

Recommendation system for football player recruitment using k-nearest neighbor

Maukar¹, Rodiah²

¹Master of System Information Management, Universitas Gunadarma, Depok, Indonesia

²Department of Informatics, Faculty of Industry Technology, Universitas Gunadarma, Depok, Indonesia

Article Info

Article history:

Received Aug 25, 2024

Revised Jun 23, 2025

Accepted Jul 10, 2025

Keywords:

Cosine similarity

Featured

Football player

Prediction

Recommendation system

ABSTRACT

In modern professional football, achieving a competitive edge depends not only on on-field performance but also on effective off-field strategies, particularly in player recruitment. This study proposes a machine learning-based recommendation system to support talent identification and optimal player placement using statistical performance data. The model analyzes a wide range of features, including shots, expected goals, expected assists, pass types, offensive contributions, and defensive actions across field zones. The dataset undergoes preprocessing steps such as normalization (per 90 minutes) and dimensionality reduction. A key innovation of this research is the use of principal component analysis (PCA) to reduce feature dimensionality, minimizing redundancy while retaining essential information, which improves model efficiency and scalability. The refined data is then processed using the k-nearest neighbors (KNN) algorithm with cosine similarity, allowing the system to identify players with similar performance profiles based on directional similarity in a high-dimensional space. This combination enhances recommendation accuracy by focusing on performance structure rather than raw values. The resulting system provides actionable insights into player suitability and potential, offering clubs a data-driven tool for informed scouting and recruitment decisions. The approach demonstrates the effectiveness of combining PCA and KNN in optimizing football player recommendation systems.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Rodiah

Department of Informatics, Faculty of Industry Technology, Universitas Gunadarma

Margonda Raya 100, Pondok Cina, Depok, West Java, Indonesia

Email: rodiah@staff.gunadarma.ac.id

1. INTRODUCTION

The application of data science and machine learning in sports, particularly professional football, has grown significantly in recent years [1]. These technologies are increasingly used not only to enhance team performance on the field but also to support strategic decisions off the field [2], such as optimizing player recruitment and placement [3]. As competition intensifies across top leagues, clubs seek to gain a competitive edge by integrating data-driven approaches into scouting processes. Football clubs must constantly evaluate and replace players due to transfers, injuries, or performance issues [4]. Traditional scouting methods, while valuable, are often subjective and costly. Machine learning offers a more scalable and objective solution, capable of evaluating vast datasets to identify players whose statistical profiles align with team needs [5].

Advanced football statistics such as expected goals (xG), expected assists (xA), take-ons, and defensive actions [6], are now commonly used to assess player performance [7]. Previous research [8], [9]

has explored predictive modeling for player transfers, performance evaluation, and role classification. For example, some studies developed predictive models based on transfer data [10], while others applied clustering and classification methods to identify player roles or analyze game strategies [11]. However, many of these approaches either overlook dimensionality issues [12] in high-dimensional datasets or fail to integrate similarity-based models effectively into player recommendation systems.

Several studies related to recommendation system for football players based on their performance and talent using data science and machine learning concepts have been carried out by previous researchers. Dinsdale and Gallagher [13] created a model to predict the performance of a football player after moving from one club to another and with a model that took input from 13 features of statistical data on the performance of a football player obtained from 26,000 samples of transfer and non-transfer data from 32 domestic leagues from 2017. Models in research [13] compared to a sample of 2,659 historical transfer data and 8,677 historical non-transfer data and compared to the predictions of a model that has been created where players are assumed to continue to produce the same performance from before and after the club transfer. Bunker and Thabtah [14] explore the use of machine learning in predicting outcomes in the increasingly important professional sports world of sports team management as well as in the gambling industry. Researchers argue that machine learning is a promising method to achieve high-accuracy predictions in this field, and researchers propose a new framework called sports result prediction cross-industry standard process for data mining (SRP-CRISP-DM) to predict sports outcomes, this framework includes several stages such as data collection and preprocessing, feature selection, model creation and model evaluation and deployment. Chavan [15] tries to find a solution to the problem to find the closest match of the player to be replaced using machine learning algorithms. Players will be classified based on ratings, in this study six machine learning algorithms were used, namely support vector machine (SVM), linear discriminant analysis (LDA), naïve Bayes, decision tree, XGBoost, and k-nearest neighbor (KNN), then a comparison was made between the algorithms and it was found that LDA and SVM had the best accuracy with 83.77% and 80.31% while the KNN algorithm produced results that had the closest match to the predicted player.

