Football Players' Stats

Mauro Ficorella - 1941639[a], Martina Turbessi - 1944497[b], Valentina Sisti - 1952657[c]

[a]ficorella.1941639@studenti.uniroma1.it
[b]turbessi.1944497@studenti.uniroma1.it
[c]sisti.1952657@studenti.uniroma1.it

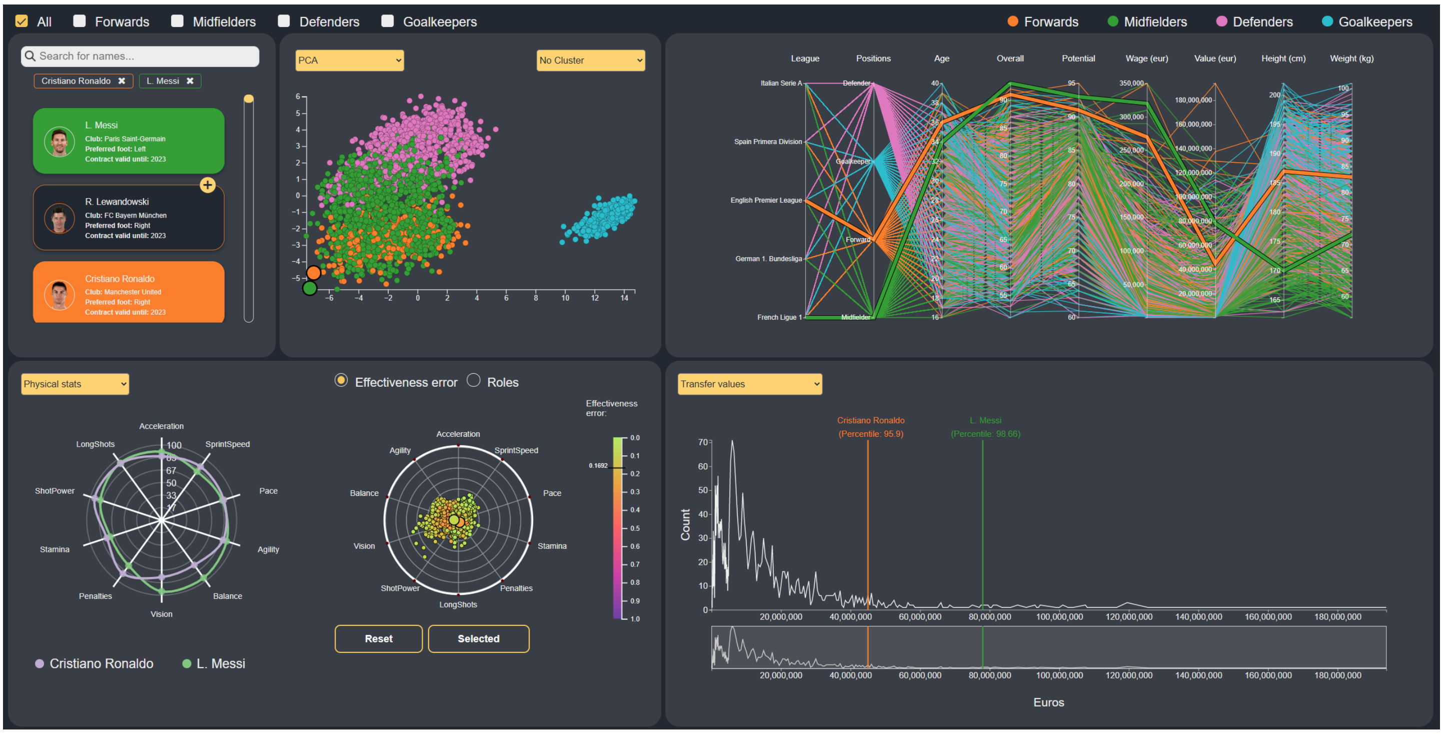


Figure 1: Overview of the complete system

Abstract

Football professionals increasingly need to search for players' statistics and to compare them in order to choose the player that suits them best. Moreover football fans want to look at their favorite players' statistics. The aim of this work is to give them an interactive dashboard made of different visualizations containing statistics and data that can help them to achieve their goals. More precisely, this work uses only data coming from the football season 2021/22 and from the top 5 European leagues.

