



# Graph theory-based mathematical modeling and analysis to predict a football dream team

Anamaya Vyas<sup>1</sup> · Arsh Parnami<sup>2</sup> · Manas Ranjan Prusty<sup>3</sup>

Received: 9 June 2022 / Revised: 21 December 2022 / Accepted: 13 February 2023 /

Published online: 8 March 2023

© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

## Abstract

The popularity of football among fans to analyze the game has been immense with the advent of internet. The concept of making a dream team in football has become a new fashion for the football lovers. The paper focuses in helping achieving this prediction of a football dream team. The aim of this research is to assess the dynamics of a complex topological structure when prompted with random entities whose attributes are known to us. Using graph theory and vectorial distances, the dream team is evaluated on the basis of individual abilities and interplayer synergy. Instead of focusing on discrete events in a match, this framework proposes an idea in which a dream team is quantified on the basis of their positional attributes. Each player is rated in accordance to the position he is playing, which eventually helps in finding the overall team rating. The second part of this research uses graph theory to evaluate structural and topological properties of interpersonal interactions of teammates. Teammates are treated as nodes of a graph, where each edge exemplifies the strength of their interpersonal interaction. The strength of the bond depends on on-field interactions via ball passing, ball receiving and communication which depend on experience of playing together, Nationality and Club. The methodology adopted in this paper can be a formidable basis for similarly situated larger setups involving much larger intricacies. Using this framework, we can see the behavior of a hypothetical topological structure whose node attributes are known to us, thus projecting its performance as a team and individual entities.

**Keywords** Graph theory · Mathematical modelings · Football dream team · Sports analysis

---

✉ Manas Ranjan Prusty  
manas.iter144@gmail.com

<sup>1</sup> School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India

<sup>2</sup> School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India

<sup>3</sup> Centre for Cyber Physical Systems, Vellore Institute of Technology, Chennai, Tamil Nadu, India

## 1 Introduction

Football is one of the most watched and followed games all around the world. It has many complexities, tactics, players, playing styles, formations and what not. Football team performance analysis [1] is not a new concept, and it often leaves the avid watchers perplexed about the dream team which would comprise of their favorite players playing in a formation that compliments their abilities.

Dream team football is a paradigm of social group functioning and performance analysis. Team sports [2] are amalgam of individual skills and team cooperation. Individual skill can be crucial in a social group, as it is central stimuli of persistence of a group of people. Human social interactions [3] and group formation take place with the advent of people with good individual skills. People with certain attributes can be beneficiary to a dream team. The idea of dream team can be exploited in any domain wherein a selected group performance is of pertinent interest. The present paper is an attempt to model a team whose performance is interesting to be realistic in tune with the available data.

People with higher individual skills would bring higher productivity to the functioning social group. The second part is about compatibility. Sports team work as an integrated system of players. These players need to work in a system where each compliments the qualities of the other for achieving a common goal, i.e., to win the match [4]. This research concept of compatibility is pragmatic to the working of social groups as well. Despite individual characteristics, players in team sports need to work in coordination [5, 6] so that the aggregate performance improves. With team sports like football, where formation transition, counter-attacking press and different decisions are taken within a fraction of seconds, good coordination between players is very cardinal. The coordination between players can be judged on discrete parameters like club, nationalities, passing, etc. Dream team is a framework to find out how a bunch of players whose attributes are known will play as a team. This mathematical modeling can be utilized in complex dynamic systems [7, 8] predicting their efficiency and feasibility. Social groups, teams sports, and any other coordinative establishment can be judged on merits using this framework. Compatibility is a quite vague term which we quantified by the means of graph theory which ultimately helped in finding the overall team strength if they played together. This mathematical modeling paper on dream team analysis is a mere effort to propose an idea of dream team in any field of work; using concepts of graph theory and individual characteristics, we can generate a dream team which can collaborate together with much more efficiency with an increase in productivity as well.

Joaõ Ribeiro et al. [9] have proposed a framework which uses social network analyses and graph theory to evaluate team performance. They considered synergistic interpersonal process between players in competitive performance environments, rather than discrete events. Using graph theory, they evaluated structural and topological properties of interpersonal interactions of teammates. The highlight of this paper is importance of interpersonal relationship for team performance, but it misses to focus on individual skills and work rate of an individual player. Pedro Silva et al. [10] have proposed that intra-team synchronization is governed by local information, which specifies shared affordances responsible for synergy formation. In this paper after experimentation and further research, they instituted those synergies were established and dispersed rapidly as a result of the dynamic creation of informational properties. By these tests the players became faster at regulating their movements with teammates. But this paper didn't focus on the asymmetric movements among the players which can be a specific strategy. Filipe Manuel Clemente et al. [11] have proposed an approach in which network metrics are used to improve the offensive processes analysis of football teams. Using

density, heterogeneity and centralization metrics, it is portrayed that it is feasible to recognize player's intra-connection and its strength. Florian Korte et al. [12] have portrayed interplay in football, a proposed playmaker indicator that focuses on real passing sequences rather than averages over a game. Additionally, it contributes to a more comprehensive understanding of players' contribution. The framework allows for the integration of other situational variables that are relevant to football performance in addition to play outcome. Filipe Manuel Clemente et al. [13] have proposed a pilot study which insinuated a set of network methods to quantify the specific merits of football teams. The results reveal that the lateral defenders, central defenders, and midfielders are the centroid players of the team. The most independent players in a regular way during all matches analyzed were the midfielders. Thus, it is safe to say midfielders offer a dynamism to the game, making them a prominent figure on the field. HalilOnal et al. [14] found that individual sports like billiards and archery require higher mathematical thinking, and in team sports football requires second highest analytical and problem-solving skills. These outcomes can be extended in support of importance of individual players in team sports to achieve a common outcome of winning. However, the experience of the players was exempted in this study. Jason D. Vescovi et al. [15] proposed correlation analysis to find the similarities between two variables in team sports. This study also highlights the importance of agility, speed and fast reactions in sports. These abilities are analyzed using correlation coefficient to find the degree of their relation. Though correlation analysis finds the degree of relation between two variables, it fails to prove the cause of similarity.

