DOI: https://doi.org/10.4038/sljas.v25i2.8132

Sri Lankan Journal of Applied Statistics, Vol (25-2)

Analyzing the Attributes of a Center Defensive Midfielder in Football using Multivariate Analysis

I.C. Nwokike[1,*], E.N. Erumaka[1], C.C. Nwaigwe[2], and M.C. Obi[1]

[1]Department of Mathematics, Federal University of Technology, Owerri, Nigeria

[2]Department of Statistics, Federal University of Technology, Owerri, Nigeria

*Corresponding author: chukwudozienwokike@gmail.com

Received: 19th June 2024/ Revised: 20th September 2024/ Published: 9th October 2024

ABSTRACT

The relevance of center defensive midfielders (CDM) in football cannot be overemphasized. During football games, they are the interlocutor between defense and attack and can be tasked with huge responsibilities. Given the high cost of acquiring an efficient and readily available CDM, we designed this study to use multivariate analysis techniques to determine the correlation between attributes (variables) of a center defensive midfielder such as crossing, short passing, volleys, dribbling, long passing, ball control, and others. With the nonexistence of a multivariate analysis of CDM attributes, we have used canonical correlation analysis (CCA) to investigate these correlations. The study produced four statistically significant $(p < 0.05)$ *canonical functions (CF), that showed that attributes such as dribbling, sprint speed, agility, and stamina should be considered more when football clubs are considering a central defensive midfielder. Principal component analysis (PCA) was conducted to discover the most important variables for a CDM. The PCA generated three components. The first component was optimal and highlighted short passing, composure, ball control, interceptions, long passing, vision, and dribbling as the most relevant attributes of a CDM.*

Keywords: Sports, Soccer, FIFA, UEFA, Canonical Correlation Analysis, Principal Component Analysis

1 Introduction

In football, the center defensive midfielder, most often referred to as the ”CDM,” is a vital position that plays a key role in both defense and attack (Hamilton, 2023). The duties of the CDM are to galvanize the defense while also assisting the attack. In assisting the attack, the CDM can sometimes be mistaken for a central attacking midfielder (CAM), but the difference is that the latter has a more advanced role and defensive duties are not so much an obligation. This can also depend on the technical formation (style of play) of the team. The CDM players that are currently most rated are Rodrigo Hernandez (Manchester City), Carlos Casemiro (Manchester United), Joshua Kimmich (Bayern Munich), N’golo Kante (Al Ittihad), Aurelien Tchouameni (Real Madrid), and a few others. Rodrigo of Manchester City demonstrates that holding midfielders are key to their teams’ success, (Karen, 2023). The holding midfield role is a bespoke one; these players need to be technically gifted, not just a destroyer of opposition attacks, but they must have a high-level understanding of the game. This helps them to know when to drop into the backline and also break forward. This also depends on the teams’ style of play (Karen, 2023; Wright, 2023). A CDM is expected to be able to effectively pass the ball to support the team; this can be in the form of initiating a counterattack (Hamilton, 2023; Wright, 2023; Karen, 2023). This is to say that they need to be relentless pressers, able to read every situation imaginable, but also talented on the ball, capable of intricately stroking the ball around in the form of short or long passes, which are the unquestionable qualities of a center defensive midfielder. The best defensive midfielders in the world are the ball-winners, anchormen, tough tacklers, and water carriers who move the needle (Mewis, 2024; Wright, 2023).

It is obvious that these players don’t come cheap these days, as can be seen in the recent acquisitions of Declan Rice (117.4 million euros), Moises Caicedo (117.1 million euros), and Aurelien Tchouameni (80 million euros) from West Ham, Brighton, and Monaco, respectively (Stead, 2023). It has not become difficult for football scouting agents and club owners to identify center defensive midfielders with the right qualities, considerable price tags, and necessary attributes to fit into their teams’ style of play. This is because these center defensive midfielders can sometimes find it difficult to replicate the performance for which they have been purchased (Ritchie, 2024).

The attributes of a CDM can come in several forms, and it is necessary for teams to identify the specifics they are looking for in their preferred center defensive midfielder. CDM attributes, such as crossing, short passing, volleys, dribbling, long passing, ball control, acceleration, sprint speed, agility, stamina, long shots, interceptions, vision, and composure, can have tendencies of correlation, (Pena, 2024). These can pose a major worry to scouts who are unaware of the correlation between these attributes while looking at the players performance statistics and attributes. This study is intended to identify

Fig. 1: Manchester City Vs. Real Madrid (Image credit: Totalfootballanalysis.com)

CDM attributes (variables) that are correlated, which can serve as substitutes for each other, and help the club make an informed recruiting decision.

Canonical correlation analysis (CCA) and principal component analysis (PCA) have been used in this study to analyze these attributes. These multivariate statistical methods can determine the correlations between these variables and highlight substitute variables. The study collected data (footballer's attribute rating data) from FIFA player ratings online sources. CCA is a measure of the interrelationships between sets of multiple dependent variables and multiple independent variables (Akbaş et al., 2005). The technique is best understood by considering it as an extension of multiple regression and correlation analysis (Meloun et al., 2011; Bai et al., 2003; Iweka et al., 2018; Kuss et al., 2003; Filho et al., 2022). The use of CCA has evolved over recent time (Yang et al., 2008). It reduced the size of the data effectively (Filho et al., 2022; Nayir et al., 2022; Lu et al., 2014; Akbaş et al., 2005). This is in line with our objective in this study. The principal component analysis is a multivariate data analysis that also has its roots in linear algebra. PCA uses sophisticated underlying mathematical principles to transform a number of correlated variables in a collection into a smaller number of variables called principal components or factors (Vidal, Ma, and Sastry, 2016; Richardson, 2009, Ramakrishna et al., 2024). The position of a CDM player on the field can be seen in the images below. The images in Fig 1 and 2 show the positions of Rodrigo Hernandez (Manchester City), and Carlos Casemiro (Manchester United) in a UEFA champions league match.