1. Introduction

Nowadays football is one of the most followed sports in the world. One of the main characters in the football industry are statistics. They are very useful for Club owners to make decisions on players' market, but also for the thousands of fans around the world who want to compare their favourite players. In this scenario, eSports games like FIFA have recently grown in popularity and so this work is based on the FIFA dataset because it's very helpful in mapping all these statistics to all the in-game players.

In particular the work aims to be a big football visualization dashboard through which every people interested in football can make comparisons between players and make better decisions.

2. Dataset

The data used for the visual enviroment are collected from the public dataset "FIFA 22 complete player dataset"[7], related to the football season 2021/2022. The original dataset has more than 100 attributes, from which have been selected only the most relevant ones. Among them there are players' information such as player name, club, league, position, and players' characteristics, such as overall, attacking, defending and goalkeeping statistics.

Moreover, the original dataset contains tuples regarding more than 19000 players, but for this work have been selected only tuples regarding players from the top 5 european football leagues: Italian Serie A, English Premier League, Spanish La Liga, German Bundesliga and French Ligue 1.

Furthermore since in the column related to players' positions there was more than one role for some players, this column has been manipulated in such a way that each player has exactly one position, and this position is not expressed anymore as an abbreviation but as one of the main roles: goalkeepers, defenders, midfielders and forwards (for example ST becomes Forward).

After the above manipulations, the dataset used in this work results in the following dimensions:

- 59 dimensions
- 2.977 tuples
- *AS:* $59 \times 2.977 = 175.643$

3. Visual Environment

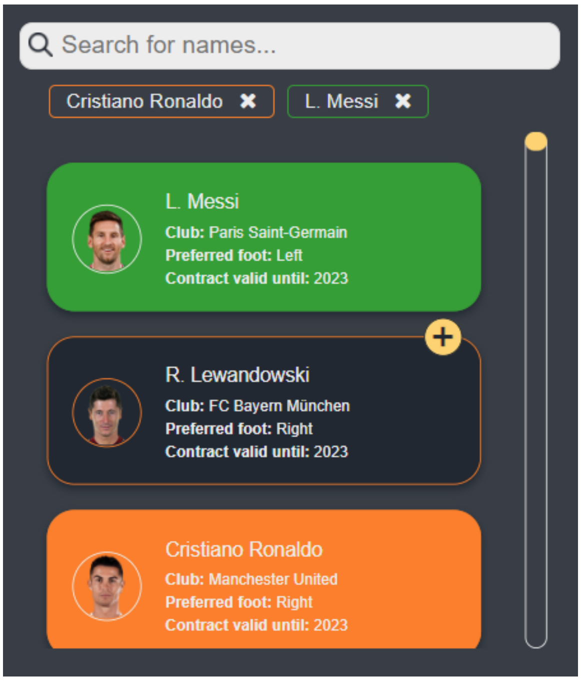
The dashboard is made of:

- Players' List showing players with the related information;
- Scatterplot showing the result of the dimensionality reduction performed on the dataset;
- Parallel Coordinates showing a subset of dataset's attributes through parallel axes in a 2D chart;
- Radar used to compare different players based on the same fixed attributes;
- RadViz used to plot multidimensional data (using a subset of dataset's attributes) on a 2D scatterplot, showing its original dimensions and proportions among attributes;
- Line Chart to show where a player ranks with respect to all the other players in terms of his Transfer Value and his Weekly Wage.

Moreover there are five checkboxes, one for each main role plus one to select all the roles, used to show, in all the other visualizations, only data relative to the selected checkboxes' role. In general, on the top right corner of the dashboard, there is a legend in which are represented the four predefined colors used for each player's role, taken from the categorical scheme provided by the module "d3-scale-chromatic"[1]. More specifically, it has been used *d3.schemeCategory10*, and the following colors have been chosen:

- Orange for Forwards;
- Green for Midfielders;
- Pink for Defenders;
- Cyan for Goalkeepers.

