CSE 6242 Final Report: Net Value (Team 012)

Yash Bhole, Sunil Ravilla, Vamsi Ravilla, Alberto Debes, Sai Rakshith Potluri, Nikolaos Kavouras

1 Introduction

The sport of soccer is constantly evolving, bringing in significant changes in the way the game is played. A notable change is the increased importance of versatility in play styles, and the ability to contribute beyond one's base responsibility. For example, goalkeepers are no longer just expected to make saves; they also need to be good with the ball at their feet and able to participate in the team's build-up play. Similarly, midfielders and defenders are more valued if they chip in with goals and assists. The game is now more complex with a greater emphasis on versatility and multi-dimensional play styles, leading to the need for more nuanced play style evaluation. The age-old distinction - attackers, midfielders, defenders, and goalkeepers - is not enough.

2 Problem Definition

Net Value is an innovative visualization tool designed to help understand the evolution of play styles for individual soccer players. Unlike existing solutions, it offers a global overview by grouping players from 53 major competitions in 5 continents, including both league and knockout formats. This allows users to scientifically compare players across competitions, teams, seasons, and observe changes in their play styles over time. To identify nuanced play styles, we employ k-Means clustering over the player performance features, visualized in lower dimensions using t-SNE technique. The information of this analysis is displayed in a dynamic interface where users can select from a range of league-season-player combinations. The performance features are scientifically combined in a manageable number of groups to help derive insights useful to potential users. Net Value is useful for club scouts, managers, and fantasy team managers, who can use it to create their own team by diversifying across different play styles. By leveraging Net Value's visualization capabilities, users can gain insights into player performance, identify talent, and make data-driven decisions.

3 Literature Survey

Unsupervised clustering, such as k-Means [1], Mixture Models, and Hierarchical clustering [2], are commonly used to identify non-trivial patterns in social groupings where no directly quantifiable responses exist. Common applications include cohort analysis, customer segmentation, etc. In sports context, clustering is commonly used to evaluate player performance [3], identify training groups for players [1, 4, 5], and classify player cohorts [6, 7]. Louvain's algorithm for community detection is routinely used to scout high-performance players [8]. Furthermore, clustering models are heavily used to group players and arrive at appropriate valuations [9, 10]. In addition to these player-focused studies, Woods et al. [11] study the evolution of general gameplay at the team level in Australian football using nonmetric multidimensional scaling and convex hull clusters.

Given the wide-ranging scope of clustering, Xu [12] provides a review of the various steps in clustering and the advances therein. A key step to obtaining good clustering is to identify appropriate feature space. Dalton-Barron et al. [13] employ multiple feature selection techniques to compensate for potential biases and aggregate base rankings via *Borda Count* voting system to select the final set. Given the advances made in the field of deep learning, architectures such as Generative Adversarial Networks (GAN) and Autoencoders are used to identify the feature space [14]. Once feature selection and clustering are performed, it is imperative to visualize the model and the features to explain the predictions. Hinneburg [15] provides several ways from scatter plots and heat maps to parallel coordinates to visualize high-dimensional data and clustering models.

A fundamental, but often overlooked, aspect of any data-driven model is the biases induced by the underlying data. Zhang et al. [16] clustered NBA players using anthropometric attributes and player experience. Although their study was able to identify player profiles with this approach, there

is a potential for ethnic/racial bias when using anthropometric data, such as the height and weight of players. Studies have shown how anthropometric data can be correlated with ethnicity/race, sex, and age [17, 18]. Hedquist [19] used NBA data for seasons and demonstrated that player positions can be successfully identified based on the player's abilities and performance rather than anthropometric features. Therefore, to minimize the risk of having implicit racial/ethnic biases in our analysis, we avoid using anthropometric features.

Gaurav & Chakraborty [20] use synthetic data from FIFA 2018, a video game, to group players. Their analysis is heavily criticized for the drawbacks in FIFA 2018 models and therefore, cannot be readily extended to the real world. Goud et al. [21] investigate the use of data obtained from wearable technology to analyze player performance. However, given the underlying privacy concerns, public access to such data is quite limited. Most leagues ban player monitoring devices during matches and the players are reluctant to public sharing.