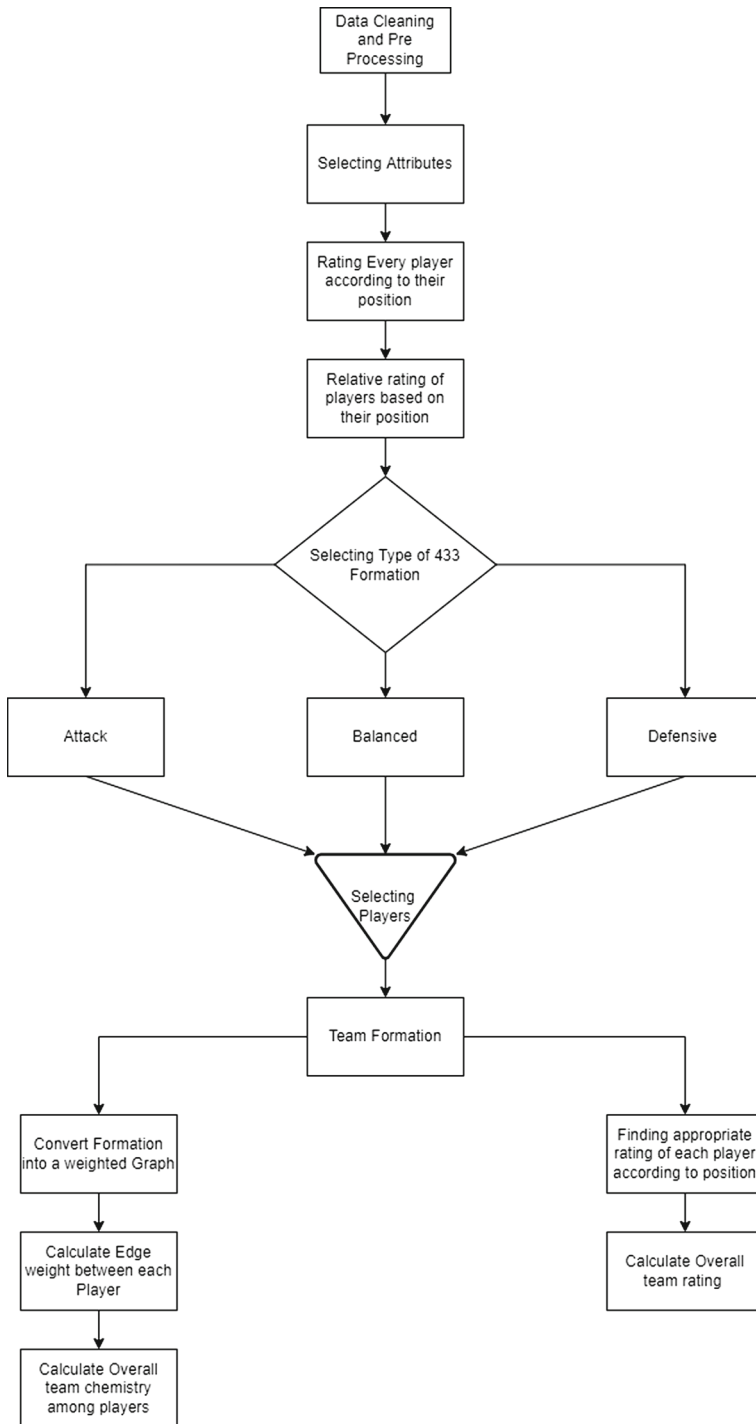
## 2 Proposed system

Figure 1 shows the block diagram of the proposed system. The proposed system starts with data cleaning and preprocessing the dataset for each playing position there on the field of football. The next step is to select attributes for each position; different positions on the field require different attributes or characteristics. Every attribute has a relevance which interposes to the overall rating of the player. After selecting the attributes, various vector distances are employed to calculate different player indices for different positions.

In the next step, the system takes the type of football formation [8]. The scope of this paper is limited to only 3 variations of 4–3–3 football field formation (4 defenders, 3 midfielders and 3 attacking forwards) [16]. The next step is to take the 11 players with the respective position at which they will play. After taking the player's names, the system is divided into two parts: the first part is about calculating the MVP (most valued player) and the team rating based on the individual abilities of every player in their respective position of play which ultimately contributes to the overall team rating. The second part is about calculating team chemistry, which is done using concepts of graph theory.

### 2.1 Dataset description

This section describes the dataset [17] used for the proposed system. The dataset used over here quantifies the properties of every player. Every player has been given 80 + attributes to judge their football skills. These encompasses attacking skills, defending skills and goalkeeping skills. There are 18,209 entries in this dataset, which means there are 18,000 + players with 80 + attributes which will ultimately help in determining the quality of each player.



**Fig. 1** Block diagram of the proposed system

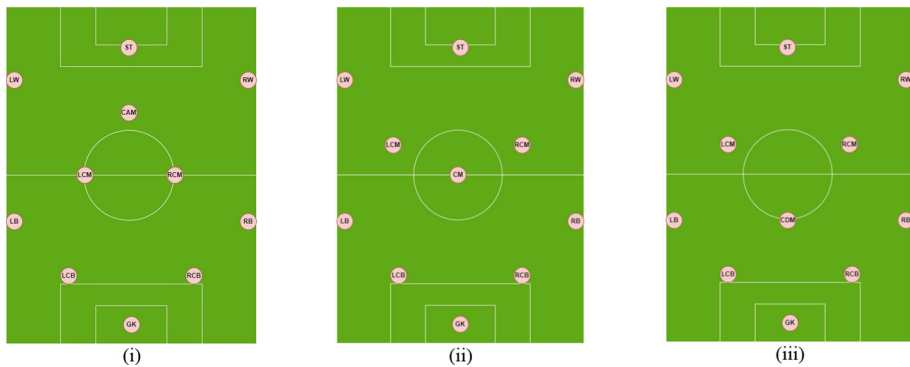
## 2.2 Football terminologies

Football as a game comes with a lot intricacy. These intricacies can be team formations, player potentials, tactics, etc. Few of these terminologies are crucial to be discussed here. The 4–3–3 football formation uses four defenders—made up of two center-backs and two full-backs—behind a midfield line of three. The most common set-up in midfield is one deeper player—the single pivot—and two slightly more advanced to either side. It is a high pressing football formation [18] in which the transitions occur as the game advances (Figs. 2 and 3).

The benefits of the 4–3–3 formation are to create natural triangles when in possession which allows several passing options to the player in possession. The player can work up the possession and can use the wingers to cut back in passing the balls with the midfielders and the overlapping full-backs to go for the goal. One of the trump cards in this formation is a ‘False-Nine’ often employed in the ideas of Pep Guardiola [19] played by Lionel Andres Messi [20]. A False-Nine is a player who is the link between the midfielders and the front attacking line. The False-Nine has the freedom to transit from the forward attacking line to the midfield to become an extra passing option. A False-Nine is a very grueling position to mark by the defenders as his position always keep transiting. This brilliant idea was used by Pep Guardiola during his early stint as a Barcelona Manager, where he became the most successful manager in the history of the club [21].

Fig. 2 4–3–3 football formation





**Fig. 3** (i) 4–3–3 Attacking (ii) 4–3–3 Balanced (iii) 4–3–3 Defensive

### 3 Methodology

#### 3.1 Player index

The first step is to give each and every player a rating. This rating would be eventually used in calculating the overall team rating. The idea behind giving each player a rating is to find out how he would play at different positions. Each position of the field has its own importance, and the requirements of every position are different. Rating every player requires parameters, and these parameters may differ from position to position. Furthermore, the number of parameters may also vary according to different positions. To have a quantification we prescribe,

$$\text{Rating}(\text{player}) = P(x_1, x_2, x_3, \dots, x_n)$$

where  $x_1, x_2$ , etc., are positional parameters derived from data set of attributes such as crossing, passing, tackling, etc. Here it is worth to note that attributes at different positions are different. For example, for a forward position tackling has not been considered as an attribute.

Here, two questions arise: first, how to select the appropriate parameters for every position and second, after selecting the parameters, how to use them to find the player index [22]. The ensuing section deals these pertinent issues.

##### 3.1.1 Parameter selection

The parameters are selected on the basis of the position the player is playing at, simply because for different positions different abilities are required. Each position has a different role, and to fulfill that role, each player should have certain qualities that are suitable for that position. To find those attributes, correlation coefficient ( $r$ ) [23] is invoked. Correlation coefficient between two variables, say  $x$  and  $y$ , indicates that if  $x$  is high and  $y$  is also high, then correlation coefficient is positive, but if one of the variables is high and the other one is low then the correlation coefficient is negative. That is,

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Having said that for every position the parameters are selected on the basis of the correlation coefficient, positive and high correlation coefficient between two parameters can be used as the degree to find the similarities of two parameters. The correlation information is then used in a proposed novel approach of average correlation coefficient (ACC). The dataset provides parameters which can be used to assess a player on every possible level. The parameters encompass goalkeeping skills, attacking skills, defending skills, dribbling skills, etc. It is worth to note that each position will use a different set of parameters depending upon the role that position has to play.

