# Genetic Algorithms for Football Scouting

8 December 2021

University of Liverpool

Michael Ross, Saksham Taneja, Priyanka Mishra, Shuhei Ishiwatari, Rizwan Ahamed, Mohamed Muradh Maricair Kader Ali, Chanakya Ratnam

## Abstract

*Football clubs around the world deploy numerous methods to search for players that fit a certain criterion, but there is no proven method to perfectly predict a player's chance of success with a new club. Traditional methods of football scouting rely heavily on intuition and have resulted in mixed results due to bias of a scout or scouts' analyses of the player. This study draws inspiration from already existing evolutionary models to improve data-driven approaches to football scouting by attempting to eliminate bias and project a player's performance in a new environment. This experiment explores the Genetic Algorithm and alternatively Lotka-Volterra model for our scouting model. This method has the potential to assist clubs with a smaller budget find players capable of performing at a high level to increase parity throughout leagues around the world.*

## Introduction

Modern football, primarily represented by the English Premier League nowadays, has become a multi-faceted business model with substantial financial and sporting demands from management since the emergence of extensive commercialisation (Nesti et al, 2012). Therefore, professional football clubs focus on pursuing their domestic and international success by attempting to identify the best possible players for their team. For this reason, professional clubs have invested an enormous amount of their financial resources into developing an effective scouting system.

Typically, football clubs employ a traditional method of scouting, where football scouts travel to multiple locations, watch live matches, and create a report for the head of recruitment advising whether the club should sign the scouted player based on their observation. The traditional methods, however, have several issues, predominantly represented by a cognitive bias of the scouts. That is, each scout has an opinion on preferred style of play, tactics and individual players, often resulting in a distorted judgement of prospective players. Furthermore, limited and unreliable human memory can be another concerning factor of the traditional method. In fact, even experienced football coaches can only recall approximately 59.2% of important occurrences during the first half of a match (Laird and Waters, 2008). This is due to the phenomenon called 'highlighting', where the recall abilities are affected by the key event of a match, such as goals. As these limitations show, cognitive bias can produce an error in observation and evaluation of players and therefore the effectiveness of the traditional scouting is questionable.

In recent history, Digital video technology has made a substantial shift in the way scouts evaluate players, from the traditional approach to a data-driven approach. Since the emergence of video scouting, scouts are now able to collect tactical performance data (shots on target, possession rate, etc.) in order to objectively evaluate team and individual performance to identify their potential transfer targets (Carling et al, 2005). Hence, professional clubs are now able to conduct a holistic approach on assessing players, considering tactical, physical and physiological aspects. Although the abundance of data available and subsequent research has advanced the football industry, the application of such data is highly complex and often lacks in practicability (Memmert and Raabe, 2018). Therefore, only a fraction of football clubs could

potentially utilise such data and advanced technologies in order to identify the best possible players for the team.

Using a data-driven approach to scouting gives clubs the opportunity to analyse a player's performance and ability, but what this method lacks is the ability to project a player's performance in a new league or club. To do this, a predator vs. prey model could be used to compare attributes of a player against other players that that prospect will face in a new environment. Without a doubt, the traditional method of scouting has had many examples of success in finding an international prospect to feature in a new league (e.g., Cristiano Ronaldo from Portugal in the English Premiere League or Zlatan Ibrahimović from Sweden in the Italian Serie A); however, there are just as many, if not more, examples of players that fail to live up to expectations that oftentimes result in a significant financial loss for a club.

## Initial Approach

The idea behind the predator vs. prey approach to scouting is to view players as a predator or prey and determine the chance of success the predator or prey will have, where success is defined by the predator's ability to capture a prey or the prey's ability to evade or fend off the predator. The player's success is defined by its traits relative to another (Schmitz, 2017). In the animal kingdom, animals have attributes such as its body size, evasiveness, or other defense mechanisms that impact their chance of success against other animals that share a common territory (see Figure 1). Such is the case in football where certain positions interact more with each other to initiate a predator vs. prey engagement.

