 competitive AFLW seasons accessible within the Champion Data dataset. After pivoting the locational component onto variables and completing the data cleaning steps, 165 variables across 1296 player seasons were included for analysis. The full process of the application of the data cleaning, clustering, and its interpretation can be seen in Figure 2.

PCA application on the first stage data of all player seasons resulted in 20 principal components that explain

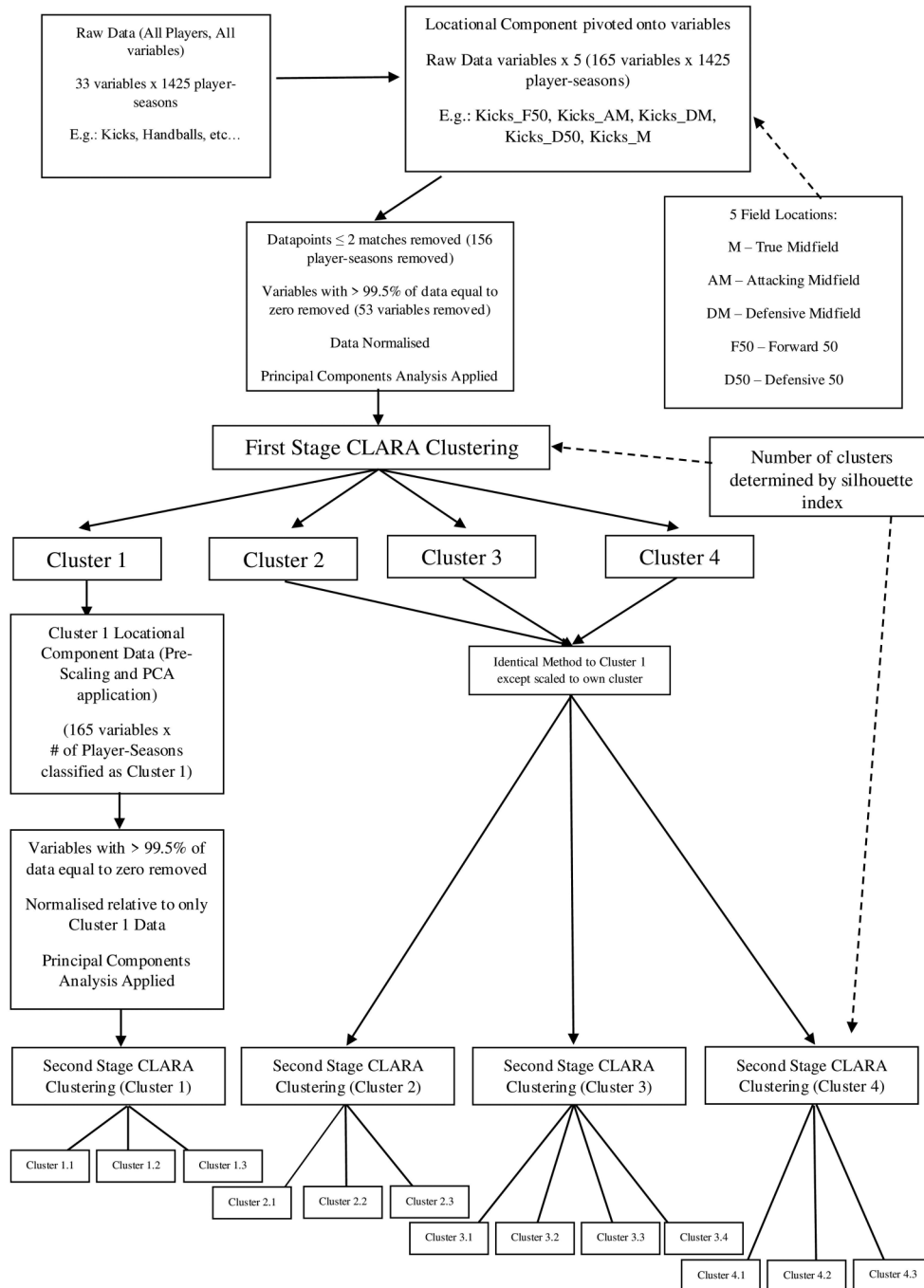


Figure 2. Data cleaning and clustering methods. PCA: principal components analysis.

71.49% of the variance in the dataset. First-stage clustering resulted in seven clusters being selected. Seven clusters resulted in a minimised silhouette metric of 0.1189. While there were lower averages in higher cluster numbers (0.102 for eight clusters), these contained empty clusters that skewed the metric.

The initial seven clusters were found to exist in pairs for each traditional position (i.e. two midfielder,

forward, and defender clusters plus one ruck cluster). However, the distinguishing factor of these pairs was determined to be by player performance in all important variables of their assignment (i.e. all important variables were just higher in value relative to its pair cluster). To combat this, the seven clusters were reduced to four, following the four primary positional categories (Forward, Defender, Midfielder, and Ruck) by combining the

Table 2. Stage 1 clusters descriptive table ordered by highest variable importance to clustering of AFLW positions.