2 Methodology

The study is an analysis of attributes associated with a football center defensive midfield player using multivariate analysis techniques: canonical correlation analysis, and principal component analysis. CCA is also concerned with two sets of variables related to each other and how much variance in one set is common with or predictable from the other set. It is the largest corre-

Fig. 2: Manchester City Vs. Real Madrid (Image credit: Besoccer.com)

lation method that applies linear combination, and the goal is to maximize the correlation as in most other techniques. It explores the relation between two multivariant sets of variables (vectors), measured on the same individual, (Iweka and Magnus-Arewa, 2018; Kuss and Graepel, 2003, Filho and Toebe, 2022). The PCA is one of the most important results of applied linear algebra (Shlens, 2005). It can effectively be used to analyze large data sets. The vector space transform is deployed by PCA to reduce the dimensionality of large data sets. Since the number of components is reduced by using principal components, the complexity of the analysis itself is also reduced by avoiding analyzing a large number of output variables (Janićijević et al., 2022; Divya et al., 2024). These two multivariate analysis techniques are adequate for this study.

2.1 Data Sources

The data used in this study was extracted from players rating website (https://fifaratings.com, accessed on 9 May 2022). The study used both Microsoft Word and Microsoft Excel for data configuration and storage.

2.2 Procedure and Data Analysis

SPSS 25 (IBM, 2017) was used to analyze the quantitative data used in this study. The data was tested for missing points. The study also tested the data for multivariate normality, linearity, multicollinearity, and singularity before performing the analyses using canonical correlation analysis and principal component analysis techniques.

3 Assumption Test Results and Discussion

Here, we shall consider some relevant results from the multivariate assumption tests, canonical correlation analysis and principal component analysis.

Table 1: Tests of Normality

	Shapiro-Wilk		
	Statistic	df	Sig.
Crossing	.981	55	.549
Short Passing	.982	55	.584
Volleys	.959	55	.056
Dribbling	.980	55	.503
Long Passing	.977	55	.356
Ball Control	.973	55	.253
Acceleration	.978	55	.400
Sprint Speed	.943	55	.051
Agility	.976	55	.335
Stamina	.972	55	.234
Long Shots	.954	55	.054
Interceptions	.976	55	.336
Vision	.983	55	.615
Composure	.987	55	.806

3.1 Multivariate Assumption Testing

The study shall subject the dataset with respect to the proposed variables for analyzing this playing position to multivariate assumption testing. The study shall ensure that the data satisfy multivariate analysis assumptions.

3.1.1 Tests of Normality

A Shapiro-Wilk test was conducted, and from Table 1, we can see the statistics for all the variables (crossing, finishing, short passing, volleys, dribbling, stamina, etc.) in the first column, with their p-values given in the last column, and showing all of the variables having values $(p > 0.05)$. This shows that the data did not show evidence of non-conformity or violation of the assumption of normality.

The boxplot in Figure 3 shows that there are no more outliers in the data. The initial data consisted of extreme points (outliers), which, if not dealt with, may affect the outcome of this study. As recommended by the multivariate assumption on outliers, we have ensured that only points within the interquartile range (IQR) were retained. Points below the $25th$ percentile and above the $75th$ (Q_3) percentile have been removed from the dataset. As we can see, variables such as agility, stamina, ball control, short passing, vision, dribbling, etc. are within the IQR. Therefore, we infer that there are no outliers in the dataset, and the assumptions have not been violated (Ramaswamy et al., 2000; Schubert et al., 2012; Grubbs, 1969; Ruan et al., 2005; Zimek et al., 2018; Hodge et al., 2004; Barnett et al., 1994).