Li *et al.* [16] characterizes the type of play of footballers in the Chinese football super league (CSL) league obtained from 960 match data from 2016-2019. The first player will be clustered into 8 positions then a one-player vector will be created for each player in each match based on player vectors using nonnegative matrix factorization (NMF). As a result, 18 types of players were found to play in the CSL and in general the type of playing forward and midfielder is directly proportional to the trend of the evolution of football performance, while the type of playing defenders must be reconsidered, the type of multifunctional play is also found among CSL players. Yean *et al.* [17] found that machine learning algorithms can also be applied to several classification problems including clinical studies, one of which is in the analysis of the emotions of stroke patients. The KNN algorithm relies on metric distance to calculate the nearest class for classification. The purpose of this study was to compare the performance of several different distance metrics to be applied to the classification of emotional electroencephalogram (EEG) between stroke patients and ordinary people. The result is that the city block distance metric has the best performance among others. Li *et al.* [16] was found that 18 types of footballers played in the CSL, the type of playing attacking players and midfielders is directly proportional to the trend of the evolution of playing style, while the playing style of defenders must be reconsidered. A wide variety of distance metrics were tried in the study [17] and it was found that the difference in the metrics used had an effect on the performance of the model.

This study addresses these limitations by proposing a principal component analysis (PCA) enhanced KNN recommendation system using cosine similarity to recommend football players based on performance similarity. PCA is employed to reduce feature dimensionality while preserving critical statistical information, thus improving model performance and interpretability. By transforming high-dimensional player performance data into a lower-dimensional space, PCA minimizes redundancy and highlights the most influential features. These optimized features are then utilized in a KNN model with cosine similarity, which calculates the angular similarity between player vectors, making it particularly effective in identifying players with structurally similar play styles, independent of raw magnitude. This methodological combination enhances both the precision and scalability of the recommendation process.

The dataset used in this study is sourced from football-reference (FBref) [18] and consists of player statistics from the top five European football leagues during the 2022–2023 season [19]. Thirteen performance features are selected for outfield players, including shots, xG, xA, crosses, total passes, and various types of defensive and offensive contributions. PCA reduces these features while maintaining data integrity. The model then identifies the most similar players based on cosine similarity scores, offering a practical tool for scouts and analysts to make informed decisions about potential recruits.

2. METHOD

This research consists of several stages of the process starting from data collection to the creation of a recommendation system based on the similarity of match historical data as can be seen in Figure 1. At the final stage, the dataset that is ready will be given into the model to be created, namely the player recommendation system. The model will try to predict the distance between data points using the KNN algorithm and the cosine similarity metric. The dataset used is data that has been dimensionality reduction before.

The research began with the process of collecting data from a player statistics collection site called FBref [18]. The data collected includes various features such as xG, xA, and several other features. This data collection process is carried out by scraping data on player statistics available on the site. After the data is successfully collected, the next step is to carry out data preprocessing where this stage will consist of several stages, the first of which begins with the data exploration stage, namely the exploration and analysis of each feature in the dataset in order to get accurate information on the best steps to be taken next. After the data exploration process is complete, the next step is to do data cleaning. This process involves removing data that is irrelevant to the purpose of the study or inaccurate data, as well as handling null data or missing data. In addition, data normalization will also be carried out, which is the process of changing the values in the dataset so that the data is on the same scale. After data cleaning, the data transformation process will be carried out. This process involves converting the data into a format that is more suitable for modeling later, such as converting categorical data into numerical data, after data transformation, feature selection will be carried out.