4 Proposed Method

Players' forms and performance levels often fluctuate throughout their careers. Therefore, our analysis highlights the progression of player performance and play styles over multiple seasons. We achieve this by clustering players across various teams and leagues. This differs from most sports clustering analyses, which do not incorporate temporal changes over many seasons, and thus tend to have a risk of recency bias. These analyses also tend to cover only a handful of leagues and teams, which limits how much their findings can be generalized. In addition to including a global outlook in our analysis, we have built a comprehensive decision-making tool that helps users interpret players' performances over years, benchmark across cluster centroids, compare any subset of players and eventually build a dream team.

We collected player performance statistics from the Football API across 95 leagues and 15 seasons. Our initial dataset had roughly 760K rows of data with 59 features, each entry representing a player's performance in a particular league and season. Each player's performance metrics were standardized using the minutes they played to fairly compare metrics between players, as some may play more than others. So, these metrics were normalized to a per-game (90-min) scale as per the duration of a soccer game. After removing entries with a significant number of null values and imputing for outliers, the resulting dataset had approximately 120K rows.

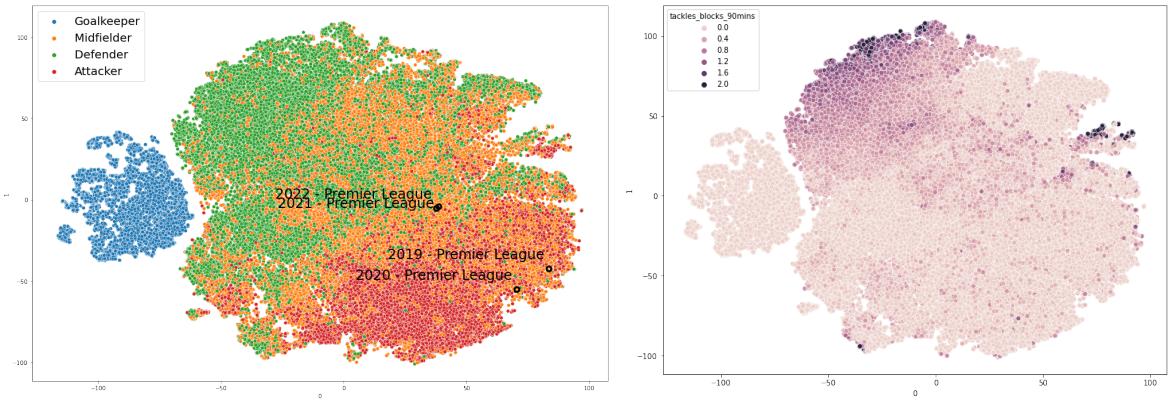


Figure 1: t-SNE scatter plot colored by playing positions and evolution of a specific player (B Fernandes) across League seasons (left). Normalized defensive metric for each player-league-season combination (right)

After downloading and cleaning the data, the first step was to verify if the existing features are informative and enable the separation of players into various clusters. To accomplish this, the data is normalized using Robust Scaler and compressed to a 2-dimensional space using t-SNE. The scatter plot of this low dimensional space is shown in Fig. 1(a), where we see a clear segregation

of goalkeepers from other players. Furthermore, as play styles for each position vary, we can see that the defenders and attackers are broadly separated, with the midfielders sandwiched between them. Figure 1(a) shows the evolution of the style of play of a specific player (Bruno Fernandes is considered for reference). During the 2019 and 2020 seasons he was the leading goal-scorer whereas, in later seasons (2021, 2022), he turned into a provider by assisting other players to score. Thus, we see that this change in play style is clearly captured in this reduced dimensional space obtained via t-SNE. Additionally, the t-SNE modes also show distinct regions in the 2D space corresponding to top performers for various performance metrics (ex. shots at goal vs passing vs tackles, etc). For example, figure 1(b) shows the top performers by a defensive metric (blocks) as a distinct region of the scatter plot.

Next, we reduced the dimensionality of the data via PCA and perform k-Means to identify unique clusters of playstyles. Lastly, we use t-SNE to project the original data to a two-dimensional space, now with the discovered cluster assignments.