Variable	Cluster 1 (mean \pm SD) Midfielders	Cluster 2 (mean \pm SD) Forwards	Cluster 3 (mean \pm SD) Backs	Cluster 4 (mean \pm SD) Rucks	MDGI
Number of players	336	367	475	91	
DISPOSAL_F50	2.19 \pm 0.94	2.66 \pm 1.36	0.34 \pm 0.4	0.78 \pm 0.75	46.13
DISPOSAL_D50	1.86 \pm 0.99	0.45 \pm 0.49	3.45 \pm 1.88	1.24 \pm 0.9	43.90
DISPOSAL_AM	4.97 \pm 1.62	2.36 \pm 0.91	1.7 \pm 0.85	2.56 \pm 1.2	40.94
EFFECTIVE_DISPOSAL_D50	1.2 \pm 0.69	0.25 \pm 0.32	2.35 \pm 1.44	0.81 \pm 0.66	36.95
KICK_F50	1.43 \pm 0.72	1.8 \pm 1.05	0.22 \pm 0.28	0.49 \pm 0.56	32.01
KICK_TO_HANDBALL_RATIO_F50	1.39 \pm 0.7	1.76 \pm 1.02	0.22 \pm 0.28	0.48 \pm 0.56	30.48
CONTESTED_POSSESSION_F50	1.13 \pm 0.62	1.5 \pm 0.88	0.15 \pm 0.23	0.54 \pm 0.54	28.41
KICK_D50	1.14 \pm 0.73	0.31 \pm 0.39	2.4 \pm 1.57	0.62 \pm 0.53	27.90
DISPOSAL_DM	5.14 \pm 1.95	1.9 \pm 0.89	3.07 \pm 1.13	2.67 \pm 1.13	26.76
EFFECTIVE_DISPOSAL_AM	3.07 \pm 1.12	1.41 \pm 0.65	1.07 \pm 0.61	1.57 \pm 0.85	25.13
EFFECTIVE_KICK_D50	0.65 \pm 0.46	0.16 \pm 0.25	1.58 \pm 1.2	0.36 \pm 0.38	25.09
KICK_TO_HANDBALL_RATIO_D50	1.11 \pm 0.7	0.3 \pm 0.39	2.31 \pm 1.5	0.6 \pm 0.5	24.95
EFFECTIVE_DISPOSAL_F50	1.05 \pm 0.54	1.34 \pm 0.77	0.17 \pm 0.21	0.39 \pm 0.41	23.62
INTERCEPT_D50	0.51 \pm 0.38	0.16 \pm 0.25	1.57 \pm 0.95	0.55 \pm 0.49	21.14
CLEARANCE_DM	1.14 \pm 0.79	0.15 \pm 0.27	0.12 \pm 0.21	0.71 \pm 0.55	18.68
CONTESTED_POSSESSION_D50	0.94 \pm 0.55	0.22 \pm 0.27	1.43 \pm 0.75	0.83 \pm 0.55	18.24
HITOUT_D50	0.01 \pm 0.06	0.04 \pm 0.19	0.02 \pm 0.1	2.3 \pm 1.18	17.64
HITOUT_DM	0.03 \pm 0.22	0.15 \pm 0.62	0.03 \pm 0.18	6.47 \pm 2.95	17.34
EFFECTIVE_DISPOSAL_DM	3.24 \pm 1.32	1.2 \pm 0.63	2 \pm 0.82	1.76 \pm 0.79	17.33
CONTESTED_POSSESSION_DM	2.43 \pm 1.12	0.84 \pm 0.48	1.28 \pm 0.6	1.55 \pm 0.73	16.60

MDGI: mean decrease Gini index; AFLW: Australian Football League Women; AM: attacking midfield; DM: defensive midfield; M: true midfield; F50: forward 50, D50: defensive 50. Cells shaded and in bold represent the cluster with the highest value for that variable across the four clusters. All variables are the count of the number the game actions performed unless indicated as being a ratio.

pairs of similar clusters into a new subset. Table 2 presents descriptive statistics of the most important variables that determined cluster classification with the highest mean highlighted in each row. These variables were the highest 20 by importance (largest MDGI) in a random forest model (accuracy: 89.28%).

Table 2 shows key statistics to determining first-stage clusters including the location of disposals (disposals in F50, D50, and attacking midfield). The location of other statistics like that of effective disposals, kicks, as well as contested possession are also indicators of player position. Intercepts in D50, clearances, and hitouts are other notable variables that further differentiated positions with less of a focus on location.

PCA applied to each rescaled subset in stage 2 resulted in variance in the dataset explained of: 71.5% (midfielders), 66.47% (forwards), 66.27% (defenders), and 77.36% (rucks) when using the first 20 components. Cluster silhouette metrics were 0.079 for three clusters (midfielders), 0.096 for three clusters (forwards), 0.081 for four clusters (defenders), and 0.13 for three clusters (rucks). Random forests for interpretation were fit with accuracies of 82.74% (midfielders), 82.02% (forwards), 78.95% (defenders), and 86.81% (rucks). Clusters, important variables, and descriptive statistics for each of these four, second-stage cluster subsets can be seen in Table 3.