This process involves selecting the most relevant and informative features to use in the model. These features are chosen based on their importance in predicting target variables, reducing the chance of underfitting as well as overfitting [20] but still represents the dataset as a whole. After feature selection, model creation and training are carried out. This model will be used to create a recommendation system where users can enter input in the form of players who will be looking for the most similar player according to the feature or statistics that represent the player. The model will then be tested to see how accurate the predictions are. Once the model has been created and trained and the results are obtained, the model will be deployed using the streamlit framework [21].

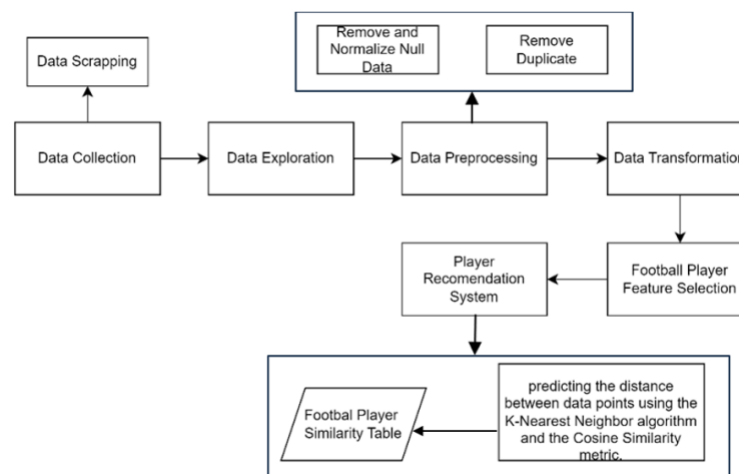


Figure 1. General chart of football player recommendation system research methods

2.1. Data collection and preparation

The development of this recommendation system begins with the collection and preparation of relevant data. High-quality and well-structured data is essential for building an accurate and reliable machine learning model, particularly in the context of football player performance analysis. This stage involves selecting a credible data source, identifying key performance features, and performing data cleaning to ensure consistency and completeness before proceeding to the modeling process.

2.1.1. Data source and selection

This study uses player performance data sourced from FBref [18], a reputable football statistics platform that aggregates advanced player metrics across major football leagues. FBref is widely adopted in professional sports analytics due to its data accuracy, breadth, and accessibility. For the purpose of this research, we focus on statistical data from the top five European leagues - Premier league, La liga, Bundesliga, Serie A, and Ligue 1, specifically from the 2022-2023 season. The dataset includes both outfield

players and goalkeepers, and is curated to ensure consistency in the metrics analyzed across all leagues. The selection of features is based on their relevance in measuring a player's overall contribution and positional effectiveness, covering offensive, defensive, and transitional aspects of the game.

2.1.2. Feature selection

Thirteen key performance metrics were selected for outfield players, including: shots, xG, xA, crosses, total passes, short passes (<32 m), long passes (≥ 32 m), passes in attacking third, penalty area entries, take-ons, defensive actions in own third, defensive actions in middle third, and defensive actions in opposition third. For goalkeepers, four position-specific metrics were considered, with PCA later reducing this to three components. The selected features ensure a comprehensive representation of each player's in-game behavior and role.

2.1.3. Data cleaning and integration

Following collection, the datasets were merged and cleaned. Players with incomplete or missing values in the selected features were excluded to preserve data integrity during modeling. The final dataset was then segmented by position (outfield vs. goalkeeper) to facilitate feature-specific dimensionality reduction and modeling.