Table 1 depicts the number of attributes each position requires to access a player in that position. Furthermore, correlation coefficient matrix's color coding also depicts the similarities between attributes. If the field is greener, the correlation coefficient between two attributes is higher. On the contrary, if the field is red, the correlation coefficient between attributes is negative.

### 3.1.2 Rating system

After selecting the parameters, a rating system is generated which uses the selected parameters to generate player rating for each player with respect to their position. Each player is represented in vectorial notation wherein a component of vector describes specified attribute of the player. With each attribute now converted into a vectorial component, different method to find the magnitude of the vector can be used to find different forms of rating systems.

$$\text{Positon(player)} = \{x_1, x_2, x_3, \dots x_n\}$$

Here,  $x_1, x_2, x_3, \dots x_n$  are the n components of the player vector. Each component in turn is a parameter of that player for a particular position.

This framework uses the following rating systems:

1. Manhattan distance
2. Euclidean distance
3. Mahalanobis distance
4. Average correlation coefficient (ACC)

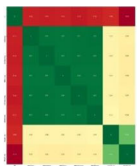
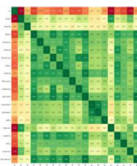
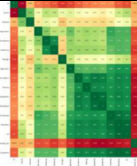
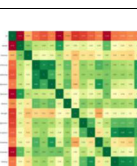
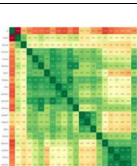
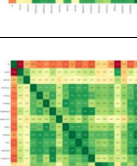
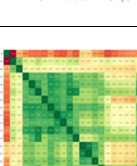
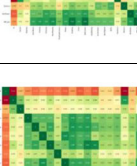
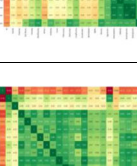
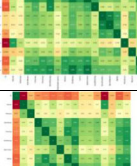
**Manhattan distance** Manhattan distance [24] is calculated as non-relative difference between 2 vectors; in other words, the sum of the absolute values is the differences of the coordinates. For instance, if  $x = (a, b)$  and  $y = (c, d)$ , the Manhattan distance  $M(a, b, c, d)$  between  $x$  and  $y$  is  $|a - c| + |b - d|$ . The framework uses the Manhattan distance as one of the rating systems.

$$\text{Manhattan}((x_1, x_2, x_3, \dots x_n)(y_1, y_2, y_3, \dots y_n)) = \sum_i^{i=n} |x_i - y_i|$$

**Euclidean distance** It is the distance between two points in Euclidean space that is represented by the length of line segment between those two points. The square root of the sum of the squares is the differences of the coordinates. For example, if  $x = (a, b)$  and  $y = (c, d)$ , the Euclidean distance  $E(x, y)$  [25] between  $x$  and  $y$  is  $\sqrt{(a - c)^2 + (b - d)^2}$ .

The framework uses the Euclidean distance as one of the rating systems. The coordinates represent the attributes in n-dimensional space, where every axis represents an attribute which is thus a component of the player vector. Euclidean distance of this provides the magnitude

**Table 1** Positional parameters and their correlation coefficients

POSITION	CORRELATION COEFFICIENT MATRIX	NUMBER OF ATTRIBUTES	POSITION	CORRELATION COEFFICIENT MATRIX	NUMBER OF ATTRIBUTES
Goalkeeper		7	Central-Attacking Midfielder		19
Center-Back		12	Central-Defensive Midfielder		16
Left-Back		21	Left Winger		19
Right-Back		21	Right Winger		19
Central-Midfielder		23	Central Striker		15

of the player vector. Higher the distance better the player.

$$\text{Euclidean}((x_1, x_2, x_3, \dots, x_n)(y_1, y_2, y_3, \dots, y_n)) = \sqrt{\sum_1^{i=n} (x_i - y_i)^2}$$

**Mahalanobis distance** The distance between two points in multivariate space is calculated with Mahalanobis distance [26]. The Euclidean representation of variables is represented by



axes which are drawn at right angle to each other. In a Euclidean plane distance between two points can be calculated using a ruler. The problem arises where the axes are correlated to each other. Here the axes are no longer perpendicular to each other. Moreover, as the dimensions increase, plotting n-dimensional coordinate system is not possible. The Mahalanobis distance solves the problem. It measures the distance between correlated points for multiple variables. The Mahalanobis distance is used to find multivariate outliers, which is a combination of two or more variables.

$$\text{Mahalanobis} = \left[ (x_B - x_A)^T \times C^{-1} \times (x_B - x_A) \right]^{0.5}$$

Here  $x_A$  and  $x_B$  is a pair of objects and  $C$  is the sample of covariance matrix.

**Average correlation coefficient (ACC)** The framework devised a novel concept average correlation coefficient or ACC. The idea behind ACC is very simple. Correlation coefficient is a measure of similarities between two attributes. If correlation coefficient is high, two variables are very similar and vice versa. The ACC is the mean strength of a parameter with other parameters. High value of ACC signifies the high similarity of a parameter with the other contributing parameters. Thus, the ACC can be called as the factor by which each parameter will contribute toward the overall rating of the player. In other terms, we can call this as the weightage of a parameter in contributing toward the overall rating of the player with respect to other parameters. High ACC means highly contributing parameter, and low ACC means less contributing parameter.

$$a_i = \frac{\sum_1^{j=i-1} C_{ji} + \sum_{i+1}^{j=n} C_{ji}}{n}$$

Here,  $a_i$  is the ACC of the  $i^{th}$  element in the array of  $n$  parameters, and  $C_{ij}$  is the correlation coefficient of  $i^{th}$  and  $j^{th}$  element.

$$\text{ACC}(x_1, x_2, x_3, \dots, x_n) = \sum_1^{i=n} a_i x_i$$

### 3.2 Relative ranking

The players are rated using multiple rating systems. But when it comes to team formation and finding overall team rating it is not possible to compare different positions. Mathematically speaking, it is impossible to compare an m-dimensional quantity with an n-dimensional object. Thus, to calculate the overall rating of the player different types of rating systems cannot be used directly as the position with larger number of attributes will always contribute more in the overall team rating. To counter this adversity, the concept of ranking is used. Every player has a rating for every possible position. The player with the higher rating will be ranked above the player with the lower one. This rank will be ultimately used in calculating the overall team rating.

Table 2 contains 10 Sects. (10 probable positions in a 4–3–3 system) having 4 distinctive types of rankings, namely Manhattan, Euclidean, Mahalanobis and average correlation coefficient (ACC). For every section (position) top 5 players and their corresponding ranking are shown.