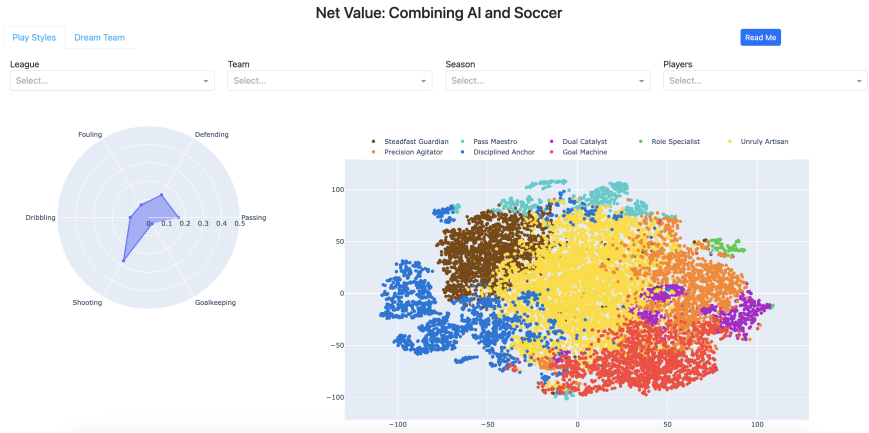


Figure 2: Cluster Plot

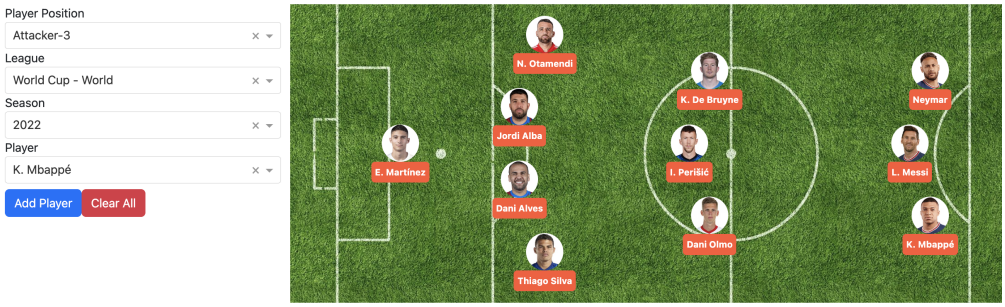
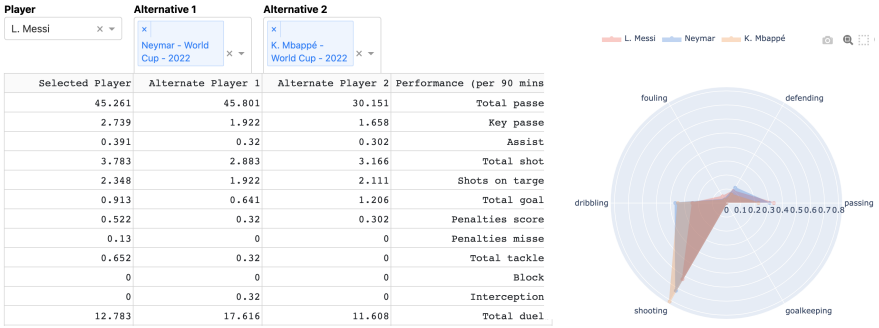


Figure 3: Dream team creator



Selected Player	Alternate Player 1	Alternate Player 2	Performance (per 90 mins
45.261	45.801	30.151	Total passe
2.739	1.922	1.658	Key passe
0.391	0.32	0.302	Assist
3.783	2.883	3.166	Total shot
2.348	1.922	2.111	Shots on targe
0.913	0.641	1.206	Total goal
0.522	0.32	0.302	Penalties score
0.13	0	0	Penalties misse
0.652	0.32	0	Total tackle
0	0	0	Block
0	0.32	0	Interception
12.783	17.616	11.608	Total duel

Figure 4: Player comparison

Our tool displays a 2-D scatter plot of players clustered by play styles (Figure 2). Users are able

to filter the data in this plot for any combination of seasons, leagues, teams, and players, to observe player progression over time. This progression is visualized as a sequence of data points visible in the scatter plot representing how the player's performance changed across seasons. When a player's data points move across different clusters, it signifies a change in their play style, which can be interpreted based on the characteristics of the clusters they move to. Moreover, users can generate a radar plot of the major soccer traits for this filtered data. The data features were combined into six radar plot features as shown in the table below for increased interpretability and ease in information consumption.