2.2. Data exploration

After completing the data collection and cleaning stages, an exploratory data analysis (EDA) was conducted to gain a deeper understanding of the dataset and to guide subsequent preprocessing decisions. This step is essential to identify structural issues, assess the completeness of the data [22], and prepare it for dimensionality reduction and modeling [23]. Exploration was conducted using the Pandas library in Python [24], which offers powerful data handling capabilities. The dataset was first inspected using the `.shape` attribute to understand its structure. For outfield players, the dataset consisted of 2,827 rows and 151 columns, representing individual players and their performance features, respectively. A key focus of the exploration was the handling of missing values, which can significantly impact model accuracy. Missing values (null) in this context typically arise due to players not recording a value in a particular statistical category often because they did not engage in that type of play during the season. To quantify this, the `isnull()` function was used in conjunction with `sum()`, revealing a total of 4,585 missing entries across various columns. Rather than imputing potentially biased values, rows with missing data in critical features were excluded to preserve the statistical integrity of the dataset. In addition to missing data, the presence of duplicate entries was also investigated. Duplicate records in this study were primarily due to players transferring between teams within the same season, which resulted in multiple entries under the same player's name. This was verified using the `duplicated()` function [25] on the 'player' column, where 70 duplicated records were identified. Instead of eliminating duplicates blindly, domain-specific considerations were applied: the most complete record or the latest club data for the season was retained to ensure relevance to recruitment analysis.

2.3. Data preprocessing

At this stage, the majority of the information on the dataset has been known thanks to several stages that have been carried out before, so that the most efficient stage can be carried out to continue the research. In this preprocessing stage, it is divided into several smaller stages, namely:

- i) Data cleaning stage for redundant and valuable data null omitted from dataset to ensure maximum model performance [26]. On the code snippet below some featured data such as nominal data such as which team the player played in, the age of the player, and the nationality of the player are omitted from the dataset so that the only data left is quantitative data that will be used as data training for the model that will be made later.
- ii) In addition to the dataset feature, a duplicate data can also be said to be duplicate, according to the exploration carried out in the previous stage, it is known that there are 70 duplicate data in the 'player' column to overcome this, each player will be given a unique id and each data owned by the player will be combined into one to each unique id.
- iii) At the data stage exploration previously, it was found that there were still 4,585 amounts of data that were null then it is necessary to take steps to eliminate these data, by using function other provided library Pandas are `fillna()` [27] valuable data null replaced with a value of zero or 0 as can be seen in the code snippet. Data that is null, this is replaced with 0 so that the data remains consistent, and the shape of the data is less skewed to one side.

2.4. Data transformation

Once the data is clean of redundant data and the value of null next is to change the whole dataset into a form that supports maximum model performance. In a football statistic, a player who has more minutes

played tends to have a higher statistical value than a player who has less minutes of play data [28]. Therefore, this can be overcome to normalize statistical data into a form per 90 minutes [29] namely normalizing the data so that all player data is considered equal even though they have different minutes of play using the (1).

$$\text{Data Per 90 minutes} = \frac{\text{Data}}{\text{Total Minutes Play}} \times 90 \quad (1)$$

2.5. Feature selection of football players

Not all featured at dataset will be part of the training data against the model, to reduce the underfitting and overfitting then it is necessary to select the data feature or feature selection [30]. The player statistical data feature is selected the best to represent each player, which is selected as many as 13 featured data for each player outfield as follows:

- i) Shots: the number of shots the player has made.
- ii) xG: the probability of a player scoring a goal in each kick taken (on a scale of 0-1).
- iii) xA: the probability of a player scoring a pass that will be converted into a goal by a teammate (on a scale of 0-1).
- iv) Crosses: the number of times a player makes crosses.
- v) Total passes: the total passes made by the player.
- vi) Total short passes (<32 m): a short pass or pass that moves shorter than 32 meters by a player.
- vii) Total long passes (≥32 m): the number of long passes or passes that move more than 32 meters by a player.
- viii) Passes in attacking thirds: the number of passes a player makes in 1/3 of the field in the opponent's area.
- ix) Penalty area entries: the number of passes a player makes into the box.
- x) Take-ons: the number of attempts to pass a player by dribbling.
- xi) Defensive actions in own third: the number of defensive actions performed by players in 1/3 of their own field.
- xii) Defensive actions in middle third: the number of defensive actions made by players in the middle of the field.
- xiii) Defensive actions in opposition thirds: the number of defensive actions a player performs on 1/3 of the field in the opponent's area.