Discussion

This research aimed to ascertain the positions and roles that are presented by employing data-driven techniques analysing player match-play technical skill data in the AFLW. Through the two-step cluster analysis, positional classifications were created for players across the 2019–2022 seasons of the AFLW, producing definitions of technical positions and roles for players in these positions through data-driven techniques rather than being subject to biases of current thinking within the competition. The existence of hybrid roles in the second-stage clustering has not previously been defined in academic literature. Roles identified in the second-stage may provide more targeted information for players and practitioners to better direct training and development practices. Greater guidance of recruitment for teams on an individual level, as well as when considering team tactical balance is also possible as suggested in previous literature.¹³ Cluster analysis on PCA data, while initially obscuring results, has been made interpretable via random forests and descriptive tables to understand important points of difference between positional clusters.

Clustering results and practical applications

Interpreting first-stage clustering, the location of statistics, particularly disposals, was key to the differentiation of

Table 3. Stage 2 clusters descriptive table ordered by highest variable importance to clustering of AFLW positional roles.

Forwards (367 players)									
Variable	Cluster 1.1 (mean ± SD)	Cluster 1.2 (mean ± SD)	Cluster 1.3 (mean ± SD)	MDGI Variable	Cluster 2.1 (mean ± SD) High scoring forwards	Cluster 2.2 (mean ± SD) General forwards	Cluster 2.3 (mean ± SD) Forward-midfielders	MDGI	
Number of players	Two-way Mids 52	Defensive Mids 134	Attacking Mids 150	Number of players	130	138	99		
DISPOSAL_D50	2.94 ± 0.92	2.25 ± 0.78	1.11 ± 0.49	59.77	DISPOSAL_D50	0.21 ± 0.26	0.96 ± 0.53	45.96	
DISPOSAL_AM	7.36 ± 1.3	4.93 ± 1.32	4.09 ± 0.91	24.77	DISPOSAL_F50	2.14 ± 0.94	1.79 ± 1.05	24.36	
EFFECTIVE DISPOSAL_AM	4.64 ± 0.99	2.91 ± 0.93	2.61 ± 0.72	8.38	CONTESTED POSSESSION_F50	1.16 ± 0.61	0.95 ± 0.7	23.59	
KICK EFFICIENCY_F50	0.43 ± 0.13	0.31 ± 0.18	0.4 ± 0.19	6.11	DISPOSAL_DM	1.39 ± 0.68	2.65 ± 0.85	14.14	
EFFECTIVE DISPOSAL_F50	1.49 ± 0.65	0.75 ± 0.36	1.14 ± 0.48	5.91	EFFECTIVE DISPOSAL_F50	1.06 ± 0.56	0.86 ± 0.59	9.81	
DISPOSAL_DM	7.7 ± 1.53	5.58 ± 1.49	3.77 ± 1.11	5.57	EFFECTIVE DISPOSAL_D50	0.1 ± 0.17	0.56 ± 0.36	8.43	
KICK_D50	1.84 ± 0.75	1.42 ± 0.61	0.63 ± 0.37	4.87	DISPOSAL_AM	1.83 ± 0.66	2.51 ± 0.95	7.74	
KICK_DM	4.52 ± 1.07	3.22 ± 1.01	2.01 ± 0.75	4.82	KICK_TO_HANDBALL_RATIO_F50	1.35 ± 0.68	1.2 ± 0.81	6.42	
CONTESTED_POSSESSION_DM	3.77 ± 0.9	2.66 ± 0.97	1.72 ± 0.71	4.20	EFFECTIVE DISPOSAL_AM	1.08 ± 0.53	1.4 ± 0.61	5.67	
EFFECTIVE_KICK_F50	0.83 ± 0.44	0.38 ± 0.26	0.57 ± 0.35	4.07	KICK_DM	0.82 ± 0.49	1.77 ± 0.7	5.51	
EFFECTIVE_KICK_AM	2.54 ± 0.67	1.53 ± 0.58	1.27 ± 0.53	3.93	KICK_F50	1.38 ± 0.7	1.22 ± 0.83	5.20	
MARK_F50	0.36 ± 0.29	0.18 ± 0.17	0.32 ± 0.25	3.58	KICK_TO_HANDBALL_RATIO_AM	1.06 ± 0.46	1.68 ± 0.78	4.70	
CONTESTED_POSSESSION_D50	1.5 ± 0.62	1.13 ± 0.43	0.57 ± 0.3	3.02	CONTESTED_POSSESSION_AM	0.91 ± 0.44	1.15 ± 0.61	3.89	