**Table 2** Top players in respective positions

<i>GOALKEEPER</i>				
Goalkeeper name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
M. Neuer	1	1	147	1
De Gea	3	3	161	2
M. terStegen	5	7	359	3
J. Oblak	4	6	114	4
T. Courtois	2	2	48	5
<i>CENTER-BACK</i>				
Center-back name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
Sergio Ramos	1	1	1672	1
D. Godin	2	2	2361	2
M. Hummels	4	4	1488	4
T. Alderweireld	5	6	3819	5
L. Bonucci	6	8	2595	6
<i>LEFT-BACK</i>				
Left-Back name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
Alex Telles	2	4	8589	1
Marcelo	1	1	1477	2
Jordi Alba	4	3	6711	4
D. Alaba	6	10	7448	6
M. Acuna	9	19	11,173	7
<i>RIGHT-BACK</i>				
Right-Back name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
Joao Cancelo	9	11	9316	10
A. Florenzi	8	12	3529	11
Dani Alves	22	25	11,740	17
D. Srna	25	27	9067	19
K. Trippier	36	36	9289	32
<i>CENTRAL MIDFIELDER</i>				
Central Midfielder name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
K. De Bruyne	2	4	8589	1
L. Modric	3	2	7322	2

**Table 2** (continued)*CENTRAL MIDFIELDER*

Central Midfielder name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
Bruno Fernandes	14	18	10,982	12
M. Reus	13	49	11,051	14
T. Kroos	21	34	3091	21

*CENTRAL ATTACKING MIDFIELDER*

Central Attacking Midfielder name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
A. Griezmann	7	11	7316	7
R. Nainggolan	13	6	9275	10
J. Rodriguez	18	54	6288	21
Roberto Firmino	25	38	9702	25
A. Vidal	48	16	4975	40

*CENTRAL DEFENSIVE MIDFIELDER*

Central Defensive Midfielder name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
N. Kante	1	1	5289	1
Sergio Busquets	2	2	8202	2
Casemiro	4	4	6902	4
Fabinho	9	9	11,197	11
Fernandinho	14	16	11,050	14

*LEFT-WINGER*

Left-Winger name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
Neymar Jr	3	3	2275	3
E. Hazard	4	4	6568	4
P. Dybala	7	7	5925	6
L. Insigne	13	13	7423	10
Coutinho	16	15	1984	14

*RIGHT-WINGER*

Right-Winger name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
L. Messi	1	1	2477	1

**Table 2** (continued)*RIGHT-WINGER*

Right-Winger name	Manhattan rank	Euclidean rank	Mahalanobis rank	Acc rank
G. Bale	5	5	3498	5
K. Mbappe	8	7	1812	8
M. Salah	14	14	6268	16
A. Robben	24	24	10,017	20

*CENTRAL STRIKER*

Central Striker name	Manhattan rank	Euclidean rank	Mahalanobis rank	ACC rank
Cristiano Ronaldo	2	2	534	2
S. Aguero	6	6	2811	6
L. Suarez	8	8	592	9
R. Lewandoski	11	11	3527	11
P. Aubameyang	14	14	2348	14

### 3.3 Team formation

As it is discussed in Sect. 2.2, the framework has utilized 3 distinctive types of 4–3–3 formations. The 3 types of formation which this framework covers are attacking 4–3–3, defensive 4–3–3 and balanced 4–3–3. The attacking 4–3–3 has 2 central midfielders (CM) and a central attacking midfielder (CAM). Defensive 4–3–3 has 2 central midfielders (CM) and one central defensive midfielder (CDM).

With a closer look at football formations, this can be inferred that the football team formations are nothing but a graph. Each player representing a node and the connection between a players in the vicinity can be observed as the edge between two players. With the conversion of the problem into graph paradigm, it can be deduced that each connection between the nodes can have a certain value if the concepts of weighted graph is introduced. This value can be called the compatibility in sporting terms. Compatibility in itself is a very vague term. The question arises how to find that whether two players are compatible or not. Only if the edge which is called compatibility can be quantified, can concrete the idea of good understanding between two players. If the edge value between two nodes is high, then both players will have a great understanding between them. Certain criteria are required to comment on the edge value between two players. This criterion will be discussed in Sect. 3.3.2.

#### 3.3.1 Graph theory induction in 4–3–3 formations

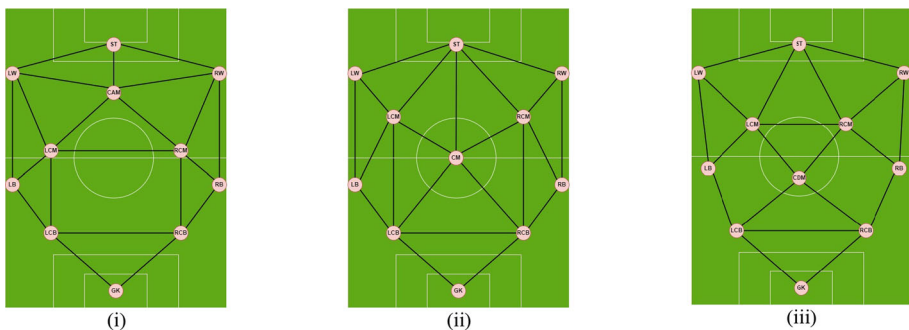
Team sports like football deploys several disciplines of graph theory. A graph  $G = (V, E)$  consists of a non-empty vertex set  $V(G)$  and a finite family  $E(G)$  of unordered pairs of elements of  $V(G)$  called edges, such that an edge  $\{v, w\}$  joins the vertices  $v$  and  $w$  [25]. Each formation in this framework has a different graph. The topology depends highly around the formation. With a defensive 4–3–3 formation [26], the defense remains highly compact and

crowded which would make it difficult for the opponent to find spaces between defensive lines to score a goal. With an attacking 4–3–3 formation [27], the central attacking midfielder plays an important role in carrying out the attack. This CAM is the link between the midfield line and the attacking line, making the team very dangerous on counter-attacking play [28], and build-up play [29] as well, but this has a drawback too. With the team attacking so well, it leaves spaces at the back which can be easily exploited once the opposition gets the ball. With a balanced 4–3–3 formation, the team has a choice to attack or defense depending on the wind of the game. This allows easy transitions and gives much more passing options to the player. There is a clear demarcation between football lines of midfield, defense and attack, midfield being the pivot of the formation which can allow the movement of the ball from back of the field to upfront, and being congested when the possession of the ball is lost, making it difficult for the opponent to attack.

Football is a game of coordination and communication. Each position works with the other one to get desired results, and every position has its own importance. As each player is important in his role, certain edges between nodes (players) cannot be prioritized over the other, because interaction of players on certain positions is negligible with the players on other positions [30]. For instance, a forward and goalkeeper rarely interact on field, because of their positions on the field. Similarly, a Left-Back will rarely have an interaction with a player on Right Wing. Thus, their weightage being infinitesimal, these edges are insignificant and can be neglected. The scope of this framework covers the compatibility between two players in vicinity.

Attacking formation should allow the players to keep the ball in opponents half as much as possible. Figure 4(i) clearly shows that with a CAM the opportunities to rotate the ball increase in the upper-half of the field. CAM is connected with 5 players in the vicinity, allowing him to pass the ball to keep the attack in progress. With this formation it is quite visible that there is ample amount of attacking options but the defense looks much stretched. If the ball possession is lost, players will have to cover greater distances to defend any possible goal scoring threat posed by the opponent.

In balanced formation there is a proper demarcation between attacking line, midfield line and defensive line. Midfield as a whole being the pivot of the formation. Midfielder are the link between attacking players and the defenders. Depending on the situation the team can transit from attack to defense providing a wide range of passing possibilities. Figure 4(ii) clearly depicts that midfielder are open to attack and defense at the same time depending on the need of the game. This type of formation is very useful in buildup play which always



**Fig. 4** (i) 4–3–3 Attacking (ii) 4–3–3 Balanced (iii) 4–3–3 Defensive

allows the player to do the one-twos with the adjacent player to build the game while always remaining in the shape.

Defensive formation is very compact in defensive lines. With a CDM playing as a pivot, he is the link which is connecting the midfield with the defense. CDM stops any possible counter-attacking opportunities and is very crucial. CDM is often responsible for providing passing options during high press by the opponent. This type of formation works out pretty well against teams which like to play high pressing game. As seen in Fig. 4(iii), the team is compact at the back with a lot of passing options but attack is very vacant. This type of formation generally attacks using counter play.