In this study, the selected players are also players who have 90 minutes or full playing numbers during a match at least 3 times to reduce the number of players who have minimal data samples using the following Pseudocode:

```
out_df = grand[grand['90s']>=3]
gk_df = gk_grand[gk_grand['90s']>=3]
```

Furthermore, it will be carried out dimensionality reduction using the PCA provided by Library sklearn [31], at this stage the dataset will again be selected for the data features that best represent the entire information from the data using the following pseudocode:

```
pca = decomposition.PCA()
pca.n_components = 13
pca_data = pca.fit_transform(out_data)
```

At the preprocessing consists of feature selection using PCA where the data dimension is reduced but still retains the majority of the information [32] data, thereby reducing the possibility of underfitting and overfitting. Dimension reduction using PCA is carried out on two data, namely player data with position outfield and also players with the goalkeeper position.

2.6. Creation of recommendation system

At this stage, the dataset is ready to be fed into the model to be created, the model will try to predict the distance between data points using the KNN algorithm and the cosine similarity metric. The dataset used is data that has been dimensionality reduction before. Library SciPy provides a function named distance which will work by performing computations between two or more data points in N-dimensional space [33]. In the implementation, each player is assigned a unique identifier, which is then used to systematically pair and compare players in a looped process. The comparison is conducted iteratively using a distance function that calculates similarity scores between all possible player pairs. Furthermore, the result data is normalized on a scale of 0-100 to obtain data that is exclusive across all components dataset. All results from the model are then fed into the form of pickle or a form of storage provided by Python so that the data from model training can be stored and reused later [34].

3. RESULTS AND DISCUSSION

To reduce dimensionality and improve model efficiency, PCA was applied separately to outfield players and goalkeepers. The goal was to retain the majority of variance ($\geq 95\%$) while minimizing the number of features to avoid overfitting and underfitting during KNN-based similarity matching. Figure 2 shows the explained variance ratio for outfield player data. Originally, 12 features were used. After performing PCA, it was observed that 10 components retained 98.7% of the total variance, making them sufficient for representing the data without significant information loss. This selection was made using the cumulative explained variance threshold, a standard practice in PCA-based modeling.

Similarly, goalkeeper data began with 4 features. Based on the results of the PCA, only three principal components were required to retain 100% of the variance in the dataset. This demonstrates PCA's effectiveness in handling both large and compact feature sets, especially for highly specialized positions such as goalkeepers. The dimensionality reduction process significantly reduced noise in the data and improved the computational performance of the model while preserving the underlying player behavior patterns. However, while PCA effectively reduces feature space and maintains high information retention, it is essential to critically examine its limitations. Suggested alternatives for future work can be seen in Table 1.

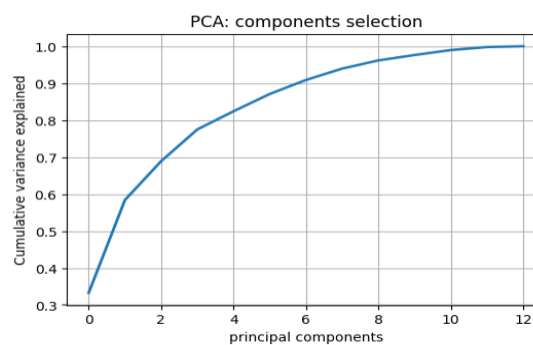


Figure 2. PCA for outfield position player data

Table 1. Suggested alternatives for future work

Technique	Strengths	Limitations
PCA	Fast, retains maximum variance, widely used	Linear, less interpretable
t-distributed stochastic neighbor embedding	Captures nonlinear relationships well	Computationally expensive, poor for new data
Uniform manifold approximation and projection	Preserves both local and global structure	May require fine-tuning, less interpretable
Autoencoders	Learns deep, nonlinear features	Requires larger datasets and training time