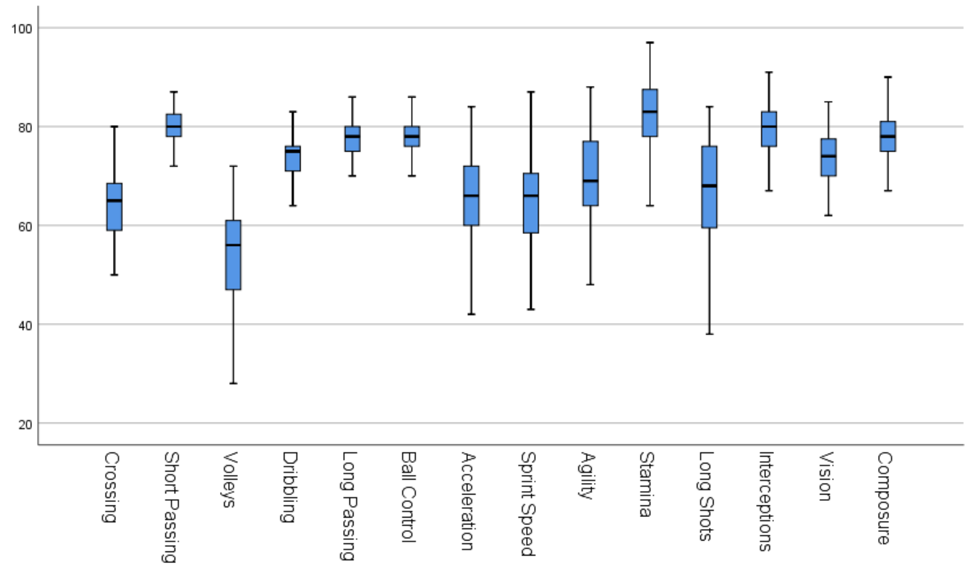

Fig. 3: Box-plot for the data outlier check

3.1.2 Test for Non-Multicollinearity

The study tested the data for multicollinearity. This occurs when the independent variables are highly correlated (Leamer, 1973; Giles, 2011; Gujarati, 2009). Hence, we have removed from the dataset one of the two variables that are correlated with values above 0.8 (Kalnins et al., 2023; Gujarati, 2009). The study shows that short passing and ball control are both significantly highly correlated with a value of 0.854, which is above 0.80. Thus, ball control has been removed from the analysis.

3.1.3 Test for Independence of Observation

Independence of observation using the Durbin-Watson test statistics was tested. The Durbin-Watson test statistic have values ranging between 0 and 4 (Durbin et al., 1950). It is important to note that a value of $0 \leq x < 2$ indicates a positive autocorrelation, 2 indicates no autocorrelation, and a value of $2 < x \leq 4$ indicates a negative autocorrelation (Durbin et al., 1951). Autocorrelation describes the correlation of a variable (data) with a delayed version of itself (Hołyst, 2017). A Durbin-Watson test statistics value of $1.5 \leq x \leq 2.5$ can be accepted to imply no autocorrelation, and independence of the observations (Dufour et al., 1985; Fahidy, 2006; King, 1983).

The test for independence of observation using the Durbin-Watson test statistics shown in Table 2 indicates that there is a negative autocorrelation between the dependent variables and the predictor variables (crossing, volleys, dribbling, long passing, ball control, acceleration, sprint speed, agility,

Table 2: Test for Independence of Observation

Model	Dependent Variables	Durbin-Watson
1	Crossing	**2.024**
2	Volleys	**2.141**
3	Dribbling	**2.211**
4	Long Passing	**2.012**
5	Ball Control	**2.047**
6	Acceleration	**2.316**
7	Sprint Speed	**2.323**
8	Agility	**2.231**
9	Stamina	**2.021**
10	Long Shots	**2.222**
11	Interceptions	**2.143**
12	Vision	**2.042**
13	Composure	**2.324**

stamina, long shots, interceptions, vision, and composure). Therefore, we infer that the assumption of independence of observations was not violated.

3.1.4 Test for Homoscedasticity

The study has tested the variance of the residuals for constancy, by plotting the standardized residuals against the predicted values and visualize the plot using a histogram, line graph, and a scatterplot (Angrist et al., 2009; Long et al., 1993; White, 1980; Engle, 1982; Gujarati et al., 2009).

Homoscedasticity test was conducted by plotting histograms of standardized residual. Each of the variables were tested independently to ensure that the assumption is not violated. The variables (crossing, volleys, dribbling, long passing, ball control, acceleration, sprint speed, agility, stamina, long shots, interceptions, vision, and composure) were plotted and investigated. The study visually inspected the histograms, which revealed that the data contained approximately normally distributed errors, which is corroborated in the normal P-P plot of the standardized residuals of the respective variables.

The P-P plots showed that these variables have an element of deviation from normality at some point along the line. Some points can be seen to not be exactly on the line but very close to it. We infer that these negligible deviations, which are expected, are also inconsequential, and we have chosen to neglect these slight deviations. We infer that the error terms of the residuals follow a normal distribution.

The scatterplots showed that the data of the variables are evenly distributed from left to right without any major clutter. This affirms the claim that the data for the various variables are normally distributed. After these investigations (histogram plots, P-P plots, and scatterplots) of data for these variables, the study, hence, infers that the data does not violate the homoscedasticity condition.

Table 3: Canonical Correlations

	Correlation	Eigenvalue	Wilks Statistic	F	Num D.F	Denom D.F.	Sig.
1	.901	4.330	.048	5.504	36.000	159.131	.000
2	.828	2.174	.257	3.116	24.000	125.314	.000
3	.393	.183	.817	.670	14.000	88.000	.798
4	.184	.035	.966	.264	6.000	45.000	.951

Table 4: Set 1 Canonical Loadings

Variables	1	2	3	4
Crossing	-.287	.399	.470	.399
Volleys	-.133	.419	.214	-.048
Dribbling	-.449	.527	-.264	-.230
Sprint Speed	-.824	-.251	-.013	-.057
Agility	-.800	-.015	.151	.057
Stamina	-.355	.419	-.341	.293
Long Shots	-.256	.415	.420	-.530
Vision	-.129	.812	.301	-.012
Composure	-.104	.837	-.142	.144

4 CCA and PCA Results

The study shall analyze the central defensive midfield data using both the canonical correlation analysis and the principal component analysis.

4.1 Canonical Correlation Analysis

Let the variables be crossing, volleys, dribbling, long passing, short passing, ball control, acceleration, sprint speed, agility, stamina, long shots, interceptions, vision, and composure.

Table 3 shows the four canonical functions generated and their canonical correlations when the CCA was conducted using the variables long passing, acceleration, interceptions, and short passing as dependent variables, while the independent variables be crossing, volleys, dribbling, sprint speed, agility, stamina, long shots, vision, and composure. Table 3 shows the canonical functions (CF) with correlations of $0.901, 0.828, 0.393$, and 0.184. The table also indicates that the first two canonical functions are statistically significant $(p < 0.001)$. With Wilks's $\lambda = 0.048$ and $F(36, 159.131) = 5.504$, canonical function 1 (CF_1) showed more significance with correlation value of 0.901. The eigenvalues $(4.330, 2.174, 0.183$, and $0.035)$ are the shared variance between the four canonical variates of dependent and independent variables.