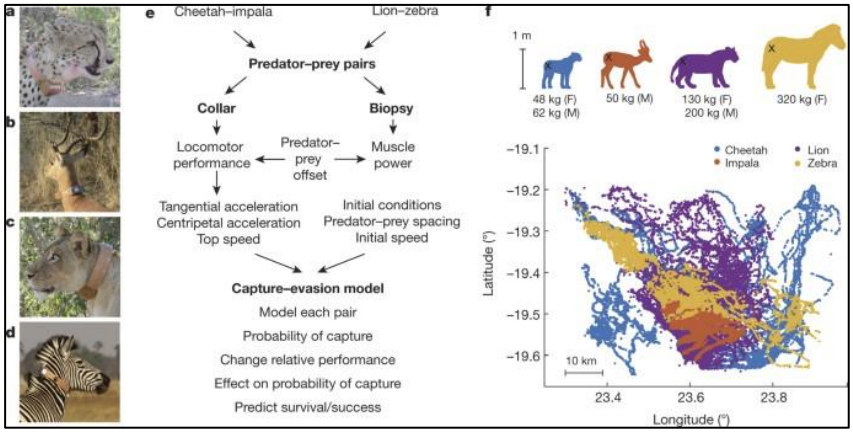


Figure 1. Predator-Prey Interaction Map (Wilson et al, 2018)

Predator vs. Prey models utilize the attributes of a predator against the attributes of a prey to predict the outcome or the winner of an engagement. By developing an algorithm to initiate the simulation of predator-prey engagements, we can then predict the outcome of an engagement between an existing predator or prey against a newly introduced animal. This same concept can then be applied to football scouting by simulating a prospective football player's performance in a new league by utilizing their attributes and traits to then use a comparison model to predict their success. This data-driven approach would then be considered as a supplement to other traditional methods of football scouting as an added insurance to increase the confidence factor when considering signing a new player.

## Lotka-Volterra Model

In order to get the best prospects from a set of data consisting of players' attributes, we initially decided to go with the predator vs prey theory based on Darwin's survival of the fittest concept.

To understand this theory, let's assume an environment with a fixed number of predators (e.g., foxes) and an excess of prey (e.g., rabbit). The fox will have an innumerable amount of food, and the number of foxes will eventually increase. This increase in predator population would lead to a decrease in the number of rabbits since more foxes mean more consumption of food (rabbit in this case). This would cause a food supply shortage, and the fox population dwindles due to that. Ultimately leading to an increase in the rabbit population and coming back to where we started. This cyclic process could be explained through the Lotka-Volterra model.

We planned to compare a football match with the animal kingdom interactions, by equating some players on the field to a predator and others to a prey.

The Lotka-Volterra model, also known as the predator-prey model, is a model used in the study of predator and prey interaction dynamics in the same environment. According to this model, the birth rate or death rate and the successful meetings between predator and prey determine the population of the predator and prey in an environment. (Restrepo and Sánchez, 2010)

The model consists of a pair of nonlinear differential equations:

$$\frac{dP}{dt} = \alpha P - \beta PD$$

$$\frac{dD}{dt} = -\gamma D + \delta PD$$

Figure 2. Lotka-Volterra Formulae

- P represents the number of prey;
- D represents the number of some predators;
- t represents time.
- α, β, γ, δ are positive real parameters describing the interaction of the two species (predator and prey)
- (dP/dt) and (dD/dt) represents the instantaneous growth rates of the two populations;

## Drawbacks of Lotka-Volterra

The model makes a few assumptions that are overly simplistic in nature and cannot be maintained in an actual (real) environment, making it imprecise (Restrepo and Sánchez, 2010). The assumptions are as follows:

- Prey has an innumerable amount of food
- Predators eat prey
- Both predators and prey have meetings that are proportional to the product between both populations

Football Analytics is a non-linear system, which makes the use of Lotka-Volterra model with its restrictive assumptions in this scenario a bit complex (Restrepo and Sánchez, 2010). The classic LV Model does not consider the random interactions, which need to be considered while implementing the model for football player analytics.

Because of these drawbacks, we looked for an alternative method and came across genetic algorithms, which is similar to Lotka-Volterra as it is based on Charles Darwin's theory of natural evolution but allows for a more focused and realistic approach. We decided not to use the Lotka-Volterra model and instead use an exploratory method like a genetic algorithm.

## Genetic Algorithm

Genetic algorithm is a heuristic optimization technique that induces variability amongst populations and creates a pool of possible outcomes to find the best solution. This algorithm is part of a larger group called evolutionary algorithms. These algorithms attempt to replicate natural evolutionary behaviors where stronger individuals with greater ease of adaptation survive, so that the final set of solutions consists of the approximated best solutions (Beasley, Bull, and Martin, 1993).