KICK EFFICIENCY_DM	0.58 ± 0.11	0.51 ± 0.12	0.57 ± 0.13	2.45	GOAL_F50	0.37 ± 0.31	0.28 ± 0.26	3.80	
EFFECTIVE_KICK_DM	2.63 ± 0.75	1.64 ± 0.62	1.15 ± 0.5	2.44	KICK_TO_HANDBALL_RATIO_DM	0.82 ± 0.49	1.74 ± 0.69	3.63	
DISPOSAL_F50	2.94 ± 1.14	1.75 ± 0.68	2.3 ± 0.84	2.43	EFFECTIVE DISPOSAL_DM	0.9 ± 0.49	1.61 ± 0.67	3.37	
KICK_TO_HANDBALL_RATIO_DM	4.01 ± 0.95	2.97 ± 0.96	1.9 ± 0.72	2.25	MARK_AM	0.43 ± 0.3	0.58 ± 0.41	3.34	
HARD_BALL_GET_DM	1.28 ± 0.47	0.84 ± 0.46	0.52 ± 0.34	2.17	CONTESTED_POSSESSION_D50	0.11 ± 0.15	0.46 ± 0.33	3.19	
KICK_TO_HANDBALL_RATIO_D50	1.77 ± 0.72	1.39 ± 0.59	0.62 ± 0.37	2.05	REBOUND_50_D50	0.14 ± 0.18	0.62 ± 0.35	2.65	
EFFECTIVE DISPOSAL_DM	4.97 ± 1.17	3.38 ± 1.06	2.46 ± 0.83	2.01	KICK_AM	1.08 ± 0.47	1.71 ± 0.8	2.58	
Rucks (91 players)									
Variable	Cluster 3.1 (mean ± SD)	Cluster 3.2 (mean ± SD)	Cluster 3.3 (mean ± SD)	Cluster 3.4 (mean ± SD)	MDGI Variable	Cluster 4.1 (mean ± SD)	Cluster 4.2 (mean ± SD)	Cluster 4.3 (mean ± SD)	MDGI
Number of players	Defender-midfielders 117	Around the ground defenders 62	General defenders 195	High disposal defenders 101	Number of players	Around the ground ruck 30	General ruck 48	Ruck-forwards 13	
DISPOSAL_AM	2.46 ± 0.68	2.41 ± 0.76	1.21 ± 0.56	1.27 ± 0.54	60.14	KICK_TO_HANDBALL_RATIO_D50	0.34 ± 0.23	0.31 ± 0.31	7.28
EFFECTIVE DISPOSAL_D50	1.42 ± 0.74	3.55 ± 1.25	1.62 ± 0.79	3.91 ± 1.22	33.63	GOAL_F50	0.03 ± 0.07	0.43 ± 0.28	5.58
DISPOSAL_D50	2.21 ± 1.02	4.98 ± 1.66	2.54 ± 1.06	5.46 ± 1.57	32.46	DISPOSAL_D50	0.77 ± 0.45	0.7 ± 0.35	5.42
EFFECTIVE_KICK_AM	0.74 ± 0.37	0.98 ± 0.44	0.4 ± 0.34	0.4 ± 0.23	15.67	EFFECTIVE DISPOSAL_D50	0.46 ± 0.35	0.47 ± 0.23	4.86
EFFECTIVE_KICK_D50	0.84 ± 0.6	2.62 ± 1.07	0.98 ± 0.66	2.8 ± 1.09	12.70	KICK_D50	0.35 ± 0.24	0.32 ± 0.31	4.37
INTERCEPT_D50	0.85 ± 0.52	2.15 ± 0.9	1.26 ± 0.58	2.54 ± 0.9	10.62	MARK_F50	0.07 ± 0.09	0.63 ± 0.27	4.18
KICK_AM	1.43 ± 0.56	1.63 ± 0.6	0.74 ± 0.47	0.74 ± 0.33	8.41	EFFECTIVE_KICK_F50	0.07 ± 0.09	0.63 ± 0.33	3.49
CONTESTED_POSSESSION_AM	1.16 ± 0.45	0.99 ± 0.44	0.55 ± 0.31	0.6 ± 0.34	7.37	DISPOSAL_AM	1.88 ± 0.7	3.32 ± 1.49	2.33
TACKLE_DM	0.91 ± 0.5	0.72 ± 0.35	0.51 ± 0.35	0.52 ± 0.28	6.98	DISPOSAL_F50	0.46 ± 0.41	2.1 ± 0.68	2.18
EFFECTIVE DISPOSAL_AM	1.54 ± 0.52	1.57 ± 0.6	0.74 ± 0.43	0.82 ± 0.41	6.48	MARK_D50	0.11 ± 0.15	0.15 ± 0.2	1.86
TACKLE_AM	0.83 ± 0.46	0.52 ± 0.27	0.42 ± 0.29	0.32 ± 0.23	5.44	KICK_AM	0.82 ± 0.46	1.69 ± 0.95	1.51
SPOIL_D50	0.21 ± 0.21	0.73 ± 0.51	0.57 ± 0.5	1.16 ± 0.68	4.99	KICK_TO_HANDBALL_RATIO_F50	0.23 ± 0.25	1.49 ± 0.55	1.31

(continued)

midfield. Cluster 2.1 was those players deemed as high-scoring forwards, with these players most likely to have scoring kicks. The second cluster (Cluster 2.2) was not as large of a source of goals or a number of statistics compared to the high-scoring forwards. It is possible that further variables could help better distinguish the differences between Clusters 2.1 and 2.2, with elements like defensive pressure applied by players and gathering of ground balls being noted as defining roles in common AFL media match-play analysis, being potentially important activities for Cluster 2.2.²³