The canonical loadings shown in Table 4 are the contributions or effects of the independent variables on the canonical functions. The table shows that four canonical functions were generated, and as seen from the second row, the

Table 5: Set 2 Canonical Loadings

Variable	1	2	3	4
Short Passing	-.249	.909	-.047	.331
Long Passing	-.104	.860	.495	-.071
Acceleration	-.989	-.149	.000	-.007
Interceptions	-.107	.801	-.389	-.442

Table 6: Proportion of Variance Explained

Canonical Variable	Set 1 by Self	Set 1 by Set 2	Set 2 by Self	Set 2 by Set 1
1	.204	.166	.265	.216
2	.265	.181	.558	.382
3	.085	.013	.100	.015
4	.068	.002	.077	.003

variables have numerical values, which plays a significant role in the ranking of the canonical functions in terms of relevance. The values of -0.287, -0.133, -0.449, and -0.824 for crossing, volleys, dribbling, and sprint speed, respectively, implies that these variables have a negative effect on the first canonical function. The table also shows that the variable, crossing, volleys, and dribbling have a positive effect on both the second canonical function with values of $0.399, 0.419$, and 0.527, respectively.

Table 5 shows the canonical loadings of the four dependent variables on the canonical functions. It can be seen that the variables short passing, long passing, acceleration, and interception have values of -0.249, -0.104, -0.989, and -0.107, respectively. This implies that they have negative effects on the first canonical function. In the second canonical functions, it can be seen that short passing, long passing, and interception all have a positive effect on the canonical function.

Table 6 shows the proportion of variance between the canonical loadings of the canonical functions in set 1 (Table 4) and set 2 (Table 5). The second and fourth columns contain the variances explained by canonical loadings of each canonical function and themselves. The third and fifth columns contain the variances of each canonical function and other canonical functions.

The diagram in Fig. 4 is a pictorial representation of the canonical correlation of canonical function 1 (CF_1) in Table 3 with respect to the independent and dependent variables shown. It also shows the corresponding canonical loadings of the given canonical functions in Tables 4 and 5. The diagram also shows the variances of the independent and dependent variables, as seen in Table 6.