3.1. Players' list



L. Messi
Club: Paris Saint-Germain
Preferred foot: Left
Contract valid until: 2023

R. Lewandowski
Club: FC Bayern München
Preferred foot: Right
Contract valid until: 2023

Cristiano Ronaldo
Club: Manchester United
Preferred foot: Right
Contract valid until: 2023

Figure 2: Players' list

3.1.1. *Visualization:*

The list in Figure 1 contains all the players from the top 5 european leagues. Each player is shown through a card that contains the most relevant information about him, that are not shown in the other visualizations, such as his name, his picture, the football club for which he plays, his preferred foot and his contract's expiration year.

3.1.2. *Interaction:*

On top of the list there is a search bar, through which the list can be filtered searching for a specific player's name. During the digitation of the name, the list is updated dynamically showing only players matching the inserted characters.

Moreover, the user can select one or more players through the "plus" button in order to obtain/highlight, in all the other visualizations, the data related to the selected players. Doing so, the cards related to the selected players get colored depending on the color associated to their role and appears an indicator for each selected player, giving to the user the possibility to notice which players are selected and deselect it with the "x" button, as shown in the Figure 1, where is selected Lionel Messi.

When one or more players are selected, those players are highlighted in the scatterplot, in the RadViz and in the parallel coordinates; moreover, in the radar the data related to them appear, and in the line chart, for each player selected, a vertical indicator appears.

3.2. Scatterplot

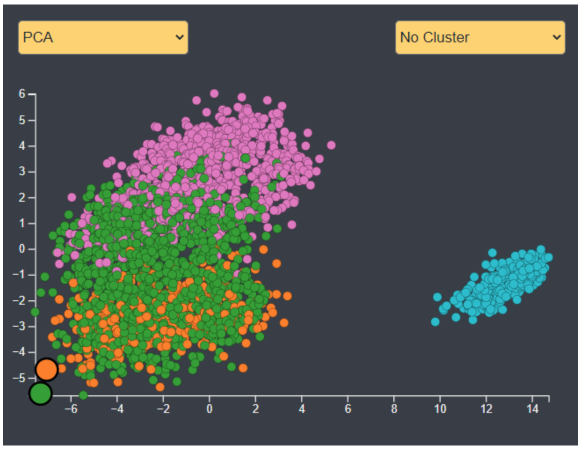


Figure 3: Scatterplot

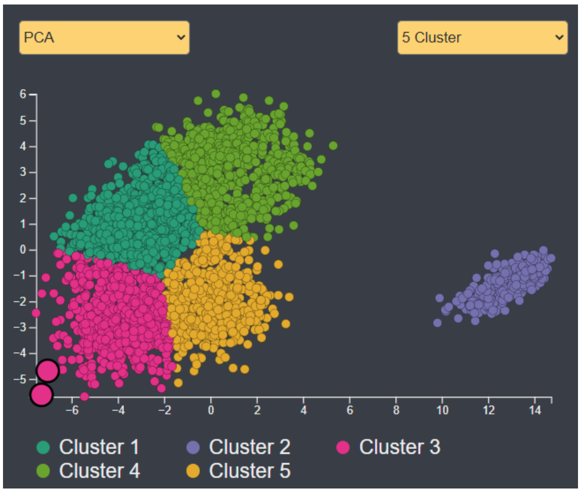


Figure 4: Scatterplot with 5 clusters selected

3.2.1. *Visualization:*

The scatterplot in Figure 2 contains the result of the dimensionality reduction, where each circle represents a player, with the possibility to choose between PCA and t-SNE algorithms, and gives the possibility to show clusters obtained through the K-means algorithm, based on the number of clusters chosen by the user.

If no cluster is selected, the scatterplot colors each circle depending on the related player's role color; otherwise, a color for each cluster has been selected from the categorical scheme provided by the module "d3-scale-chromatic"[1]. More specifically, it has been used *d3.schemeDark2*. When a certain number of clusters is selected, below the scatterplot is shown a legend with the related clusters' colors.

3.2.2. *Interaction:*

The user can interact with the scatterplot through brushing and by focusing the players with the mouse. The brush highlights a certain number of players, and this filters the players shown in the parallel coordinates and in the list, and highlights those players in the RadViz. When a user focuses a player with the mouse cursor, a tooltip appears showing his name. Moreover the user can choose the algorithm used for the dimensionality reduction between PCA and t-SNE through the top-left dropdown menu, and he can also choose the number of clusters to show through the top-right dropdown menu; in particular, when the user selects a certain number of clusters, the K-means algorithm is triggered using this number of clusters.