Trait	**Description**
Passing	Total passes, key passes, pass accuracy and assists
Dribbling	Dribbling attempts, successes, penalties won and fouls drawn
Shooting	Goals scored, penalties scored, total shots and shots on target
Defending	Total duels, duels won, tackles, blocks and interceptions
Fouling	Fouls committed, yellow cards, red cards, double yellow and penalties committed
Goalkeeping	Total saves, goals conceded, penalties saved

Table 1: Combined features for Radar Plots

Additionally, our tool lets users compare the attributes of different players to help in the team creation process. Users can easily select the player they wish to compare, choose two other players to compare them against, and analyze the corresponding radar plots and feature data to determine which player would be a better fit for their team.

Finally, the tool allows users to create custom teams by plotting players onto a football pitch and laying them out spatially according to their position. Then, we display each player's cluster to expose the distribution of play styles across positions in the custom team.

5 Experiments and Evaluation

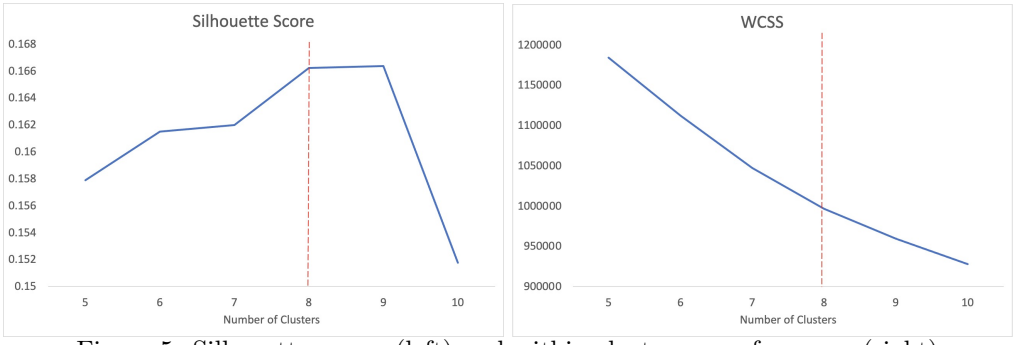


Figure 5: Silhouette scores (left) and within-cluster sum of squares (right)

Cluster	**Description**	**%Players**
Goal Machine	Excellent in attack with high goal contributions	13.3
Disciplined Anchor	Low performing but good at game discipline	15.4
Precision Agitator	Good shooting, excellent dribbling, committing and drawing a lot of fouls	11.3
Unruly Artisan	Jack of all trades, but poor discipline	34.4
Adaptable Ace	Extremely versatile performances	0.8
Dual Catalyst	Good attacking and midfield contributions	5.4
Pass Maestro	Good passers but mediocre performance level	4.9
Steadfast Guardian	Good at defence but non-versatile	14.5

Table 2: Cluster Description and Player Share

Our analysis starts by reducing the dimensionality of the data via PCA to avoid long computation time during clustering. We were able to reduce the number of features from 27 to 15, while retaining 99.5% of the variance of the original data. We then use the k-Means algorithm to identify the play style groupings. To evaluate the effectiveness of our clustering approach, we use metrics such as silhouette score and within-cluster sum of squares. We computed and compared these metrics for k-Means clustering with number of clusters ranging from 5 to 10. We did not want

to use less than 5 clusters, since there already exist 4 unique player positions (attacker, midfielder, defender and goalkeeper), and we wanted to find more nuanced play styles across positions. On the other hand, we capped the potential number of cluster to 10 to avoid clusters that describe the same play styles. We found the best results using 8 clusters, since it has the largest silhouette score and low within-cluster sum of squared. This number of clusters was enough to find unique play styles, while optimizing for the relevant clustering metrics.

We interpreted the clusters in terms of player attributes and play style, and assigned a label and description to better understand the kind of player they describe as shown in Table 2.

The final results of our model can be seen in the fig 6 below. We are able to observe player progression and can compare how an individual's play style varied across different leagues and seasons they participated in. For example, in Fig. 6, we analyze Kylian Mbappe's performance. We notice that during most of Mbappe's time playing in France's Ligue 1, he focuses on shooting and goal-scoring. Meanwhile, during international tournaments, he diversifies his role, focusing on more dribbling and defending, or more passing.