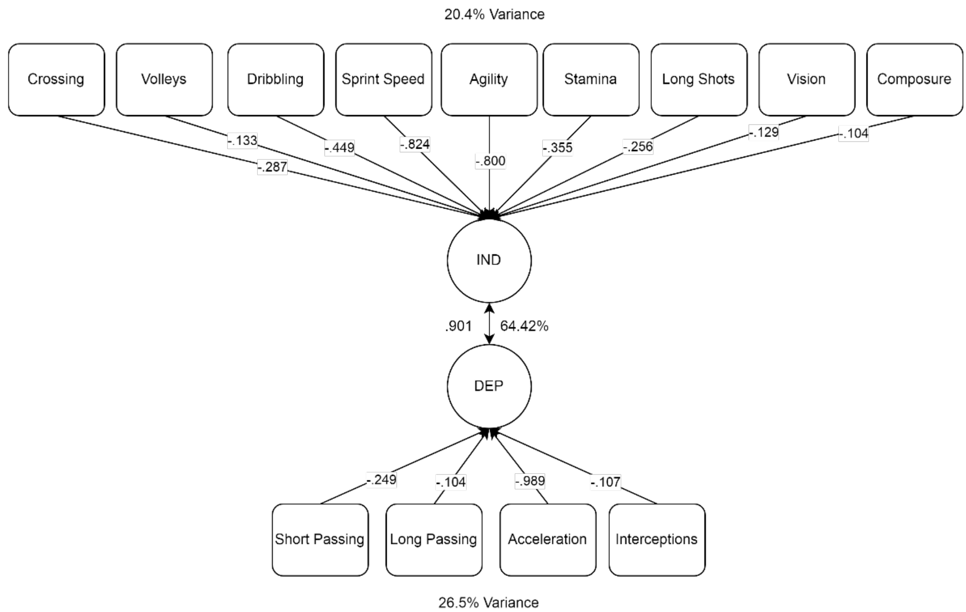

Fig. 4: CC CF_1 diagram

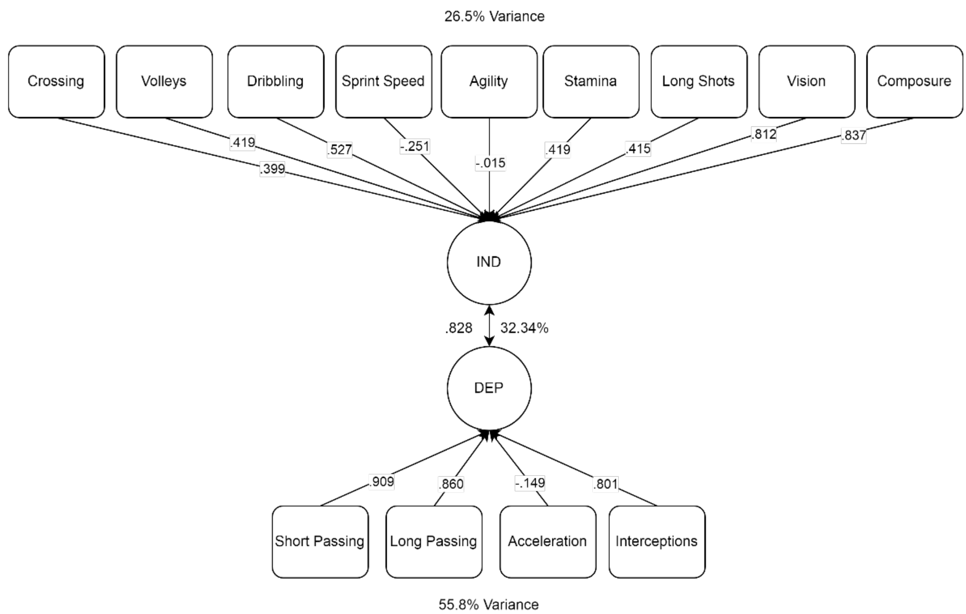

Fig. 5: CC CF_2 diagram

Table 7: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.776
Bartlett's Test of Sphericity	Approx. Chi-Square	446.655
	df	91
	Sig.	.000

Table 8: Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative
1	5.541	39.576	39.576	5.541	39.576	39.576
2	2.479	17.709	57.285	2.479	17.709	57.285
3	1.428	10.203	67.488	1.428	10.203	67.488
4	.961	6.867	74.355			
5	.705	5.038	79.392			
6	.638	4.554	83.947			
7	.524	3.741	87.687			
8	.479	3.419	91.107			
9	.345	2.466	93.573			
10	.275	1.963	95.536			
11	.202	1.445	96.981			
12	.198	1.411	98.392			
13	.129	.919	99.311			
14	.096	.689	100.000			

Extraction Method: Principal Component Analysis.
a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

The diagram in Fig. 5 is a pictorial representation of the canonical correlation of canonical function 2 (CF_2) in Table 3 with respect to the independent and dependent variables shown. It also shows the corresponding canonical loadings of the given canonical functions in Tables 4 and 5. The diagram also shows the variances of the independent and dependent variables, as seen in Table 6.

The study shall only interpret the first two canonical functions if the second canonical function is of considerable significance (Sherry et al., 2005).

4.2 Principal Component Analysis

The PCA was used to reduce the variables in the data for each of the playing positions. The PCA groups the variables into components called principal components (Forkman et al., 2019; Hotelling, 1993; Jolliffe, 2002; Stewart, 1993). The results of the PCA can be seen below:

Table 7 shows a Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy value of 0.776 and a significant value less than 0.0001 in the Bartlett's test, with both results indicating that the data are suitable for the principal component analysis (Kaiser et al., 1974; Cureton et al., 2013; IBM, 2024; Dziuban et al., 1974).

Table 8 shows the total variance explained by the variables in the center defensive midfield analysis. The initial eigenvalue compartment in the

Table 9: Component matrix

	Component		
	1	2	3
Short Passing	.860	-.172	-.204
Ball Control	.856	-.113	-.167
Long Passing	.786	-.275	
Vision	.769	-.198	
Dribbling	.712	.214	-.122
Composure	.700	-.334	-.234
Interceptions	.698	-.262	-.245
Crossing	.609	.153	.497
Stamina	.448	.237	-.291
Acceleration	.250	.890	-.190
Sprint Speed		.820	-.155
Agility	.352	.716	
Volleys	.553		.694
Long Shots	.553	.184	.587

Extraction Method: PCA
a. 3 components extracted.

table contains three columns. This compartment accounts for the variances explained by the full set of initial factors. For optimality, we shall choose the components whose eigenvalues are greater than one (IBM, 2024). The extraction sums of squared loadings compartment show that the PCA has extracted three components, which can account for 67.488 percent of the variances in the analysis. This implies that the PCA has reduced the factors to three.

Table 9 shows the correlations between the variables and the three unrotated factors (components). We can see the three components generated by the PCA for the center defensive midfield analysis. The columns show the correlation between the variables and the components under consideration. The table shows that the correlation between the variable short passing and the first PCA component is 0.860.

Table 10 shows the partial correlations between the variables and the three rotated factors (components) generated by the PCA for the center defensive midfield analysis after 4 iterations using the Oblimin with Kaiser normalization method. This study will consider values above 0.45 (IBM, 2024; Velicer, 1976; Larsen et al., 2010; Kaiser et al., 1960; Velicer, 1976). The columns show that 10, 7, and 8 variables have been grouped into the first, second, and third components, respectively.

Table 11 shows the correlations among each of the three components and themselves. The study will only consider correlation values that are ≤ 0.4 (Taherdoost et al., 2014; IBM, 2024). The table shows that the correlation between component one and the other two components is weak since the values are seen to be 0.128, and 0.382, respectively. The correlation between components two and three is 0.154. This result also implies that the PCA is adequate for the center defensive midfield data.

Figure 6 shows the 3D component plot of the first three components extracted by the PCA for the center defensive midfield data. The axes are scaled

Table 10: Pattern matrix

	Component		
	1	2	3
Short Passing	.880		
Composure	.837	-.147	
Ball Control	.830		
Interceptions	.813		
Long Passing	.741	-.149	.208
Vision	.731		.154
Dribbling	.555	.344	.139
Stamina	.440	.381	-.117
Acceleration		.945	
Sprint Speed	-.178	.842	
Agility		.758	.146
Volleys			.905
Long Shots			.812
Crossing	.116		.733

Extraction Method: PCA
Rotation Method: Oblimin with Kaiser Normalization.
a. Rotation converged in 4 iterations.