### 3.3.2 Team chemistry

In team sports, the strength of interaction between teammates can be measured using weighted graph [30]. This strength of interaction can be called as compatibility between two players. Before defining compatibility or chemistry mathematically, its literal meaning should be clear. So, compatibility is a state in which two things exist together without any conflict. In footballing terms if two players are compatible with each other, chances of error and miss communication during the game would be reduced. Compatibility is a very important factor in team sports. The players playing together should know each other quite well which will ultimately boost the game of the other. For instance, if a player has a lot of skills but is not able to communicate with the players in the vicinity, his talents won't be a use to the team. He will ultimately miss passes, goal scoring opportunities and would lose possession rather cheaply. With this it is quite clear that other than individual skills players should be compatible with each other. The question arises what dynamics can shape the compatibility between two players which can ultimately make compatibility quantified?

Factors affecting compatibility:

1. Passing
2. Ball control
3. Nationality
4. Club
5. Experience

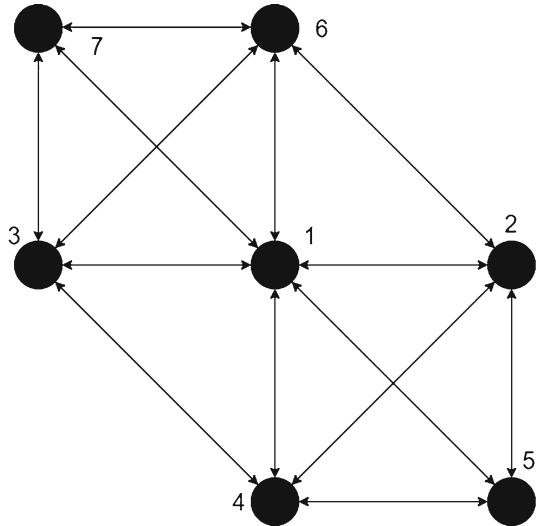
The framework has divided compatibility between two players into two parts: A) individual passing and ball receiving quotient and B) communication compatibility.

On-field interaction is a two-way communication, so while calculating compatibility between the nodes (players), it is important to consider attributes of both the nodes (Fig. 5).

### 3.4 Passing and ball control factor (PBC Factor)

Passing is one of the most important parts in the game of football. Passing the ball [31] keeps the game in continuous motion, and it is one of the most frequent ways through which players interact with each other. If passing percentage of a player is high, it means that most of the passes are completed by a player. On the other hand, if the ball control of a player is high, it means he would handle the passed ball quite nicely and would keep the game moving. There are two types of passing in football: a) long-passing (LP), which is used to pass the ball across the field or to give a lob pass to a player, and b) short-passing (SP), which is used to pass the ball to players in vicinity with less power, just to avoid losing possession and to build up the game. Both are important and thus play a decisive role in generating PBC factor.

**Fig. 5** Schematic representation of bidirectional graph representing bidirectional nature of compatibility between players



Passing is relative in nature, i.e., it depends upon the position the player is playing. Types of passes also vary according to position. Defensive players generally play short passes to avoid losing possession. Attacking players and midfielder depend on both SP and LP. The framework uses the prescribes the following way to calculate PBC,

$$PBC(\text{Defensive player}) = \text{Ball Control} + \text{SP}$$

$$PBC(\text{Attackingplayer}) = 2 \times \text{Ball Control} + \text{SP} + \text{LP}$$

Passing and ball receiving is a two-way process, to calculate the first half of compatibility or edge value between two nodes (players); both player’s PBC factor should be taken in account. We call this first half of compatibility as *C* factor. To calculate *C* factor between *P*<sub>1</sub> and *P*<sub>2</sub>, we prescribe the following formula:

$$C(P_1, P_2) = PBC(P_1) + PBC(P_2)$$

With passing of ball being relative to position, it is not appropriate to compare PBC factor of two players who play in completely different positions than the other. For instance, GK-LCB will have a different PBC factor than a RCM-LCM. To counter this problem, relative PBC factor is used. For each position minimum and maximum value of PBC is calculated. This minimum and maximum PBC is used to calculate minimum and maximum *C* factor for every possible edge in the graph. *C*<sub>min</sub> and *C*<sub>max</sub> denote the minimum and maximum *C* values of two players, respectively, in their respective positions. The interval *C*<sub>min</sub>, *C*<sub>max</sub> is further partitioned into equally space point in the interval. Every subinterval is assigned an index computed from an algorithm. Any *C* factor lying in the subinterval is accorded respective index as discussed above. Table 3 depicts the *C* factor intervals for every possible pair of position. The pair of players can be marked from 1 to 5 using these intervals, giving the first part of the compatibility factor.

**Table 3** Equally spaced interval between  $C_{min}$  and  $C_{max}$  for every probable edge between two nodes

Edge	min	T1	T2	T3	T4	max
CB-GK	70	109	148	187	226	265
CB-CB	104	148	192	236	280	324
CB-RB	202	295	388	481	574	667
CB-LB	203	296.8	390.6	484.4	578.2	672
CB-CM	258	346	434	522	610	698
CM-CM	320	404	488	572	656	740
RB-CM	264	353	442	531	620	709
LB-CM	265	354.8	444.6	534.4	624.2	714
RB-RW	267	355.2	443.4	531.6	619.8	708
LB-LW	268	357	446	535	624	713
CM-ST	299	383.4	467.8	552.2	636.6	721
CM-RW	323	406.2	489.4	572.6	655.8	739
CM-LW	325	406.4	487.8	569.2	650.6	732
ST-RW	302	385.6	469.2	552.8	636.4	720
ST-LW	304	385.8	467.6	549.4	631.2	713
CDM-CB	133	174.4	215.8	257.2	298.6	340
CDM-CM	163	203.4	243.8	284.2	324.6	365
CAM-CM	296	383	470	557	644	731
CAM-ST	275	362.4	449.8	537.2	624.6	712
CAM-LW	301	385.4	469.8	554.2	638.6	723
CAM-RW	299	385.2	471.4	557.6	643.8	730

### 3.5 Communication quotient (CQ)

In football maintaining formation, advancing, track-back, etc., are very important. Players need to maintain discipline and coordination to do this. Coordination and communication are the keys to a very well-drilled game. So, using this idea communication quotient (CQ) is generated. CQ depends upon two factors: Nationality and Club. If two players have same nationality, their CQ is higher. If two players are of same club, their CQ depends upon the experience of them playing together and players with higher experience of playing together will have high CQ than the players with less experience of playing together. Using CQ and C factor the framework rates each edge out of 10. The rating of every edge represents the synergy between the two players. Good synergy portends fluidity in game, i.e., easy passing and ball receiving between the players, good understanding of game situation and ultimately good cognizing of each other's game.