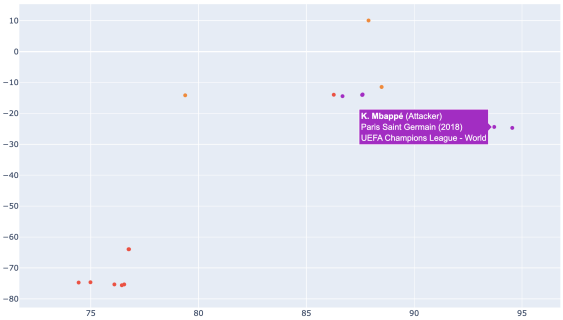


Figure 6: Mbappe play style progression (right)

By combining these two evaluations with stress tests to gauge the scalability of the website, we aim to ensure that our solution produces informative and accurate play style clusters that are effectively presented on the website for users to engage with.

We optimized our dashboard functions to ensure our application had fast loading times, despite utilizing a dataset with over 123K data points. We recorded the execution times when viewing play style clusters with different numbers of players, and in different combinations of the filters, to ensure user convenience. The major functions used to retrieve information and render results on various sections of the Net Value interface are listed in Table 3 along with their corresponding mean execution times or more than 100 trials each. These trials were done after hosting it on the web. When we reset the filters, the default view shows the plots for all the data points, leading to slow loading times of around 2-4 seconds. All the other functions have low loading times making it almost imperceptible to the average user. Computing the clusters in advance and storing and indexing data points for quick filtering and retrieval, helped reduce the loading times and make the experience better.

Function Name	**Description**	**Time(s)**
update_radial_chart	Generates a radar chart for visualizing player, league, and centroid data across six essential soccer attributes.	6.5e-2
update_scatter_plot	Creates t-SNE-based scatter plot of players, with color coding based on playstyles, and filters for league, team, season	1.541
update_team_options	Retrieves and returns a list of teams or players based on specified league and season parameters.	3.05e-1
update_table	Generates a player comparison table displaying various player performance metrics.	0.0068
update_selected_players	Executes a query and renders a specific player on the Dream Team Soccer Ground visualization.	9.344e-3

Table 3: Execution times of different dashboard functions

We shared our tool with a group of fantasy soccer enthusiasts and received 24 responses. Users were asked to rate our application using a few metrics. The mean ratings are presented in Table 1, where the scale goes from 1 (worst) to 5 (best). The results of our user study were positive and reflect the ease of use of the application and the quality of visualization provided. Based on the qualitative feedback, users found the analysis useful as they were able to find suitable players for their fantasy teams and compare it with real-world knowledge. This was an essential win for the team, as the execution goals were successful. By including further data points and more useful metrics, as we discuss in the next section, we believe this tool will be useful to other potential users in making informed decisions.

Metric	**Metric Score**
Visualization Quality	4.64
Clarity in Information	4.41
Analysis and Insights	4.56
Loading Speed	4.35
Ease of Use	4.78

Table 4: User Survey Results

6 Conclusions and Discussion

Our tool, Net Value, suggests a novel method for clustering soccer players based on their individual stats and play style, rather than on their traditional positions. It differs from other similar tools because of the emphasis put on the year-by-year progression and the inclusion of a greater range of leagues and tournaments. Our interactive visualizations will allow the user to observe a player's performance progression in 2-D space and let them draft their own team and assess its overall play styles.

Our visualization provides insights useful for decision-makers in the sports industry or fantasy enthusiasts. Soccer club managers and talent scouts can evaluate player performance by conducting standardized comparisons. This may be useful for salaries or initial offer valuation for players. Moreover, this tool can help fantasy managers create and improve their teams. Since our tool uses many of the performance indicators used in fantasy leagues, it can help managers find alternative players under budget limitations. Managers can also find good players in lesser-known leagues through our tool.

Although many player performance statistics are captured in our analysis, not all metrics are. For example, "off-the-ball movement" metrics, which are not included in our dataset, show how players contribute to the team when they are not directly interacting with the ball. There are other performance metrics such as crosses, passes in the final third, expected goals and assists, clean sheets, big chances created, etc. The tool could be further improved by including machine learning enabled, intelligent player suggestions to the user. This approach could be applied to other sports like basketball and volleyball.

All team members have contributed a similar amount of effort towards this project.

References

[1] Zachary Shelly, Reuben F Burch, Wenmeng Tian, Lesley Strawderman, Anthony Piroli, and Corey Bichey. Using k-means clustering to create training groups for elite american football student-athletes based on game demands. *International Journal of Kinesiology and Sports Science*, 8(2):47–63, 2020.