The flow chart of the algorithm we are using is outlined in Figure 3. A genetic algorithm starts with the initial population generation where a random set of solutions is formed. To select the best football players, a fitness function is used to represent the quality of the solution according to a given constraint. The players with the highest fitness function are chosen for the mating pool as the parents. There are several variation operators to apply for each pair of parents selected, such as crossover (recombination) and mutation. During crossover, a random location on the chromosome is chosen and players from the parents are swapped (Fig. 4). The resulting chromosomes are the offspring. Another variation operator is mutation where values of genes are selected and changed. During football scouting, we select one set of solutions and replace a random position for another valid player (Fig. 4). These processes cause the next generation of players to vary from the previous generation. Because only the best players from the first generation are selected for breeding, the population's average fitness will generally improve as a result of this method. Once a stop criterion is met the generational process stops. Our stop criterion for football scouting occurs when the fitness of the highest-ranking solution has reached a point where more iterations do not improve the result.

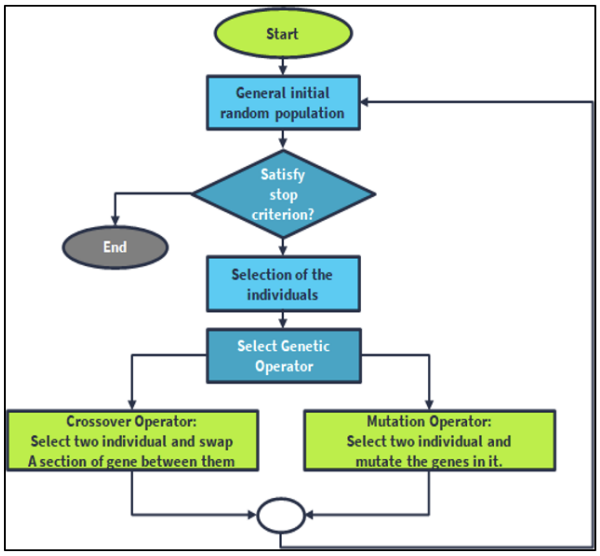


**Figure 3. Genetic Algorithm Flowchart**

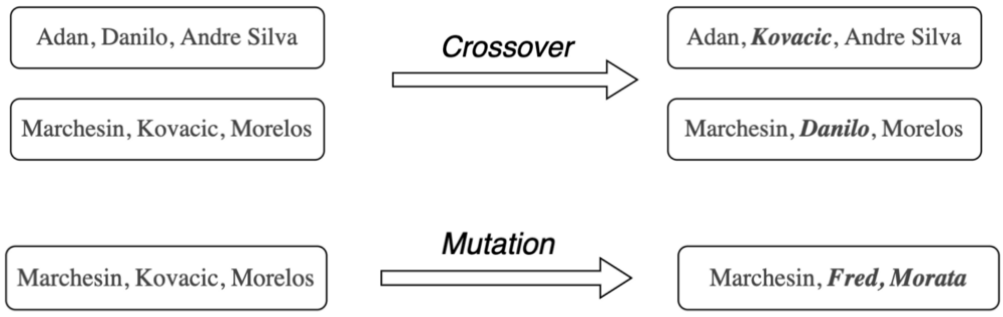


Figure 4. Crossover and Mutation operators during football scouting in the position order [goalkeeper, central midfielder, striker].

## Why Genetic Algorithm is Appropriate for Football Scouting

Football analytics datasets consist of high dimensional data with many variables, many of which could be irrelevant, noisy, or redundant. Using genetic algorithm, we can further focus on the relevant attributes and adopt the dramatic shift in 'familiar environment' into the algorithm. Additionally, genetic algorithm supports multi-objective optimization which is essential for football scouting since many variables are required to find the optimal football player.

## Breakdown of the Football Algorithm

Each player is assigned a fitness score by the fitness function, which evaluates the ability of the player to compete with other players. Player selected for the next iteration is based on its fitness score (Sarda et al., 2015).

$$Fitness\ function = \alpha(Budget - \sum_{p}^{players} value_p) + \beta\left(Avg(players_{age})\right) + \gamma(Max(players_{defending})) + \theta(Max(player_{vision})) + \delta(Max(players_{shooting}))$$

Dominant attributes such as Player's value and Player's age have been considered initially for scouting. The player's value is directly proportional to the quality that the player possesses, which helps us to ignore several other attributes available for the player's data.

However, the most skilled player can be available at a cheaper price if they are older as the player's age inversely affects their resale value. Therefore, with Player's value and Player's age attribute, a team can scout best players who can fit into the club's budget.