Defensive player clustering led to four role clusters, more than any other second-stage application. The highest subset of players ($n=475$) was found to be defensive players in stage 1, with this either stemming from a higher number of defensive players in the competition or potentially those outliers that were harder to cluster being categorised as defenders which can occur when all data-points need to be clustered as in this technique.²² In Cluster 3.1, a similar phenomenon as with the forwards was seen, with those primarily considered defenders but also accumulating higher statistics in attacking zones compared to other clusters of defensive players. Cluster 3.2 also defended higher up the ground, but not in F50, suggesting that these are more of a half-back type of defenders rather than those who defend closer to the goal. Cluster 3.4 were those that had a high number of disposals in D50, indicating that they had a large role in rebounding the ball from the backline up the field as well as defending primarily in the D50. Cluster 3.3 did not have the highest average in any statistic although they had a large sample size of players. These could be defenders that are primarily focused on defensive actions that are not necessarily contained in the Champion Data statistics (perhaps as they are marking the most dangerous goalscoring forwards of the opposition team) and less focused on having many possessions or rebounding the ball out of defence.

Rucks were classified into three clusters. These were those who were more active in getting disposals around the ground (Cluster 4.1), those who were more traditional rucks with a primary focus on winning hitouts (Cluster 4.2), and those who scored more goals, which suggests they shared ruck duties with a second ruck and spent more time up forward (Cluster 4.3). Another key characteristic of Cluster 4.1 is the presence of D50 statistics including marks suggesting the importance of this type of ruckman in defending. The rucks are naturally the smallest subset, with typically only one to two ruck players on a team in a match. Despite the low sample size, clear second-stage clusters were able to be created.

When comparing the clustering of the four major positional groups, it was observed that midfielders were the smallest, non-ruck, cluster in terms of the number of players. This was an unexpected result given that AFLW team line-ups have six nominal midfielders compared to

five nominal forwards and five nominal defenders, and more players with midfield roles are generally available on the interchange. A possible reason for this is the presence of part-time midfielders in defensive (117 players in Cluster 3.1) and forward (99 players in Cluster 2.3) second-stage clusters, which if they were to be considered midfielders, would make midfielders the biggest subset. This prominent presence of players who spend some time in the midfield is likely reflective of reality, given the high need for rotations in this position due to their greater running demands and physiological workloads.²⁴

Player position when considering the structure of a whole team is an important facet to consider, both to understand the role of positions within a team strategy and to check whether reasonable positions are being derived. Figure 3 represents two examples of team composition that can be created from the clustering results. It is an example of the team structure of two competing teams in the 2021 season using both first and second-stage clusters of players. Future research can build on this by looking into team composition and game style properties over time using these cluster definitions.

Previous women's AF literature has identified multiple variables as being important in match-play success. These include disposals, disposal efficiency, contested and uncontested possessions, marks, marks inside 50, contested marks, inside 50s as well as contributions from key players in a team.^{5,6,8} While the aim of this research differs in objective to these precedents of Black et al.,⁵ Cust et al.,⁶ and Dwyer et al.,⁸ overlap can be seen in important variables. The importance of variables of disposals, kicks, disposal efficiency/effective disposals, and contested possession were still evident in determining clusters through the methodology applied in this research even when accounting for the additional locational component in this analysis.

Other variables that modelling deemed important in determining either first-stage positions (forward, midfield, defender, or ruck) or second-stage roles within the position in the present study included clearances, intercepts, rebound 50s, and hitouts, which were not previously found to be important in previous research into AFLW team match-play success.^{5,6,8} This validates the methodological decision to not perform variable selection based on match-play success as key differentiators of positions are not necessarily the same variables. Therefore, this is a method to consider when performing more exploratory research, whereas a different dataset or treatment of the data may be needed in research to determine match outcomes. With that in mind, key variables in match-play success still being present suggest that some important actions performed by players in each position are similar to those required for team success. This indicates that our analysis shares some similarities in its results and conclusions with previous research identifying key technical variables