3.3. Parallel Coordinates

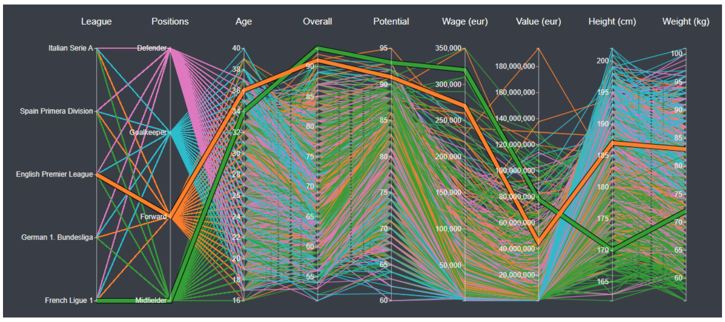


Figure 5: Parallel Coordinates

3.3.1. *Visualization:*

The parallel coordinates in Figure 3 show a subset of dataset's attributes through parallel axes in a 2D chart. More specifically each axis represents the following attributes: League, Positions, Age, Overall, Potential, Wage (eur), Value (eur), Height (cm), Weight (kg). Each line represents a player, and intersects each axis at its corresponding value. Also here each line is colored depending on the role of the player that it represents.

3.3.2. *Interaction:*

Since parallel coordinates suffer from over-plotting, the brushing is implemented on each of their axis in order to filter only certain players, making it easier to look for their characteristics. The brush also filters the players shown in the list, and highlights those players in the RadViz and in the scatterplot.

The user, besides brushing, can also interact with parallel coordinates by focusing the players with the mouse and by dragging each axis. In the first case, the focused player highlights and it appears a tooltip showing his name; in the second case, he can change the position of each axis and the graph updates itself accordingly, and this makes easier to notice the trade-off and the correlation between each axis, moving them near to each other.

3.4. Radar Chart

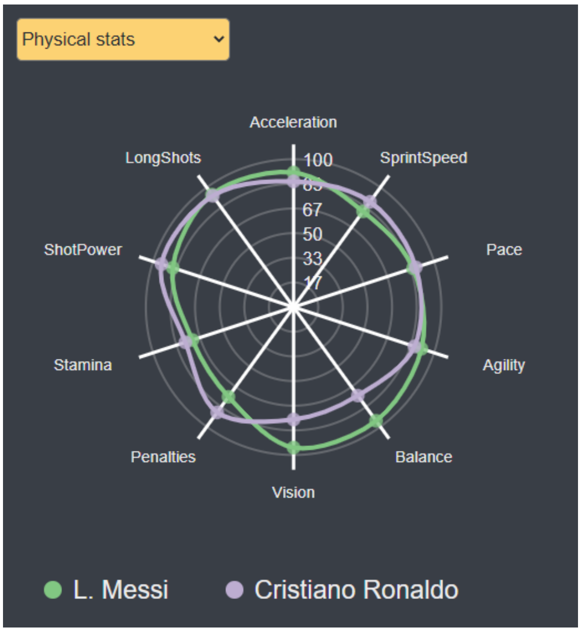


Figure 6: Radar Chart

3.4.1. *Visualization:*

The radar in Figure 4 shows each player represented as a polygon where each point of intersection with each axis represents the exact value of a certain attribute related to that player. This chart allows a direct comparison among players based on their stats, which can be chosen between Overall stats, Attack stats, Defense stats, Keeper stats and Physical stats.

A color for each player has been selected from one of the qualitative sets provided by "ColorBrewer"[8]. Moreover there is a legend containing the name of the selected player and the related color used in the radar.

3.4.2. *Interaction:*

The user can change the stats shown in the radar through the dropdown menu. Moreover he can focus a specific player through the mouse cursor, and in this case a tooltip is shown, containing his name, and the other players decrease their opacity. Furthermore there are circles in each intersection with each axis, and if the user focus them, it appears a tooltip that shows the exact value of the related attribute.