Fitness function has been modified to add the single most important attribute for each position to narrow down to specific players with relevant skillset for that position. The most influential characteristic of a player for a specific position is chosen as the attribute for identifying the best players.

## Experiment

The experiment was carried out in Windows-10 64-bit operating system with 2.60 GHz intel i7 processor and 8GB memory. It was also carried out in web IDE - Google Collab as it stores the data in cloud and runs much faster.

## Data

To validate our scouting model, we have used FIFA 20 players dataset which contains statistical attribute information of almost any player in the world. A sample data has been shown in Figure 5.

| sofifa_id | short_nan | age | nationalit | club | overall | potential | value_eur | wage_eur | player_po | preferred_ | pace | shooting | passing | dribbling | defending | physic | attacking_ | attacking_ | attacking_ | attacking_ | attacking_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 158023 | L. Messi | 32 | Argentina | FC Barcelo | 94 | 94 | 95500000 | 565000 | RW, CF, ST | Left | 87 | 92 | 92 | 96 | 39 | 66 | 88 | 95 | 70 | 92 | 88 |
| 20801 | Cristiano R | 34 | Portugal | Juventus | 93 | 93 | 58500000 | 405000 | ST, LW | Right | 90 | 93 | 82 | 89 | 35 | 78 | 84 | 94 | 89 | 83 | 87 |
| 190871 | Neymar Jr | 27 | Brazil | Paris Saint | 92 | 92 | 105500000 | 290000 | LW, CAM | Right | 91 | 85 | 87 | 95 | 32 | 58 | 87 | 87 | 62 | 87 | 87 |
| 200389 | J. Oblak | 26 | Slovenia | AtlÃ©tico | 91 | 93 | 77500000 | 125000 | GK | Right | | | | | | | 13 | 11 | 15 | 43 | 13 |
| 183277 | E. Hazard | 28 | Belgium | Real Madr | 91 | 91 | 90000000 | 470000 | LW, CF | Right | 91 | 83 | 86 | 94 | 35 | 66 | 81 | 84 | 61 | 89 | 83 |
| 192985 | K. De Bruy | 28 | Belgium | Manchest | 91 | 91 | 90000000 | 370000 | CAM, CM | Right | 76 | 86 | 92 | 86 | 61 | 78 | 93 | 82 | 55 | 92 | 82 |
| 192448 | M. ter Ste | 27 | Germany | FC Barcelo | 90 | 93 | 67500000 | 250000 | GK | Right | | | | | | | 18 | 14 | 11 | 61 | 14 |
| 203376 | V. van Dijk | 27 | Netherlan | Liverpool | 90 | 91 | 78000000 | 200000 | CB | Right | 77 | 60 | 70 | 71 | 90 | 86 | 53 | 52 | 86 | 78 | 45 |
| 177003 | L. ModriÄ‡ | 33 | Croatia | Real Madr | 90 | 90 | 45000000 | 340000 | CM | Right | 74 | 76 | 89 | 89 | 72 | 66 | 86 | 72 | 55 | 92 | 76 |
| 209331 | M. Salah | 27 | Egypt | Liverpool | 90 | 90 | 80500000 | 240000 | RW, ST | Left | 93 | 86 | 81 | 89 | 45 | 74 | 79 | 90 | 59 | 84 | 79 |
| 231747 | K. Mbappé | 20 | France | Paris Saint | 89 | 95 | 93500000 | 155000 | ST, RW | Right | 96 | 84 | 78 | 90 | 39 | 75 | 78 | 89 | 77 | 82 | 79 |
| 201024 | K. Koulibal | 28 | Senegal | Napoli | 89 | 91 | 67500000 | 150000 | CB | Right | 71 | 28 | 54 | 67 | 89 | 87 | 30 | 22 | 83 | 71 | 14 |
| 202126 | H. Kane | 25 | England | Tottenhan | 89 | 91 | 83000000 | 220000 | ST | Right | 70 | 91 | 79 | 81 | 47 | 83 | 75 | 94 | 86 | 81 | 85 |
| 212831 | Alisson | 26 | Brazil | Liverpool | 89 | 91 | 58000000 | 155000 | GK | Right | | | | | | | 17 | 13 | 19 | 45 | 20 |
| 193080 | De Gea | 28 | Spain | Manchest | 89 | 90 | 56000000 | 205000 | GK | Right | | | | | | | 17 | 13 | 21 | 50 | 13 |
| 215914 | N. KantÃ© | 28 | France | Chelsea | 89 | 90 | 66000000 | 235000 | CDM, CM | Right | 78 | 65 | 77 | 81 | 87 | 83 | 68 | 65 | 54 | 86 | 56 |
| 138956 | G. Chiellini | 34 | Italy | Juventus | 89 | 89 | 24500000 | 215000 | CB | Left | 68 | 46 | 58 | 60 | 90 | 82 | 54 | 33 | 83 | 65 | 45 |
| 153079 | S. AgÃ¼erc | 31 | Argentina | Manchest | 89 | 89 | 60000000 | 300000 | ST | Right | 80 | 90 | 77 | 88 | 33 | 74 | 70 | 93 | 78 | 83 | 85 |
| 155862 | Sergio Ran | 33 | Spain | Real Madr | 89 | 89 | 31500000 | 300000 | CB | Right | 72 | 68 | 75 | 73 | 87 | 85 | 66 | 63 | 92 | 80 | 69 |
| 176580 | L. SuÃ¡rez | 32 | Uruguay | FC Barcelo | 89 | 89 | 53000000 | 355000 | ST | Right | 73 | 89 | 80 | 84 | 51 | 84 | 78 | 91 | 83 | 82 | 90 |
| 188545 | R. Lewand | 30 | Poland | FC Bayern | 89 | 89 | 64500000 | 235000 | ST | Right | 77 | 87 | 74 | 85 | 41 | 82 | 62 | 88 | 85 | 82 | 88 |
| 189511 | Sergio Bus | 30 | Spain | FC Barcelo | 89 | 89 | 55000000 | 300000 | CDM, CM | Right | 42 | 62 | 80 | 80 | 85 | 80 | 62 | 67 | 68 | 89 | 44 |
| 194765 | A. Griezma | 28 | France | FC Barcelo | 89 | 89 | 69000000 | 370000 | CF, ST, LW | Left | 81 | 86 | 84 | 89 | 57 | 72 | 83 | 89 | 84 | 85 | 87 |
| 211110 | P. Dybala | 25 | Argentina | Juventus | 88 | 92 | 76500000 | 215000 | CAM, RW | Left | 83 | 82 | 84 | 90 | 43 | 64 | 82 | 80 | 64 | 87 | 88 |
| 195864 | P. Pogba | 26 | France | Manchest | 88 | 91 | 72500000 | 250000 | CM, CDM | Right | 74 | 81 | 86 | 85 | 66 | 86 | 80 | 75 | 75 | 86 | 84 |
| 210257 | Ederson | 25 | Brazil | Manchest | 88 | 91 | 54500000 | 185000 | GK | Left | | | | | | | 20 | 14 | 14 | 56 | 18 |
| 202652 | R. Sterling | 24 | England | Manchest | 88 | 90 | 73000000 | 255000 | RW, LW | Right | 93 | 79 | 78 | 89 | 45 | 57 | 78 | 83 | 38 | 84 | 67 |