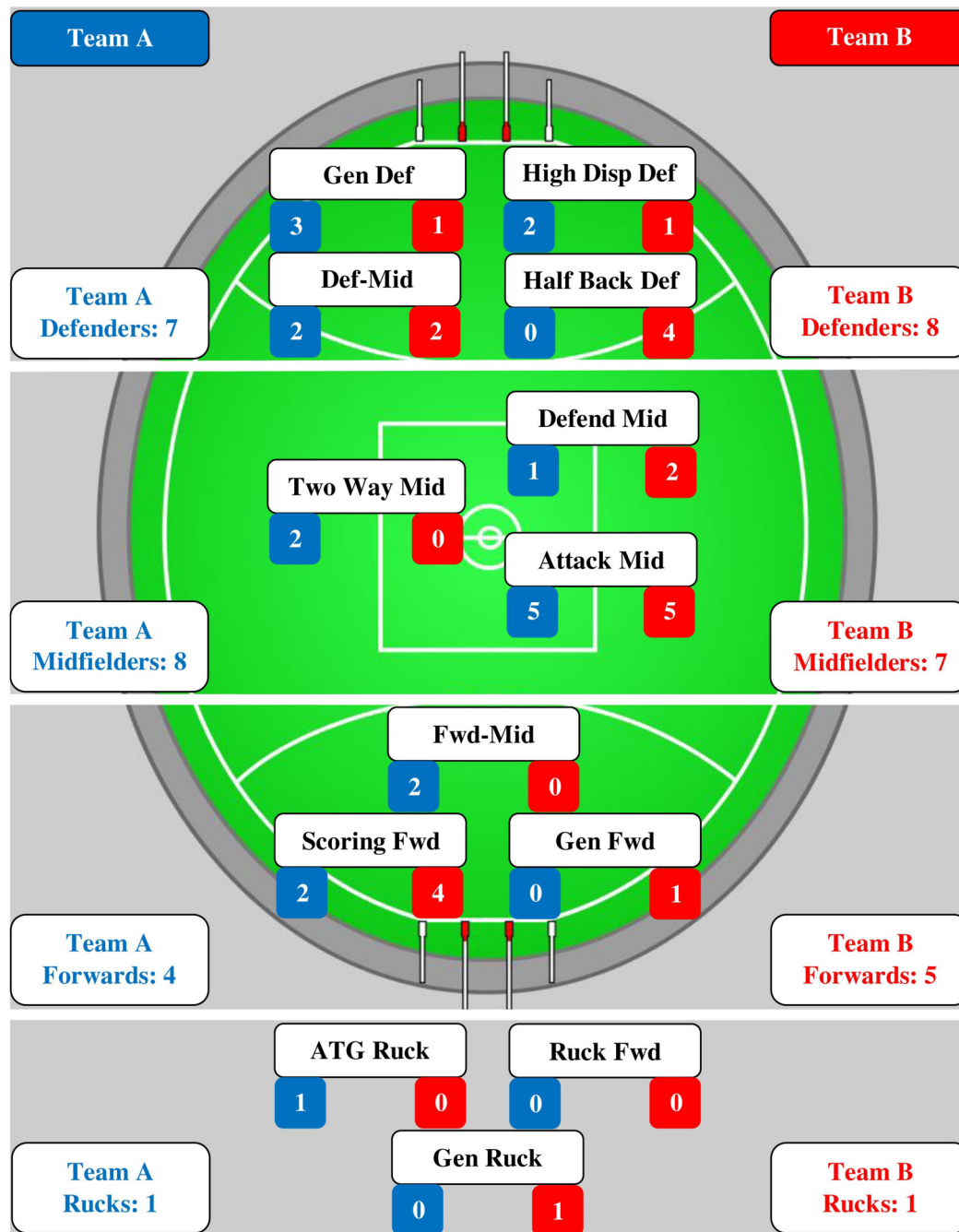


Figure 3. Example team structures of both first and second-stage cluster results. Defenders, midfielders, forwards, and rucks represent first-stage cluster positions for teams A and B on the far left and right, respectively. Internal labels represent the number of players clustered into each role within a position in second-stage clustering for each team in this example match (Team A on the left below each positional role box and Team B on the right). Gen: General, Def: Defender, High Disp: High Disposal, Mid: Midfield, Fwd: Forward, ATG: Around the Ground. This work was built ('remixed') upon the previous open licence work available at https://commons.wikimedia.org/wiki/File:AFL_stadium.svg. Attribution: Cflm001, CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0/>, via Wikimedia Commons.

influencing match-play success. As a result, this research assists practitioners, particularly coaches and sports scientists, by finding key game actions players need to perform in match-play by position which can assist in training practices, as well as defining new roles.

Research in men's AF at the elite AFL level,¹² the overlap of elite/sub-elite,¹³ and junior levels,¹⁴ have taken the same objective of analysis of position albeit primarily in a supervised manner with a reliable existing positional label for comparison. Variables key in distinguishing

position in previous men's research include intercepts, spoils, hitouts, pre-clearance, and post-clearance disposals¹²; uncontested and contested possession, clearances, disposals, kicks, inside 50s, contested marks, and effective disposals¹⁴; and spoils and tackles in all areas of the field, along with possessions in F50/D50, intercepts in the mid-field, bounces in D50, smothers in F50, total clearances, hitouts, contested marks, and shots at goal.¹³

Derivatives of all of these important variables were also reflected in our research in the AFLW. Overall, there is a similarity to previous men's results in positional classification despite differences in match-play environments in terms of variables. Given the unconstrained nature of our methodology to pre-existing labels of position, this allowed for the discovery of more hybrid positions, particularly through second-stage clustering of roles. These new roles give a more specific definition of what actions a player performs in match-play, with previous general definitions of forwards, defenders, midfielders, and rucks, which may have obscured key statistics that characterise different roles in the same general classification. This comes at the cost of verification of results, as the validity of a semi-supervised approach is hard to check without a pre-defined label.²² In the future, there is room for subjective assessment of clusters by an expert.