3.5. Radviz

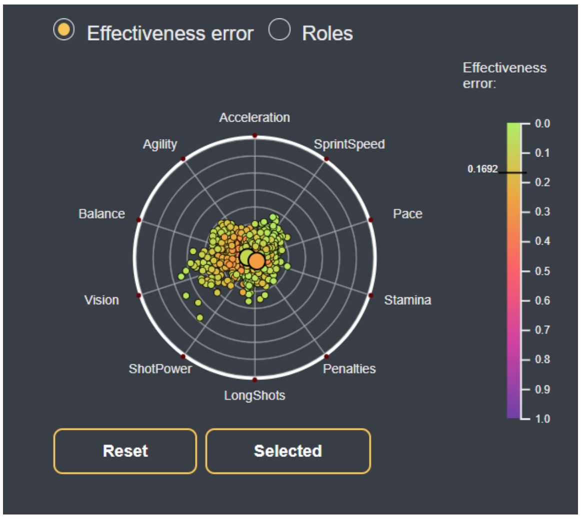


Figure 7: Radviz

3.5.1. *Visualization:*

The RadViz[4] in Figure 5 plots multidimensional data (using a subset of dataset's attributes) on a 2D scatterplot, showing its original dimensions and proportions among attributes. Those attributes are the same that are used in the Radar, and also here they can be chosen between Overall stats, Attack stats, Defense stats, Keeper stats and Physical stats. The dimension arrangement (DA) follows the heuristic described in the paper which minimizes the effectiveness error that degradate the visualization of the radviz. Regarding colors, by default each point encodes the value of Effectiveness Error, ranging in [0,1], from green to violet, as shown in the legend besides, which shows also the Effectiveness Error numerical value. Moreover there is the possibility to change the color, according to each player's role.

3.5.2. *Interaction:*

Also here the user can change the stats shown in the RadViz through the dropdown menu, that is the same used for the radar. If the user focuses a player with the mouse cursor, the related circle is highlighted and it appears a tooltip showing his name. The user can select one or more players from the list, and through the "Selected" button the user can update each circle's position according to the heuristic applied only on the selected players. Moreover the user can drag each anchor in order to see how the effectiveness error changes according to their position. At this point, the user can also click on "Reset" button in order to reset the starting dimension arrangement.

3.6. Line Chart

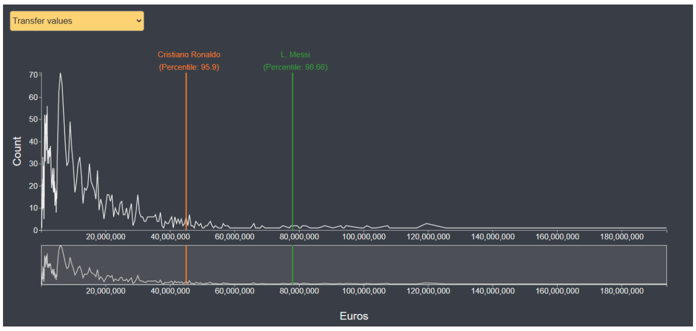


Figure 8: Line Chart

3.6.1. *Visualization:*

The Line Chart in Figure 6 shows a line that represents, depending on the value chosen from the dropdown, the number of occurrences of players that are worth a certain market value or that earn a certain wage. It's used to show where a player ranks with respect to all the other players in terms of his Transfer Value and his Weekly Wages, through a line plotted when the user selects one or more players from the list. Below the graph there is a small preview used to zoom only on certain values.

3.6.2. *Interaction:*

The user can choose whether to show players' transfer values or weekly wages through the dropdown menu. Moreover the user can see, focusing the edges of the plotted line, the percentile of each transfer value or wage. For example, in the Figure 6, it can be seen that Cristiano Ronaldo has the percentile value of around 96, meaning that the 96% of players has a transfer value lower than him. After having selected a player from the list, the user can focus only a certain player's value by moving the mouse cursor over it, and the other lines related to other players' values decrease their opacity.
Furthermore the user can perform a brush through the range selector below the line chart, and doing so the line chart is zoomed in that specific range. This range selector is useful because it allows the user to visualize all selected players' values in the preview also if he zoomed on a certain range that does not contain them all.