## 4 Results and discussion

Table 4 is used to check the credibility of the proposed framework. In Table 4, the actual data from the real-time matches are compared with the results of the framework. The first and the second column represents the teams which played the match and the date on which



**Table 4** Real-life match results vs framework results

Match	Match date	Winning team	Compatibility	Team rating
Liverpool vs Tottenham Hotspur	1 <sup>st</sup> June 2019	Liverpool	Tottenham	Liverpool
Barcelona vs Real Madrid	3 <sup>rd</sup> March 2019	Barcelona	Barcelona	Barcelona
PSG vs Bayern Munich	21 <sup>st</sup> July 2018	Bayern Munich	PSG	Bayern Munich
Tottenham Hotspur Vs Everton	13 <sup>th</sup> January 2018	Tottenham Hotspur	Tottenham Hotspur	Tottenham Hotspur
Manchester City vs Manchester United	11 <sup>th</sup> November 2018	Manchester City	Manchester City	Manchester City
Liverpool vs Barcelona	8 <sup>th</sup> May 2019	Liverpool	Barcelona	Barcelona
Real Madrid vs Atletico Madrid	9 <sup>th</sup> Feb 2019	Real Madrid	Real Madrid	Real Madrid
Juventus Vs Bologna	24 <sup>th</sup> Feb 2019	Juventus	Juventus	Juventus
Barcelona vs Tottenham Hotspur	4 <sup>th</sup> Oct 2018	Barcelona	Barcelona	Barcelona
FC Bayern Munich vs Manchester United	5 <sup>th</sup> August 2018	Bayern Munich	Bayern Munich	Bayern Munich
Arsenal vs Manchester City	3 <sup>rd</sup> Feb 2019	Manchester City	Manchester City	Manchester City
PSG vs Nimes	1 <sup>st</sup> September 2018	PSG	PSG	PSG
France vs Germany	16 <sup>th</sup> October 2018	France	France	France
Spain vs Croatia	11 <sup>th</sup> September 2018	Spain	Spain	Spain

the match was actually played in real life, respectively. The third column is team which won the respective match in real life. When the matches were simulated using the framework, the team compatibility and team ratings were the output and the last two columns depict the team with better compatibility and better team rating. Out of 14, 11 results showed the team with better compatibility and team rating (as predicted by the model) has actually won the game.

With the credibility of the framework that has been tested, the framework can now be tested on the players who are from different clubs and countries, and it can help in inferring how will the team dynamics will look once these players play together. In Table 5, ten different hypothetical dream teams are simulated using the proposed framework, each of them having different formations and different players playing at different positions.

Table 5 shows the lineup of ten random teams and their formation in which they will play. Table 6 depicts different team ratings which were discussed earlier and team compatibility of the simulated teams from Table 5. Table 7 shows the most important player for every team. Top 3 most valuable players (MVP) are given for every team.

Table 8 depicts the visualization for all teams is present. It portrays team formation and player rating vs positional graph for every team. This is a mere effort to show how these teams would behave if they play in future under the given circumstances with similar properties attached to them.

Table 5 Team lineups

Position	Team A	Team B	Team C	Team D	Team E	Team F	Team G	Team H	Team I	Team J
Formation	Balanced	Attacking	Defensive	Balanced	Attacking	Defensive	Balanced	Attacking	Defensive	Balanced
GK	G. Buffon	F. Forster	H. Lloris	M. Staniszewski	W. Szczesny	R. Bürki	T. Courtois	R. Sandford	D. Gerken	M. Neuer
LCB	L. Bonucci	R. Funes Mori	R. Varane	F. Mancinelli	J. Boateng	O. Solet	Piqué	J. Thompson	N. Onuoha	Sergio Ramos
RCB	Thiago Silva	G. Bellusci	N. Otamendi	J. Turley	H. Maguire	M. de Luna	Sergio Ramos	S. Old	V. Cabrera	D. Godin
LB	Alex Telles	P. van Aanholt	Alex Sandro	S. Tracy	K. Asamoah	D. Christensen	Marcos Alonso	T. Rocha	M. Pouraliganji	Alex Sandro
RB	Dani Alves	B. Jones	L. Bender	S. Al Khayri	K. Walker	K. Mbabu	D. Wass	A. Tarmin	S. Alakouch	K. Trippier
CAM	-	M. Ekiçi	-	-	Fred	-	-	M. Flynn	-	-
RCM	Coutinho	J. Gbamin	Arthur	J. Kimmich	Luiz Gustavo	N. Rigoni	J. Rodríguez	C. Butcher	D. Gorter	A. Vidal
LCM	R. Niang-golan	M. Estigarribia	David Silva	N. Dams	G. Bonaventura	M. Gradel	P. Pogba	H. Finndell	J. Martinez	M. Dembélé
CM	E. Bruno Fernandes	-	-	C. Tolisso	-	-	L. Modrić	-	-	A. Rabiot
CDM	-	-	Iniesta	-	-	Timo Triano	-	-	A. Martinelli	-
RW	G. Bale	O. Watkins	R. Mahrez	D. Loader	T. Delaney	G. Lezcano	G. Bale	H. Palacios	P. Twumasi	L. Messi
LW	E. Hazard	E. Mor	I. Piatti	Coutinho	David Silva	J. Helmer	Cristiano Ronaldo	A. Ceeseay	J. Helmer	E. Hazard
ST	S. Aguerro	D. Benedetto	E. Cavani	H. Kane	Rui Patrício	G. Hooper	K. Benzema	A. Morelos	J. Kliment	L. Suárez

**Table 6** Team statistics

Team name	Manhattan rating (out of 18,207)	Euclidean rating (out of 18,207)	Mahalanobis rating (out of 18,207)	ACC rating (out of 18,207)	Team compatibility (out of 100)
Team A	18,196.08	18,194.26	12,437.20	18,185.15	55.39
Team B	17,146.18	17,139.54	6,617.91	17,345.99	36.63
Team C	18,087.18	18,098.18	13,537.82	18,071.64	50.71
Team D	13,727.36	13,629.82	10,864.36	14,806.98	35.30
Team E	16,597.09	16,698.00	11,739.27	16,578.10	43.76
Team F	15,554.82	15,581.55	6865.09	16,394.53	34.59
Team G	18,187.18	18,182.73	13,686.45	18,176.82	55.21
Team H	10,497.18	10,511.09	9,840.90	12,211.12	28.64
Team I	14,221.45	14,134.91	5,933.18	14,725.21	32.95
Team J	18,182.18	18,191.45	12,735.36	18,169.81	54.60

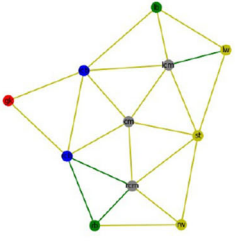
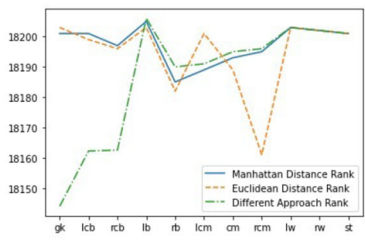
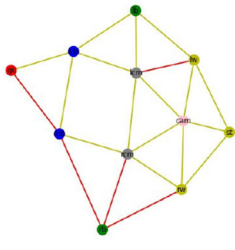
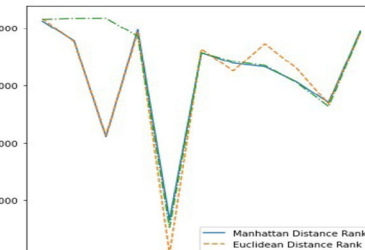
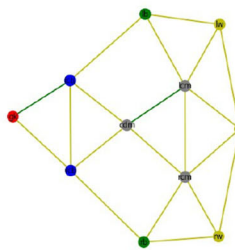
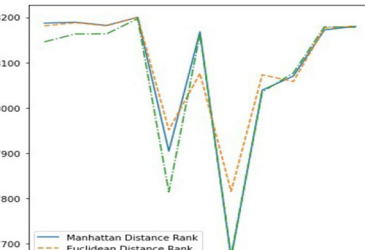
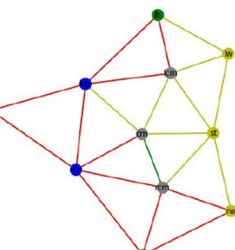
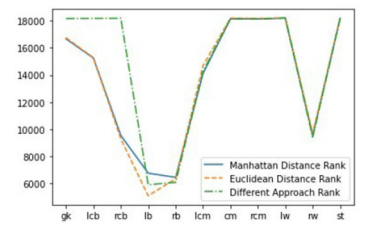
**Table 7** Most valuable players (MVP)