Figure 5. FIFA 20 Dataset

The dataset was cleaned by removing insignificant columns, handling missing data, and standardizing values to bring in a uniform dataset. This technique ensures high quality of processed data and minimizes the risk of wrong or inaccurate conclusions.

## Configuration Parameters

There are a few parameters which can be configured for scouting players based on the team's needs. Target position must be specified so that model looks for players in that specific position. Also, Total budget a team decides to spend for buying a player in the market.

We can tune the coefficients like α, β, γ, θ, and δ which augments the importance of certain constraints. The higher the weights, the more importance it is given to the respective constraint.

- **'α'** represents the weightage factor for the first constraint**, *Budget*:** The amount of budget that needs to be used.
- **'β'** represents the weightage factor for the second constraint**, *Age*:** Younger players are preferred over older players due to the longevity of their careers.
- **'γ'** represents the weightage factor for the third constraint, ***Defending*:** How good a defender can mark/tackle the opponent in his area of coverage
- **'θ'** represents the weightage factor for the fourth constraint, ***Vision*:** Ability to pick out defense splitting passes to create goal scoring chances.
- **'δ'** represents the weightage factor for the fifth constraint**, *Finishing*:** Goals scored by a player do not reflect the number of attempts on goal to reach that tally, therefore Player's finishing attribute is a better fit.