4. Analytics

The logical part of this work is implemented through both preprocessing and user triggered analytics, using for dimensionality reduction the PCA and t-SNE algorithms, for clustering the K-means algorithm, and in the line chart there is the count of the number of occurences of players that are worth a certain market value or that earn a certain wage and the computation of percentile.
The user can trigger the analytics by clicking on one or more checkboxes selecting one or more roles, and doing so, the above mentioned computations are performed on demand, on a subset of the dataset defined by the roles selected, through a back-end service developed on a server that continuously listens for requests. This back-end is a RESTful service developed in Python, using the Flask framework. On the other hand, the other computations have been implemented using pandas and numpy for csv manipulations, and using sklearn library for dimensionality reductions and clustering.
The system adopts PCA and t-SNE algorithms for the dimensionality reduction. More specifically, PCA because is useful to give to the user information about players that are far in the original dataset: in fact, after PCA, two points are far away from each other if they were far away from each other in the original dataset.
t-SNE, on the other hand, is useful to give to the user information about players that have similar characteristics: it maps the data in a lower dimensional space, where a small distance between two points means that they were close in the original space; t-SNE only gives reliable information on the closest neighbours, whereas large distance information is almost irrelevant. This is also a good algorithm to separate data into clusters.
So using both of them is helpful because in this way the user can get the best of both worlds.
The clustering is implemented through the K-means algorithm. Since this algorithm requires the number of clusters as parameter, the user can select this number through the dropdown menu above the scatterplot, in order to trigger the computation and have an immediate visual feedback. Its principal disadvantage is that it requires to select in advance the number of clusters. In order to mitigate this, there is the possibility for the user to choose on demand the number of clusters on which to execute the clustering, giving him the possibility of fine tuning the process with an immediate visual feedback.

5. Insights

In this section are presented some insights which can be obtained through the system, depending on system's **intended users**, which can be:

- *Coach of a football club:* he noticed that the club needs a certain type of player, depending on specific characteristics that he should have, so he asks to the Director of his club to buy it;
- *Director of a football club:* he has a certain budget to spend for a player, so he search for the most valuable one, staying within budget, both in terms of cost and characteristics;
- *Football fan:* he wants to discover new facts and statistics about football players.

5.1. Correlation between players of different roles

Suppose that a Coach of a Serie A football club, during the winter transfer market, is using the system in order to find a new defensive midfielder in order to improve his club's midfield.

So he selects the checkbox related to the midfielders, in order to filter them with respect to all the other roles. He knows Cristante, so he searches and selects him, but, before checking his defensive stats, he wants to know if he is also good in the attacking phase. So he filters the midfielders selecting only Serie A and overall greater than 75 from the parallel coordinates, and then he selects two other midfielders, such as Luis Alberto and Çalhanoğlu, because he knows them and he knows that they are good in the attacking phase. Through the radar, selecting attack stats, he notices that they are quite similar in the attacking phase, but that Cristante is more solid in the defensive phase, and so he realizes that is the player that could be the right one for him. Then, in order to check if he is very good in the defending phase, he selects also the checkbox of the defenders in order to compare him with some defenders. He knows that among them, there is Bonucci that is a good defender with also good passing stats, and Koulibaly that is a very strong and solid defender. So, after having selected also them, he notice, through the t-SNE output plotted in the scatterplot, that Bonucci and Cristante are very close to each other, and this is already a good sign meaning that Cristante is very good at defending. After that he performs clustering using 5 clusters, and so he notices that Cristante, Bonucci and Koulibaly fall in the same cluster. This is another confirmation of the fact that Cristante is very solid at defending.

Finally, through the RadViz he notices that Cristante Bonucci and Koulibaly are almost overlapped in the center of the graph related to the defense, meaning that they have similar defensive characteristics, and that, as can be seen from the radar, they have also very similar values regarding defensive stats.

So, after all these conclusions obtained thanks to the system, the coach reaches a significative insight telling him that Cristante is the defensive midfielder that he was searching for.

5.2. Spend money on a player wisely

Suppose that a Director of a football club, is using the system in order to find a new defender for his club that meets his budget of around 30-45mln of euros and that should be tall at least 187cm.

He starts selecting the checkbox related to the defenders, and then, after having noticed, through scatterplot's PCA, that the best defenders are on the right portion of the graph, he decides to brush only that area. Then, since he has a budget of 30-45mln of euros, he brushes, on the parallel coordinates, only this price range in order to see which defenders he can afford. Moreover, since he wants a good defender, he brushes an overall greater or equal than 84, and since he wants a defender tall at least 187cm, he brushes also this value on the height axis of the parallel coordinates.