Table 11: Component correlation matrix

Component	1	2	3
1	1.000	.128	.382
2	.128	1.000	.154
3	.382	.154	1.000

Extraction Method: PCA
Rotation Method: Oblimin with Kaiser Normalization.

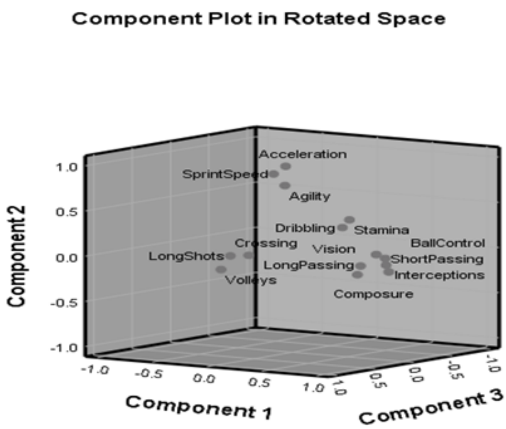

Fig. 6: 3D component plot

from -1 to 1. We can see the variables represented on the plot. Coefficients close to -1 or 1 indicate that the variable strongly influences the component, either negatively or positively (IBM, 2024; Barnett et al.,1987).

5 Discussion

The canonical correlation analysis was conducted using the variables short passing, long passing, acceleration, and interception as the dependent variables and crossing, volleys, dribbling, sprint speed, agility, stamina, long shots, vision, and composure as the independent variables to evaluate the multivariate shared relationship between the two variable sets. The CCA yielded four canonical functions(CF) with canonical correlations of $0.901, 0.828, 0.393$, and 0.184 for each successive canonical function. Only two of the canonical functions (CF_1) and (CF_2) are statistically significant $(p < 0.001)$, but canonical function 1 (CF_1) showed more significance if we consider Wilks's $= 0.048$ and $F(36, 159.131) = 5.504$, as shown in Table 3. The variance between the dependent and independent canonical variates for CF_1 is 64.42% (Table 6).

The canonical loadings (Tables 4 and 5), which are the values contributed to the canonical functions by each of the variables in the independent set (set 1) and dependent set (set 2), have values of $-0.287, -0.133, -0.449, -0.824, -0.800, -0.355, -0.256, -0.129$, and -0.104, respectively, for the control variables. The variance of these nine variables is 20.4%. The dependent variables in Table 5 (set 2) contain the canonical loadings of the dependent variables on CF_1 as $-0.249, -0.104, -0.989$, and -0.107. Also, from Table 6, column 2, the variance of these four variables is 26.5%. We can see from above that the highest contributors to the canonical function CF_1 are sprint speed, and agility, with values of -0.824, and -0.800, respectively. While in the response set (set 2), we can see that acceleration is most affected, with a response coefficient of -0.989. This is not to say that the other predictor and response variables are inconsequential.

A quick look at CF_2 shows that the independent variables have values of $0.399, 0.419, 0.527, -0.251, -0.015, 0.419, 0.415, 0.812$, and 0.837, respectively. The CF_2 dependent variables have coefficients of $0.909, 0.860, -0.149$, and 0.801. The variance of these nine control variables is 26.5%, while the variance of these three response variables is 55.8%. We can see that the highest control variable coefficients of CF_2 are vision and composure, with values of 0.812 and, -0.837, respectively. The most affected in the response were set short passing, acceleration, and interception $(0.909, 0.860$ and $0.801)$. The variance between the dependent and independent canonical variates is $32.\%$. The study shall not consider CF_3 and CF_4 because the variance between the dependent and independent canonical variates is low.

The PCA was conducted using the variables crossing, short passing, volleys, dribbling, long passing, ball control, acceleration, agility, stamina, vision, sprint speed, and composure. The PCA reduced the factors to three, showing the results for the three retained components (factors). The component correlation matrix (Table 11) showed that the three extracted components have weak correlations between themselves. Table 11 shows the pattern in which the variables correlate with the three rotated factors (components) generated by the PCA for the center defensive midfield data. We shall be considering variables with correlation values above 0.45, which implies that we shall accept only the first, second, and third components. The optimal variables in the first component are short passing, composure, ball control, interceptions, long passing, vision, and dribbling, with stamina falling short with a value of 0.44. The second component will contain the variables acceleration, sprint speed, and agility. The third component will contain variables such as volleys, long shots, and crossing. We infer that after 4 iterations, the PCA grouped the variables into three components as the most relevant variables (attributes) for a center defensive midfield player.

6 Conclusion

The CCA showed that the first and second canonical functions were the most significant. The study showed in the first canonical function, that attributes such as dribbling, sprint speed, agility, and stamina should be considered more when considering a central defensive midfielder, especially when acceleration is also a desire. This is because these attributes have a significant influence on acceleration. The study also showed in the second canonical function, that crossing, volleys, dribbling, stamina, long shots, vision, and composure should be considered when selecting a central defensive midfielder, as these attributes have huge influences on performance attributes such as short passing, long passing, and interception. The canonical function is more realistic and considers more attributes associated with a central defensive midfielder, which is therefore more acceptable. The PCA produced three principal factors. The results showed that the three extracted components have weak correlations between themselves. The first component can be considered optimal, with variables such as short passing, composure, ball control, interceptions, long passing, vision, and dribbling considered by the analysis as the most relevant attributes of a center defensive midfielder.

Conflicts of Interest: The authors declare no conflicts of interest.