[2] Sayan Roy and Binoy Sasmal. Integration of hierarchical clustering method and dendrogram method with expectation maximization for identification of the best player cluster. In *2021 7th International Conference on Optimization and Applications (ICOA)*, pages 1–8. IEEE, 2021.

[3] Megan Muniz and Tulay Flamand. A weighted network clustering approach in the nba. *Journal of Sports Analytics*, (Preprint):1–25, 2022.

[4] Eduardo A Abade, Bruno V Gonçalves, Alexandra M Silva, Nuno M Leite, Carlo Castagna, and Jaime E Sampaio. Classifying young soccer players by training performances. *Perceptual and Motor Skills*, 119(3):971–984, 2014.

[5] Iker J Bautista, Ignacio J Chirosa, Joseph E Robinson, Roland van der Tillaar, Luis J Chirosa, and Isidoro Martínez Martín. A new physical performance classification system for elite handball players: cluster analysis. *Journal of Human Kinetics*, 51(1):131–142, 2016.

[6] Serhat Akhanli and Christian Hennig. Clustering of football players based on performance data and aggregated clustering validity indexes. *arXiv preprint arXiv:2204.09793*, 2022.

[7] Pierpaolo D'Urso, Livia De Giovanni, and Vincenzina Vitale. A robust method for clustering football players with mixed attributes. *Annals of Operations Research*, pages 1–28, 2022.

[8] Argimiro Arratia and Martí Renedo Mirambell. Clustering assessment in weighted networks. *PeerJ Computer Science*, 7:e600, 2021.

[9] Radu S Tunaru and Howard P Viney. Valuations of soccer players from statistical performance data. *Journal of Quantitative Analysis in Sports*, 6(2), 2010.

[10] Miao He, Ricardo Cachucho, and Arno J Knobbe. Football player's performance and market value. In *Mlsa@ pkdd/ecml*, pages 87–95, 2015.

[11] Carl T Woods, Sam Robertson, and Neil French Collier. Evolution of game-play in the australian football league from 2001 to 2015. *Journal of sports sciences*, 35(19):1879–1887, 2017.

[12] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[13] Nicholas Dalton-Barron, Anna Palczewska, Dan Weaving, Gordon Rennie, Clive Beggs, Gregory Roe, and Ben Jones. Clustering of match running and performance indicators to assess between-and within-playing position similarity in professional rugby league. *Journal of Sports Sciences*, 40(15):1712–1721, 2022.

[14] Elie Aljalbout, Vladimir Golkov, Yawar Siddiqui, Maximilian Strobel, and Daniel Cremers. Clustering with deep learning: Taxonomy and new methods. *arXiv preprint arXiv:1801.07648*, 2018.

[15] Hinneburg Alexander. *Visualizing Clustering Results. In: Liu, Ling and Özsu, M Tamer (eds) Encyclopedia of Database Systems.*, volume 6. Springer, 2009.

[16] Shaoliang Zhang, Alberto Lorenzo, Miguel-Angel Gómez, Nuno Mateus, Bruno Gonçalves, and Jaime Sampaio. Clustering performances in the nba according to players' anthropometric attributes and playing experience. *Journal of sports sciences*, 36(22):2511–2520, 2018.

[17] Ziqing Zhuang, Douglas Landsittel, Stacey Benson, Raymond Roberge, and Ronald Shaffer. Facial anthropometric differences among gender, ethnicity, and age groups. *Annals of occupational hygiene*, 54(4):391–402, 2010.

[18] Morgana Mongraw-Chaffin, Sherita Hill Golden, Matthew A Allison, Jingzhong Ding, Pamela Ouyang, Pamela J Schreiner, Moyses Szklo, Mark Woodward, Jeffery Hunter Young, and Cheryl AM Anderson. The sex and race specific relationship between anthropometry and body fat composition determined from computed tomography: evidence from the multi-ethnic study of atherosclerosis. *PLoS One*, 10(10):e0139559, 2015.

[19] Alexander L Hedquist. *Redefining NBA Basketball Positions Through Visualization and Mega-Cluster Analysis*. PhD thesis, Utah State University, 2022.

[20] Vishal Gaurav and Goutam Chakraborty. Scouting in soccer with applied machine learning. 2019.

[21] P Sri Harsha Vardhan Goud, Y Mohana Roopa, and B Padmaja. Player performance analysis in sports: with fusion of machine learning and wearable technology. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pages 600–603. IEEE, 2019.