Clustering, data, and methodology application

Data employed for clustering were performed using the basis of the season average per player for each statistic, despite statistics representing each match and quarter being available. This was done as initial analysis using match or quarter length data demonstrated greater variance in match-play performance which made clustering harder to achieve given the need to have consistent player clusters and that the best indicator in an unsupervised clustering is tight, low-variance clusters.^{22,25} However, it is acknowledged that using a basis of a player's season average in each variable for clustering as in our analysis can lead to outlier results, particularly in players who have moved positions throughout a season; although this still provides greater insight than the alternative approaches. There exists the potential to revisit this clustering technique using individual matches or quarters in the future as this dataset increases in size, although management of the increased variance through a time series property or otherwise may need to be investigated to keep results consistent.

Increased sources of data including physical and anthropometric measures could also be included in the future to produce better representations of positions that may not be evident here. Rucks are generally the tallest players on the field, while midfielders typically have greater running demands and associated physiological capacities.³ Re-investigating physical attributes exhibited by players in these new positional classifications should be

considered in future research, to give a more holistic perspective of match-play demands for a player in physical, technical, and tactical capacities. Doing so could also give further clarity to current results. For example, Global Positioning System (GPS) data could be used to investigate whether players classified in hybrid roles are spending more time in specific areas of the field, or whether greater running capacity allows them to get to more areas of the field while remaining in the same overall position. Unfortunately, physical performance data is collected on an individual club basis, meaning that using this in future clustering would be difficult as it would be heavily biased toward the results of the number of clubs collected.⁷

More granular data of exact locational coordinates of players or more field zones could also improve clustering. For example, some commonly acknowledged player roles, such as winger, are potentially unable to be distinguished easily without the presence of the wing location on the field in data. Further contextual information on match-play would also boost the descriptive power of clustering to better understand roles and in what phase of play players perform specific actions. This increased information in the data in the future will also cause additional difficulty in creating representative clusters due to increasing dimensionality and should consider the commentary on the methodological insights gained from this research.²²

Applying PCA on datasets was key to modelling, as the applied clustering algorithm was unable to deal with many highly correlated variables well, leading to a result not representative of match-play positional splits. The use of PCA adds difficulty to cluster interpretation, which was overcome with random forests to determine which specific variables were most important, providing an easier-to-use output that can be communicated to coaches and practitioners. Previous similar work in men's elite rugby league used discriminant functions¹⁹ to distinguish cluster variable importance, although this was done on the PCA score data, giving a list of important variables rather than specific variables as presented in this analysis. Previously principal component equations were also presented as part of the analysis. Although, in this scenario, the higher number of variables made equations less interpretable, so alternatives were opted for.

It is also noted that different clustering algorithms could have been tested, with the current use of the PAM method of clustering which has Euclidean distance as the datapoint difference method potentially missing important contextual patterns that other non-linear datapoint difference methods could discover.²⁶ Current results were found to make sense from a statistical and match-play perspective, although re-running analysis with different algorithms could lead to alternate, insightful clustering patterns arising. The potential exists for a follow-up study repeating this methodology in the future, with the addition of an easier method of interpretation that allows the visualisation of temporal change in

clusters. Use of these positional classifications in the greater analysis of team tactical phenomena can also be conducted, while the definitions provided can ensure accurate classification and understanding of the technical demands for player positions in future performance research in the AFLW competition.

Strengths and limitations

The approach used to cluster positions and roles found distinct clusters without the need for bias introduced via a priori methods given the objective to understand existing data-driven relationships which unconstrained, unsupervised learning only allows.²² This allowed for more specific roles to be identified in second-stage clustering in addition to hybrid roles which would not be able to be produced with predetermined groupings or variable selection derived from variables previously found to be influential in match-play success. While this identification of hybrid roles is a novel contribution, it is unable to comment on a player's capability when performing other roles. The number of variables and seasons within the dataset is also a strength of this analysis, with no previous analysis of the AFLW having more than three seasons worth of data or employing data from the most recent seasons.⁷ In addition, the results produced are interpretable and actionable. As the AFLW competition is still evolving, the clustering definitions may not completely hold for long as further development occurs and match-play trends change.⁷ A strength of this approach is that the methodology used can be repeated into the future to ensure that results are up to date with the current trends of the AFLW.

There are some inherent limitations in the data and methodology used, much of which stems from the sample size issues derived from the relatively short seasons and history of the AFLW competition. Seasonal averages were utilised for both ethical clearance and data dimensionality reasons given limited computing power. In the future, more granular data (both temporally and additional variables) could be employed, which may provide a more robust definition. However, such an approach may also be more subject to issues surrounding increasingly sparse and noisy data in line with issues associated with the curse of dimensionality.^{25,27} Using only the seasonal average of players can obscure true positions in cases where players play different positions across different games in a season.

Conclusion

This study is the first to report on the technical skill positional roles within the AFLW competition using data-driven techniques, with the added benefit of a robust dataset used for classification. Two-stage clustering analysis provided a basis for classifying players into four

primary position groups which followed generally accepted definitions (defender, midfielder, forward, and ruck). A further 13 groups, each representing a role within the four primary positions, were determined, which consisted of four defensive roles, and three midfielder, forward, and ruck roles. The existence of hybrid roles defining forwards and defenders that likely spend time playing as a midfielder was also revealed. Unsupervised clustering enabled a viable approach to discover these new roles rather than being constrained to pre-defined existing positional classification as has been seen in previous literature.