7 References

Akbaş, Y., & Takma, Ç. (2005). Canonical correlation analysis for studying the relationship between egg production traits and body weight, egg weight and age at sexual maturity in layers. *Czech Journal of Animal Science*, 50(4), 163–168. https://doi.org/10.17221/4010-cjas.

Angrist, J. D., & Pischke, J. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. Doi:10.1515/9781400829828. ISBN 978-1-4008-2982-8.

Bai, Z., & Krishnaiah, P.R. (2003). *Reduction of Dimensionality*. Encyclopedia of Physical Science and Technology (Third Edition). Academic Press. Pages 55-73. ISBN 9780122274107. https://doi.org/10.1016/B0-12-227410-5/00466-X.

Barnett, T. P., & Preisendorfer R. (1987). Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. Monthly Weather Review. https://doi.org/10.1175%2F1520

Barnett, V., & Lewis, T. (1994), *Outliers in Statistical Data* (3 ed.), Wiley, ISBN 978-0-471-93094-5.

Cureton, E. E., & d'Agostino, R. B. (2013). Factor Analysis. doi:10.4324/9781315799476.

Divya, T., & Praveen, L. (2024). A method for human behavior identification based on integrated sensor data using XGBoost classifier with PCA techniques. *Physica Scripta*. https://doi.org/10.1088/1402-4896/ad328c

Dufour, J.-M., & Dagenais, M. G. (1985). Durbin-Watson tests for serial correlation in regressions with missing observations, *Journal of Econometrics*, Volume 27, Issue 3, 1985, Pages 371-381, ISSN 0304-4076, https://doi.org/10.1016/0304-4076(85)90012-0.

Durbin, J., & Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression, I. *Biometrika*. 37 (3–4): 409–428. Doi:10.1093/biomet/37.3-4.409.

Durbin, J., & Watson, G. S. (1951). Testing for Serial Correlation in Least Squares Regression, II. *Biometrika*. 38 (1–2): 159–179. Doi:10.1093/biomet/38.1-2.159.

Dziuban, C. D., & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin*. 81 (6): 358–361. doi:10.1037/h0036316.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*. 50 (4): 987–1007. Doi:10.2307/1912773. ISSN 0012-9682. Peter Kennedy, A Guide to Econometrics, 5th edition, p. 137.

Fahidy, T. Z. (2006). An application of Durbin–Watson statistics to electrochemical science, *Electrochimica Acta*, Volume 51, Issue 17, 2006, Pages 3516-3520, ISSN 0013-4686, https://doi.org/10.1016/j.electacta.2005.09.046.

Filho, A. C., & Toebe, M. (2022). Sample size for canonical correlation analysis in corn. *Bragantia*, 81(2017), 1-14. https://doi.org/10.1590/1678-4499.20210335.

Forkman J., Josse, J. & Piepho, H. P. (2019). Hypothesis tests for principal component analysis when variables are standardized. *Journal of Agricultural, Biological, and Environmental Statistics*. 24 (2): 289–308. doi:10.1007/s13253-019-00355-5.

Giles, D. (2011). Econometrics Beat: Dave Giles' Blog: Micronumerosity. *Econometrics Beat*. Retrieved 3 September 2023.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*. 11 (1): 1–21. An outlying observation, or "outlier," is one that appears to deviate markedly from other members of the sample in which it occurs. doi:10.1080/00401706.1969.10490657.

Gujarati, D. (2009). Multicollinearity: what happens if the regressors are correlated? Basic Econometrics (4th ed.). *McGraw−Hill*.
Pp. 363. ISBN 9780073375779.

Gujarati, D. N. & Porter, D. C. (2009). Basic Econometrics (5th ed.). Boston: *McGraw-Hill* Irwin. ISBN 978-0-07-337577-9.

Hamilton A. (2023). The Complete Guide to Soccer Positions: Center Defensive midfield (CDM). https://www.teamplayr.us/post/the-complete-guide-to-soccer-positions-cdm

Hodge, V. J., & Austin, J. (2004), A Survey of Outlier Detection Methodologies, *Artificial Intelligence Review*, 22 (2): 85–126, doi:10.1023/B:AIRE.0000045502.10941.a9.

Hołyst, R., Poniewierski, A., & Zhang, X. (2017). Analytical form of the autocorrelation function for the fluorescence correlation spectroscopy. *Soft Matter*. 13 (6): 1267–1275. doi:10.1039/C6SM02643E. ISSN 1744-683X. PMID 28106203.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, and 498–520.

Hotelling, H (1936). Relations between two sets of variates. Biometrika. 28 (3/4): 321–377. doi:10.2307/2333955.

http://www.brainmapping.org/NITP/PNA/Readings/pca.pdf

IBM. (2024). KMO and Bartletts Test. https://www.ibm.com/docs/en/spss-statistics/28.0.0?topic=detection-kmo-bartletts-test

IBM SPSS Statistics for Windows, version 25 (IBM Corp., Armonk, N.Y., USA)

Iweka, F., & Magnus-Arewa, A. (2018). Canonical Correlation Analysis, A Sin Quanon for Multivariant Analysis in Educational Research. *International Journal of Humanities, Social Sciences and Education*, 5(7), 116–126. `https://doi.org/10.20431/2349-0381.0507013`

Janićijević, S., Mizdraković, V., & Kljajić, M. (2022). Principal component analysis in financial data science. Advances in principal component analysis. http://dx.doi.org/10.5772/intechopen.102928

Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics. New York: *Springer-Verlag*. doi:10.1007/b98835. ISBN 978-0-387-95442-4.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*. 20 (1): 141–151. doi:10.1177/001316446002000116.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*. doi:10.1177/001316447403400115. .