Team name	Player 1	Player 2	Player 3
Team A	Alex Telles	E. Hazard	G. Bale
Team B	F. Forster	P. van Aanholt	D. Benedetto
Team C	Alex Sandro	R. Varane	H. Lloris
Team D	H. Kane	Coutinho	C. Tolisso
Team E	W. Szczęśny	J. Boateng	David Silva
Team F	R. Bürki	K. Mbabu	G. Lezcano
Team G	Sergio Ramos	Cristiano Ronaldo	L. Modrić
Team H	A. Morelos	R. Sandford	H. Palacios
Team I	D. Gerken	D. Gorter	V. Cabrera
Team J	L. Messi	Sergio Ramos	M. Neuer

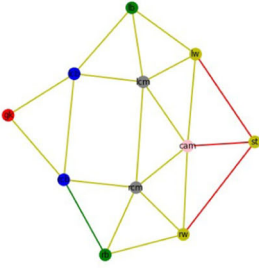
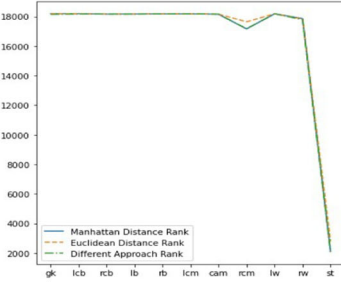
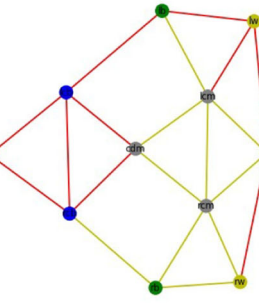
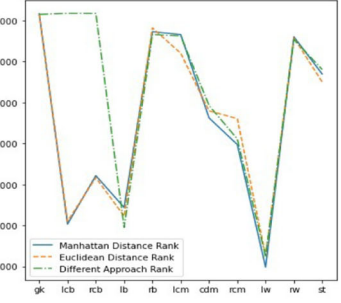
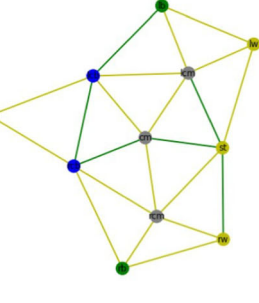
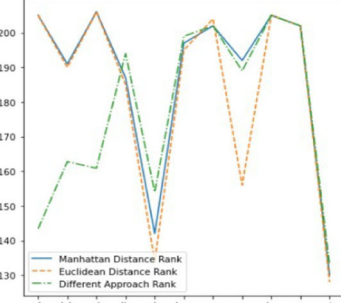
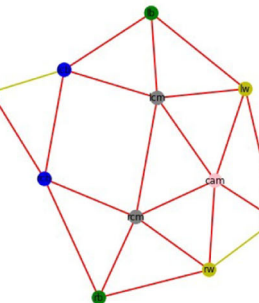

## 5 Conclusion

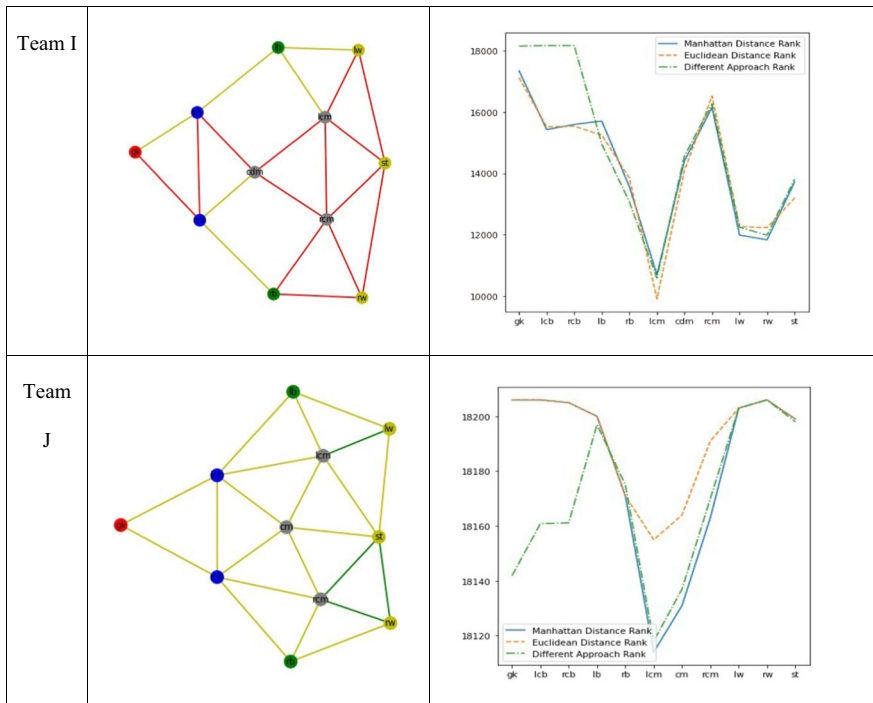
This manuscript advocates the importance of individual players in the process of team formation. Using the prescribed framework, performance analysis of a hypothetical topological structure can be done and its key players and their compatibility with each other can be evaluated. This framework gives a freedom to select different structures to find the optimum results with the given entities. Also, the suggested work can be used to check whether a given player would contribute in the betterment of the team. This analysis helps team to check whether they should select a specific player in a specific position in the match against a specific opponent. Such analysis leads to improved team selection and also allows the management to try various formations for increasing the efficiency of the team. When simulated with real time matches, out of fourteen, eleven results showed the team with better compatibility and team rating (as predicted by the model) has actually won the game. This framework can be used to

**Table 8** Team formation and team ratings

Team	Team Formation	Pictorial Representation of Position vs Rating
Team A		
Team B		
Team C		
Team D		

**Table 8** (continued)

<p>Team E</p>		
<p>Team F</p>		
<p>Team G</p>		
<p>Team H</p>		

**Table 8** (continued)

build a team with better efficiency and can be deployed by major footballing clubs to improve the efficiency of their starting 11 against different opponents. The suggested methodology helps in improving the team dynamics and also allows the management to try few variations for getting better. However, the given framework also has some limitations. The experiment did not highlight the importance of the manager on the field and also on the inter player synergies. Moreover, it also did not consider the league in which the player plays which can ultimately impact the interplayer synergy. Also, other footballing setups are overlooked and preferentially only 4–3–3 and its variations have been considered. The model and approach adopted in this paper can be applied to any field where selection process out of the given data is important. It may have further refinement with more added complexities, etc.

**Author contribution** All the authors have made substantial contributions to the conception of the work, drafted, revised and approved the version to be published.