After that, he has only three defenders left in the list, more specifically Hummels, Maguire and Ginter. So he first compare them through Radar and RadViz in terms of their defending and physical statistics, and he notices that they are very similar to each other. So he can start thinking only about the differences among their market values and wages.

He immediately notices that Maguire earns a much bigger wage with respect to the other two players and he has also the contract valid longer than the other, making it more difficult to negotiate with his club, and so he decides to look only at Hummels and Ginter, since they offer very similar characteristics at a lower wage and with contract expiring very soon (2022). Then, selecting players' wages in the line chart, he notices that Ginter earns a much lower wage than Hummels, and, more importantly, that has a percentile of 79, meaning that the 79% of defenders earn a wage lower than him, that is much lower than the 95th percentile given by Hummels, meaning that is one of the defenders that earn more wage. In addition to this, brushing on line chart's range selector, he notices that Ginter has a transfer value of 42,5mln of euros while Hummels has a value of 44mln of euros. Finally, he notices that Ginter is much more younger than Hummels, with an age of 27 years old against 32 years old, and so, given all the other considerations made before, he decides to go for Ginter, especially with a view to the future.

5.3. Discover new facts about players

Suppose that a football fan is using the system in order to see football players' statistics and to discover new facts about them.

He notices immediately, from height axis of the parallel coordinates, that all the lines of players that are concentrated in the top part of this axis belong to goalkeepers, and so he can find out the insight that the tallest players are the goalkeepers. Moreover, he can see that, in the axis related to transfer value and wage, the majority of players is concentrated in a transfer value less than 40mln euros and in a wage less than 100,000 euros. Then, in order to deepen into these data, he looks at line chart and he can see that, regarding both transfer values and wages, respectively, on 40mln the percentile has a value of about 95, meaning that the 95% of players has a value lower than that, and on 100,000 the percentile has the same value of about 95, meaning also here that 95% of players has a wage lower than that. Staying on topic, the user can notice that there is a line that stands out with respect to the others regarding the transfer value. So he moves the mouse over it in order to see to which player is related, and he discovers that is related to Mbappé, and so that he has a value a lot higher than the other players, since the second highest transfer value has a difference of more than 40mln.

Then the user selects goalkeepers, and, from the PCA output in the scatterplot, he notices that there is a goalkeeper that is an outlier, and he wants to know why. So he moves the mouse over his circle in the scatterplot in order to see his name that is Milinković-Savić. After having searched for him, he searches for other goalkeepers in order to better understand why he is an outlier, and he notices, through the radar on the attack stats, that, unlike any other goalkeeper, he has very good shooting stats (freekicks accuracy and curve shooting).

6. Related works

This work aims to be a big football visualization dashboard through which every people interested in football can make comparisons between players and coaches and directors of football clubs can make better decisions.

This section explores the related works, explaining this work's contributions with respect to the different available interactive football visualizations and current solutions used for summarizing players statistics.

Football professionals, such as coaches, are starting to change mentalities and to increasingly trust sports analysts. This has encouraged some people to make more understandable data unsing interactive visualizations.

From now on, this work will be referred as *"FPS"*. *"Analyzing In-Game Movements of Soccer Players at Scale"*[6] aims at representing the physical performance of players by analyzing and summarizing the movement of players in a game. While they computed the similarities of players based on their movement characteristics, *FPS* aims at representing those similarities through the overall attributes, in order to give to football professionals a much more complete overview of each player. Moreover, *FPS* and [6] share the goal of find similarities between players using the scatterplot visualization, with the difference that they use uniqueness and consistency obtained through their algorithm, while *FPS* uses dimensionality reduction's output obtained through all players' attributes.

From now on this section will analyze works that are more related to players' statistics. First, in 2010, *"Soccer Scoop"*[9] was one of the first softwares that tried to extensively use data to support choices of football professionals. More specifically it aimed to easily measure the performance of a player and compare him to other players. The basic idea is very similar to *FPS*, and the same is valid also for the intended user. In fact, they based their intended user on football directors that would use their system before signing players in their club. However the visualization was very focused on numerical values, neglecting aesthetics and easiness of use.