Variable importance that determines the classification of a player into each cluster was also presented, suggesting technical areas of focus for each specific positional group that can be applied by coaches and performance staff from developmental to elite levels of women's AF. Data analysts could also derive benefit from the application of clustering in this environment with key data and methodological considerations to make in the future while producing interpretable, reproducible, and comparable results for future research with suggestions of how to best use available data.



Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Australian Government Research Training Program Scholarship.

ORCID iDs

Braedan van der Vegt  <https://orcid.org/0000-0001-6028-1797>
Justin Keogh  <https://orcid.org/0000-0001-9851-1068>

Supplemental material

Supplemental material for this article is available online.

References

1. McIntosh S, Kovalchik S and Robertson S. Multifactorial benchmarking of longitudinal player performance in the Australian football league. *Front Psychol* 2019; 10. DOI: 10.3389/fpsyg.2019.01283
2. Black GM, Gabbett TJ, Johnston RD, et al. Physical fitness and peak running periods during female Australian football match-play. *Sci Med Footb* 2018; 2: 246–251.
3. Clarke AC, Ryan S, Couvalias G, et al. Physical demands and technical performance in Australian Football League Women's (AFLW) competition match-play. *J Sci Med Sport* 2018; 21: 748–752.
4. Clarke AC, Whitaker M and Sullivan C. Evolving peak period, match movement, and performance demands in elite women's Australian football. *J Sci Med Sport* 2021; 24: 683–688.

5. Black GM, Gabbett TJ, Johnston RD, et al. A skill profile of the national Women's Australian Football League (AFLW). *Sci Med Footb* 2019; 3: 138–142.
6. Cust EE, Sweeting AJ, Ball K, et al. The relationship of team and individual athlete performances on match quarter outcome in elite women's Australian rules football. *J Sci Med Sport* 2019; 22: 1157–1162.
7. van der Vegt B, Gepp A, Keogh J, et al. Methods of performance analysis in women's Australian football: a scoping review. *PeerJ* 2023. <https://doi.org/10.7717/peerj.14946>
8. Dwyer DB, Di Domenico I and Young CM. Technical performance in elite women's Australian football—comparisons with men's football, identifying important performance characteristics and apparent trends. *Int J Perform Anal Sport* 2022; 22: 29–37.
9. AFL. AFLW expansion: four new clubs, no more AFL overlap. <https://www.womens.afl/features/aflw-expansion-four-new-clubs-no-more-afl-overlap> (2021).
10. Coutts AJ, Kempton T, Sullivan C, et al. Metabolic power and energetic costs of professional Australian Football match-play. *J Sci Med Sport* 2015; 18: 219–224.
11. Sullivan C, Bilsborough JC, Cianciosi M, et al. Factors affecting match performance in professional Australian football. *Int J Sport Physiol Perform* 2014; 9: 561–566.
12. McIntosh S, Kovalchik S and Robertson S. Examination of player role in the Australian Football League using match performance data. *Int J Perform Anal Sport* 2018; 18: 451–462.
13. Barake AJ, Mitchell H, Stavros C, et al. Classifying player positions in second-tier Australian football competitions using technical skill indicators. *Int J Sports Sci Coaching* 2022; 17: 73–82.
14. Woods CT, Veale J, Fransen J, et al. Classification of playing position in elite junior Australian football using technical skill indicators. *J Sports Sci* 2018; 36: 97–103.
15. Emmonds S, Heyward O and Jones B. The challenge of applying and undertaking research in female sport. *Sports Med - Open* 2019; 5. DOI: 10.1186/s40798-019-0224-x
16. Robertson S, Gupta R and McIntosh S. A method to assess the influence of individual player performance distribution on match outcome in team sports. *J Sports Sci* 2016; 34: 1893–1900.
17. Maechler M, Rousseeuw P, Struyf A, et al. *Cluster: cluster analysis basics and extensions*. 2.1.0 ed. 2019.
18. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018.
19. Wedding C, Woods CT, Sinclair WH, et al. Examining the evolution and classification of player position using performance indicators in the National Rugby League during the 2015–2019 seasons. *J Sci Med Sport* 2020; 23: 891–896.
20. O'Donoghue P. Principal components analysis in the selection of key performance indicators in sport. *Int J Perform Anal Sport* 2008; 8: 145–155.
21. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
22. James G, Witten D, Hastie T, et al. *An introduction to statistical learning*. 1 ed. New York: Springer, 2013.
23. Atkinson C and Lawson S. Inside the Game: Anticipation and movement without the ball the key to AFL's pressure forwards.
24. Black GM, Gabbett TJ, Johnston RD, et al. The influence of rotations on match running performance in female Australian football midfielders. *Int J Sport Physiol Perform* 2018; 13: 434–441.
25. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning*. New York: Springer, 2009.
26. Rajaraman A and Ullman JD. *Mining of massive datasets*, 2011, pp.1–315.
27. Bellman RE. *Adaptive control processes*. Princeton, NJ: Princeton University Press, 1961.