## Experiment Results

The simulation program is executed in sequential order as per the genetic algorithm flowchart. Sample pool of candidates gets updated in each iteration based on the fitness score.

Once saturation level is achieved when the same set of players are returned in multiple successive iterations, the model returns the best set of players for the given weights in respect to their attributes. This model can be utilized by any clubs irrespective of their budget limitations

The candidate solution with the lowest fitness value over the different iterations should the considered the best solution, as it was the one that better met our restrictions.

Example 1:

- **Positions required:** [Striker, Right Winger]
- **Budget:** 70000000
- **Criteria:** Budget fulfillment: 0.5 (α)| Age: 0.2 (β) | Marking: 0 (γ) | Vision: 0 (θ) | Finishing: 0.3 (δ)

| Required Position | Name | Age | Overall | Budget | Budget Used |
|---|---|---|---|---|---|
| Striker | P. Cutrone | 21 | 77 | 70000000 | 57000000 |
| Right Winger | O. Dembélé | 22 | 84 | | |

Figure 6. Returned Results

This was the example used above – we can see that the players selected were **[P. Cutrone, O. Dembélé]**. We can observe that recommended players are very young since the weightage factor is set for age.

## Future Work

The Genetic Algorithm model can be improved drastically by creating a graph database for the real-time player's data rather than using FIFA dataset in relational database. Graph database is widely used by social media companies for better recommendations **(Kolomeets et al., 2019).** Our model can combine with state of art recommendation model to improve the quality of scouting around the world.

The FIFA dataset used is not an ideal dataset to make or evaluate predictions since the players may be subject to biased ratings by different scouts. It is, therefore, necessary to investigate other alternative datasets which consider the real-time statistical data of all the players.

## Conclusion

Genetic Algorithms used for football scouting have the potential to assist clubs at any level identify prospective players for the improvement of overall team performance. It is crucial to note that the various methods of football scouting all have their pros and cons, and that one method is not necessarily more effective than the other. There are many factors that impact whether a signing can be deemed to be successful in the future and it is simply impossible to be able to accurately predict through data or traditional methods of football scouting all attributes of the player such as their mentality or ambition - all of which are overlooked aspects that directly influence the prospect's performance. However, deploying multiple methods of football scouting can potentially provide a greater level of insurance for football clubs to increase confidence levels when deciding to sign new players. In this experiment, we drew inspiration from the genetic algorithm concept to use club-specific requirements to identify a smaller pool of candidate prospects.

## References


- Beasley, D., Bull, D.R. and Martin, R.R., 1993. An overview of genetic algorithms: Part 1, fundamentals. University computing, 15(2), pp.56-69.
- Carling, C., Williams, A.M. and Reilly, T., HANDBOOK OF SOCCER MATCH ANALYSIS: A SYSTEMATIC APPROACH TO IMPROVING PERFORMANCE.
- Kolomeets, M., Chechulin, A. and Kotenko, I.V., 2019. Social networks analysis by graph algorithms on the example of the VKontakte social network. J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl., 10(2), pp.55-75.
- Laird, P. and Waters, L., 2008. Eyewitness recollection of sport coaches. International Journal of Performance Analysis in Sport, 8(1), pp.76-84.
- Memmert, D. and Raabe, D., 2018. Data analytics in football: Positional data collection, modelling and analysis. Routledge.
- Nesti, M., Littlewood, M., O'Halloran, L., Eubank, M. and Richardson, D., 2012. Critical moments in elite premiership football: Who do you think you are?. Physical Culture and Sport, 56, p.23.
- Restrepo, J.G. and Sánchez, C.M.V., 2010, September. Parameter estimation of a predator-prey model using a genetic algorithm. In 2010 IEEE ANDESCON (pp. 1-4). IEEE.
- Sarda, V., Sakaria, P. and Deulkar, K., 2015. Football team selection using genetic algorithm. International Journal of Engineering and Technical Research, 3(2), pp.153-156.
- Schmitz, O., 2017. Predator and prey functional traits: understanding the adaptive machinery driving predator–prey interactions. F1000Research, 6.
- Wilson, A.M., Hubel, T.Y., Wilshin, S.D., Lowe, J.C., Lorenc, M., Dewhirst, O.P., Bartlam-Brooks, H.L., Diack, R., Bennitt, E., Golabek, K.A. and Woledge, R.C., 2018. Biomechanics of predator–prey arms race in lion, zebra, cheetah and impala. Nature, 554(7691), pp.183-188.