The current way amateurs get access to the summary statistics of their favorite player is through static websites. There is a big choice of these static websites, but the principal ones are *"Transfermarkt"*[3] and *"SoFIFA"*[2], which show in a very clear and effective way players' statistics. Regarding Transfermarkt, the only interactivity is a simple tooltip of a line chart that shows the value of the player during the years. On the other hand, SoFIFA implemented the interactivity, also in this case through a line chart, with a tooltip showing overall and potential of a player during the years and a range selector used to zoom on a certain period of time to analyze overall and potential of that player. *FPS* used the same approach of the line chart used by SoFIFA. More specifically, *FPS* uses the range selector to zoom on certain transfer values/wages and a vertical line to rank a certain player with respect to his transfer value/wage, plus a tooltip to indicate player's name.

Another work based on player's statistics is *"Infootmation"*[5], that, among other things, gives the possibility to football amateurs to learn the statistics of their favorite players and to see how a player compares with the average. They used visualizations similar to *FPS*. More in details, they used a player card to show player's info such as his name, club in which he plays, favourite foot and his photo, and a search bar to search for a player. Moreover they used also a radar chart to show player's overall characteristics, compared to the average of the other players. *FPS*, instead, implemented the radar in a slightly different way, showing characteristics not only regarding overall but also regarding attack, defense, keeper and physical statistics, and, more importantly, giving the possibility to compare a player with other players at the same time. The last work analyzed was *"Gaussian mixture clustering model"*[10], that used Principal Component Analysis in conjuction with a model-based Gaussian clustering method with the purpose of characterizing professional football players. They represented in a scatterplot the projection of the clusters obtained by their model, and this showed a very similar result to *FPS*'s PCA output plotted in the scatterplot. Moreover, they used a radar chart to represent mean values for each attribute grouped by players' roles. In this case, this work uses a different way to represent data in the radar chart, showing only data related to one or more players, using a considerably smaller number of attributes, in order to make easier to compare them each other.

7. Conclusions and future works

This work tried to offer an interactive dashboard to all football fans and professionals. It took advantage of multiple visualizations coordinated each other in order to better guide the final users to reach their goals and satisfy their needs. Thanks to this work, they can discover correlations between different players and obtain useful insights both from technical statistics and economic perspectives. Last, but not least, this system gave the possibility to the user of performing on-demand analytics, through which the user can further refine their analysis.

The related works yielded interesting hints for further future improvements. For example it could be interesting to add information about players' scored goals and players' heatmap in order to give more information about their position during the various matches of the season.

Another nice addition would be to update dynamically the information offered by the system for the next and previews seasons, in order to let the user notice how statistics have changed during the years.

References

[1] Data driven documents, "d3 scale chromatic references. https://github.com/d3/d3-scale-chromatic.

[2] Sofifa. https://sofifa.com.

[3] Transfermarkt. https://www.transfermarkt.it.

[4] Marco Angelini, Graziano Blasilli, Simone Lenti, Alessia Palleschi, and Giuseppe Santucci. Effectiveness error: Measuring and improving radviz visual effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[5] Yann Dubois. Infootmation: Visualize soccer player statistics. https://www.cs.ubc.ca/~tmm/courses/547-17/projects/yann/report.pdf.

[6] László Gyarmati and Mohamed Hefeeda. Analyzing in-game movements of soccer players at scale. *CoRR*, abs/1603.05583, 2016.

[7] Stefano Leone. Fifa 22 complete player dataset. https://www.kaggle.com/stefanoleone992/fifa-22-complete-player-dataset.

[8] Cynthia A Brewer Mark Harrower. Colorbrewer. org: an online tool for selecting colour schemes for maps. 40(1), 2003/6.

[9] Adrian Rusu, Doru Stoica, Edward Burns, Benjamin Hample, Kevin McGarry, and Robert Russell. Dynamic visualizations for soccer statistical analysis. In *2010 14th International Conference Information Visualisation*, pages 207–212, 2010.

[10] César Soto-Valero. A gaussian mixture clustering model for characterizing football players using the ea sports' fifa video game system. *RICYDE. Revista internacional de ciencias del deporte*, 13:244–259, 07 2017.