**Funding** No financial support.

**Availability of data and material** Not applicable.

**Code availability** Not applicable.

## Declarations

**Conflict of interest** Not applicable.

## References

1. Mackenzie R, Cushion C (2013) Performance analysis in football: a critical review and implications for future research. *J Sports Sci* 31(6):639–676. <https://doi.org/10.1080/02640414.2012.746720>
2. Lago-Ballesteros J, Lago-Peñas C (2010) Performance in team sports: identifying the keys to success in soccer. *J Hum Kinet* 25:85–91. <https://doi.org/10.2478/v10078-010-0035-0>
3. Xiao X, Wu Z-C, Chou K-C (2011) A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. *PLoS ONE* 6(6):20592. <https://doi.org/10.1371/journal.pone.0020592>
4. Taylor JB, Mellalieu SD, James N, Shearer DA (2008) The influence of match location, quality of opposition, and match status on technical performance in professional association football. *J Sports Sci* 26(9):885–895. <https://doi.org/10.1080/02640410701836887>
5. Clemente FM, Martins FML, Couceiro MS, Mendes RS, Figueiredo AJ (2014) A network approach to characterize the teammates' interactions on football: a single match analysis. *Cuadernos de Psicología del Deporte* 14(3):3
6. Rawson ES, Clarkson PM, Tarnopolsky MA (2017) Perspectives on exertional rhabdomyolysis. *Sports Med* 47(Suppl 1):33–49. <https://doi.org/10.1007/s40279-017-0689-z>
7. Glazier PS (2010) Game, set and match? Substantive issues and future directions in performance analysis. *Sports Med* 40(8):625–634. <https://doi.org/10.2165/11534970-000000000-00000>
8. Nobuyoshi H (2006) Modeling tactical changes of formation in association football as a zero-sum game. *J Quant Anal Sports* 2(2):1–22
9. Ribeiro J, Silva P, Duarte R, Davids K, Garganta J (2017) Team sports performance analysed through the lens of social network theory: implications for research and practice. *Sports Med* 47(9):9. <https://doi.org/10.1007/s40279-017-0695-1>
10. Silva P, Chung D, Carvalho T, Cardoso T, Davids K, Araújo D, Garganta J (2016) Practice effects on intra-team synergies in football teams. *Hum Mov Sci* 46:39–51. <https://doi.org/10.1016/j.humov.2015.11.017>
11. Clemente FM, Couceiro MS, Martins FML, Mendes RS (2015) Using network metrics in soccer: a macro-analysis. *J Hum Kinet* 45:123–134. <https://doi.org/10.1515/hukin-2015-0013>
12. Korte F, Link D, Groll J, Lames M (2019) Play-by-play network analysis in football. *Front Psychol* 10:1738. <https://doi.org/10.3389/fpsyg.2019.01738>
13. Clemente FM, Couceiro MS, Martins FML, Mendes RS (2014) “Using network metrics to investigate football team players' connections: a pilot study. *Motriz Rev Educ Fis*. 20:262–271. <https://doi.org/10.1590/S1980-65742014000300004>
14. Onal H, Inan M, Bozkurt S (2017) A research on mathematical thinking skills: mathematical thinking skills of athletes in individual and team sports. *J Educ Train Stud* 5(9):133–139
15. “Relationships between sprinting, agility, and jump ability in female athletes - PubMed.” <https://pubmed.ncbi.nlm.nih.gov/17852692/> (accessed Jun. 02, 2022)
16. Bradley PS et al (2011) The effect of playing formation on high-intensity running and technical profiles in English FA Premier League soccer matches. *J Sports Sci* 29(8):821–830. <https://doi.org/10.1080/02640414.2011.561868>
17. “FIFA 19 player dataset.” <https://www.kaggle.com/chaitanyahivlekar/fifa-19-player-dataset> (accessed Jun. 02, 2022)
18. Carling C (2011) Influence of opposition team formation on physical and skill-related performance in a professional soccer team. *Eur J Sport Sci* 11(3):155–164. <https://doi.org/10.1080/17461391.2010.499972>
19. Alonso-Gonzalez A, Alamo-Hernandez P, Peris-Ortiz M (2017) Guardiola, Mourinho and Del Bosque: three different leadership and personal branding styles. In: *Proceedings of sports management as an emerging economic activity*, vol 1. Springer, Springer, pp 329–244. [https://doi.org/10.1007/978-3-319-63907-9\\_20](https://doi.org/10.1007/978-3-319-63907-9_20)
20. Castañer M, Barreira D, Camerino O, Anguera MT, Canton A, Hileño R (2016) “Goal scoring in soccer a polar coordinate analysis of motor skills used by lionel messi. *Front Psychol*. <https://doi.org/10.3389/fpsyg.2016.00806>
21. Ming-jun G, “The Research on Barcelona and Spain Tiki-Taka Football Style,” *undefined*, 2013, Accessed: Jun. 02, 2022. [Online]. Available: <https://www.semanticscholar.org/paper/The-Research-on-Barcelona-and-Spain-Tiki-Taka-Style-Ming-jun/fa9b8722f34c64bb2652f5cc5a46854976604a68>
22. Nesser TW, Huxel KC, Tincher JL, Okada T (2008) The relationship between core stability and performance in division I football players. *J Strength Cond Res* 22(6):1750–1754. <https://doi.org/10.1519/JSC.0b013e3181874564>
23. Taylor R (1990) Interpretation of the correlation coefficient: a basic review. *J Diagn Med Sonogr* 6(1):35–39. <https://doi.org/10.1177/875647939000600106>

24. Warner S, Bowers M, and Dixon M (2012). Team dynamics: a social network perspective. *J Sport Manag* 26(1):53-66. <https://doi.org/10.1123/jsm.26.1.53>
25. Bondy JA and Murty USR. (1976). *Graph Theory with Applications*. Elsevier, New York
26. Duarte R, Travassos B, Araújo D, and Richardson M (2013) The influence of manipulating the defensive playing method on collective synchrony of football teams,” In: *Performance Analysis of Sport IX*, Routledge
27. Leontijevic B, Jankovic A, Tomić L (2019) “Attacking performance profile of football teams in different national leagues according to Uefa rankings for club competitions. *Facta Univ, Ser Phys Educ Sport*. <https://doi.org/10.22190/FUPES180404062L>
28. Turner BJ, Sayers MGL (2010) The influence of transition speed on event outcomes in a high performance football team. *Int.J Perform Anal Sport* 10(3):207–220. <https://doi.org/10.1080/24748668.2010.11868516>
29. Casal CA, Maneiro R, Ardá T, Marí FJ, Losada JL (2022) Possession zone as a performance indicator in football the game of the best teams. *Front Psychol*. <https://doi.org/10.3389/fpsyg.2017.01176>
30. Ruohonen K, Tamminen J, Lee KC, and Piche R (2006), “Graph Theory;” Accessed: Jun. 02, 2022. [Online]. Available: <https://researchportal.tuni.fi/en/publications/graph-theory>
31. Gonçalves B, Coutinho D, Santos S, Lago-Penas C, Jiménez S, Sampaio J (2017) Exploring team passing networks and player movement dynamics in youth association football. *PLoS ONE* 12(1):0171156. <https://doi.org/10.1371/journal.pone.0171156>

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



**Anamaya Vyas** is currently pursuing his Bachelor in Technology degree at Vellore Institute of Technology (VIT), Vellore, India from the School of Computer Science and Engineering. His research interest lies in the field of mathematical modelling and machine learning.





**Arsh Parnami** is currently pursuing his Bachelor in Technology degree at Vellore Institute of Technology (VIT), Chennai, India from the School of Computer Science and Engineering. His research interest lies in the field of mathematical modelling and machine learning.



**Manas Ranjan Prusty** received his Ph.D. in Computer Science and Engineering from Homi Bhabha National Institute (HBNI) at Indira Gandhi Centre for Atomic Research (IGCAR), Kalpakkam, India in 2017. Currently, he is working as an Assistant Professor (Sr. Grade) in Vellore Institute of Technology (VIT), Chennai, Tamil Nadu, India and also, a Research Faculty at the Centre for Cyber Physical Systems at VIT Chennai. His research interest includes signal and image processing, machine learning & deep learning.