Kalnins, A., & Praitis Hill, K. (2023). The VIF Score. What is it Good For? Absolutely Nothing. Organizational Research Methods. Doi:10.1177/10944281231216381. ISSN 1094-4281.

Karen C. (2023). Rodri's Manchester City role shows holding midfielders are key to success. https://www.theguardian.com/football/blog/2023/mar/30/rodri-indispensable-manchester-city-hunt-for-trophies-holding-midfielders.

King, M. L. (1983). The Durbin-Watson test for serial correlation: Bounds for regressions using monthly data, *Journal of Econometrics*, Volume 21, Issue 3, 1983, Pages 357-366, ISSN 0304-4076, https://doi.org/10.1016/0304-4076(83)90050-7.

Kuss, M., & Graepel, T. (2003). The geometry of kernel canonical correlation analysis. *Biological Cybernetics*, http://www.kyb.mpg.de/publication.html?publ=2233

Larsen, R., & Warne, R. T. (2010). Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods*. 42 (3): 871–876. doi:10.3758/BRM.42.3.871.

Leamer, E. E. (1973). Multicollinearity: A Bayesian Interpretation. *The Review of Economics and Statistics*. 55 (3): 371–380. Doi:10.2307/1927962. ISSN 0034-6535.

Long, J. S., & Trivedi, P. K. (1993). Some Specification Tests for the Linear Regression Model. In Bollen, Kenneth A.; Long, J. Scott (eds.). Testing Structural Equation Models. London: Sage. Pp. 66–110. ISBN 978-0-8039-4506-7.

Lu, Y., & Foster, D. P. (2014). Large scale canonical correlation analysis with iterative least squares. Advances in Neural Information Processing Systems.

Meloun, M. & Militký, J. (2011). *Statistical analysis of multivariate data. Statistical Data Analysis*. Woodhead Publishing India. Pages 151-403. ISBN 9780857091093. https://doi.org/10.1533/9780857097200.151.

Mewis J. (2024). Ranked! The 10 best defensive midfielders in the world. https://www.fourfourtwo.com/features/ranked-the-10-best-defensive-midfielders-in-the-world-right-now.

Nayir, F., & Saridas, G. (2022). The relationship between culturally responsive teacher roles and innovative work behavior: Canonical Correlation Analysis. *Journal of Educational Research and Practice*, 12(1), 36–50. https://doi.org/10.5590/jerap.2022.12.1.03

Pena, D. (2024). CDM Soccer Possition Explained. https://artoffootballblog.com/cdm-soccer-position/

Ramakrishna, K.K., Muddu, M., Fouzi, H., Anoop, V., & Ying, S. (2024). Robust Fault Detection in Monitoring Chemical Processes Using Multi-Scale PCA with KD Approach. Chem Engineering. https://doi.org/10.3390/chemengineering8030045

Ramaswamy, S.; Rastogi, R.; & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* – SIGMOD ’00. P. 427. Doi:10.1145/342009.335437. ISBN 1581132174.

Richardson, M. (2009). Principal component analysis.*6*(16), 4. `http://people.maths.ox.ac.uk/richardsonm/SignalProcPCA.pdf` (last access: 3.5. 2013). .

Ritchie, C. (2024, March 24). 16 worst premier league signings in football history (ranked). Givemesport. `https://www.givemesport.com/football-soccer-premier-league-worst-signings-ever/`

Ruan, D., Chen, G., & Kerre, E. (2005). Intelligent data mining: Techniques and applications. studies in computational intelligence Vol. *5. Springer*. ISBN 978-3-540-26256-5.

Schubert, E., Zimek, A., & Kriegel, H. -P. (2012). Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection. Data Mining and Knowledge Discovery. 28: 190–237. doi:10.1007/s10618-012-0300-z.

Sherry, A., & Robin K. H. (2005). Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer, *Journal of Personality Assessment*. https://doi.org/10.1207/s15327752jpa8401

Shlens, J. (2005). A Tutorial on Principal Component Analysis. http://www.brainmapping.org/NITP/PNA/Readings/pca.pdf

Stead, M. (2023). The most expensive midfielders ever: Man City new boy crashes top 20 on deadline day. `https://www.football365.com/news/most-expensive-midfielder-signings-ever-liverpool-fernandez`

Stewart, G. W. (1993). On the early history of the singular value decomposition. SIAM Review. 35 (4): 551-566. doi:10.1137/1035134.

Taherdoost, H., Sahibuddin, S., & Jalaliyoon, N. (2014). Exploratory Factor Analysis; Concepts and Theory. Advances in Applied and Pure Mathematics, 27, WSEAS, pp.375-382, 2014, *Mathematics and Computers in Science and Engineering Series*, 978-960-474-380-3.

Velicer, W.F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika. doi:10.1007/bf02293557.

Vidal, R., Ma, Y., & Sastry, S. (2016). Principal Component Analysis. doi:10.1007/978-0-387-87811-9_2.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 48 (4): 817–838. doi:10.2307/1912934.

Wright N. (2023). Declan Rice is making an instant impact and driving Arsenal forward following his £105m move from West Ham.

Yang, H. J., Jing, Z. L., & Zhao, H. T. (2008). Tensor canonical correlation analysis. Shanghai Jiaotong Daxue Xuebao/*Journal of Shanghai Jiaotong University*, 42(7), 1124–1128.

Zimek, A., & Filzmoser, P. (2018). There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.
doi:10.1002/widm.1280. ISSN 1942-4787.