3.1. Impact on recommendation results

The results of the reduction by selecting features using PCA succeeded in recommending players with patterns that closely match based on football-specific behavior. For example, as shown in Figure 3, the recommendation system identified Riyad Mahrez as the most similar player to Lionel Messi, with a similarity score of 91.85%. This result illustrates the strength of PCA in capturing latent behavioral patterns across high-dimensional football performance data. By filtering out less relevant features and retaining those that contribute most to the variance, PCA enables the recommendation system to match players with remarkable precision. The similarity score of 91.85% between Lionel Messi and Riyad Mahrez in Figure 3, reflects how PCA preserves nuanced playing styles while enhancing computational efficiency.

recommendations similar to Lionel Messi (Paris S-G)					
Player	Similarity	Position	League	Age	90s
1 Riyad Mahrez (Manchester City)	91.85%	FWMF	Premier League	32	18.5000
2 Marco Asensio (Real Madrid)	90.9%	FWMF	La Liga	27	14.7000
3 Pablo Sarabia (Paris S-G)	90.4%	FWMF	Ligue 1	31	4.1000
4 Thomas Müller (Bayern Munich)	89.96%	FWMF	Bundesliga	33	17.8000
5 Badredine Bouanani (Nice)	89.37%	FWMF	Ligue 1	18	8.3000

Figure 3. Example of recommendation system after feature selection using PCA

3.2. Results of shots position distribution representation analysis on the football field

For example, Figure 4 shows a representation of shots position distribution on the football field [35]. Players with the goalkeeper position, the feature will be separated into three, namely:

- i) Average pass length: the number of goalkeepers pass distances (in meters).
- ii) Average goal kick length: the total distance of a goalkeeper (in meters).
- iii) Crosses stopped: the number of crosses that the goalkeeper has stopped.

The results of the findings show that based on all statistics, the players who are represented according to the position and role played in a team are selected accurately and can be categorized into four categories that describe the abilities of outfield players that can be seen as in Table 2. One of the profiles of champions league football players Alexander Isak from Borussia Dortmund can be seen in Figure 5 when visiting Athlitikos Podosferikos Omilos Ellinon Lefkosias (APOEL) Nicosia in October 2017 [36]. As can be seen in one example of a European league profile, Sweden striker Alexander Isak has 40 caps with two goals in the Premier league from his first three games. The player still scores 10 goals in the league (eight non-penalty goals). Based on the statistics in Figure 5, out of a total of 52 shots (32 right foot and 9 left foot), head is 11 with an accumulation of xG of 6.7 and xG per shot of 0.14 and puts the player as a center forward.

PCA also proves beneficial in analyzing distribution-based features such as shots and positional behaviors. As demonstrated in Figure 5, the profile of Alexander Isak, PCA abstracts complex feature interactions (e.g., foot preference, xG efficiency, and heading ability) into concise components. These components retain the essential variance required to differentiate player types and roles. This abstraction enables the recommendation engine to match players not only on absolute shooting metrics but also on nuanced, category-level behavior patterns. The preservation of player characteristics across feature categories, such as shooting ability and survivability, further illustrates PCA's strength in maintaining football-specific context within a reduced dimensional space.

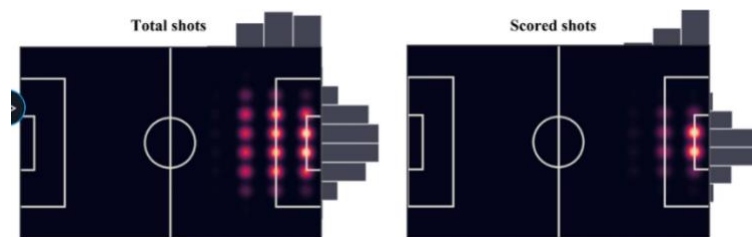


Figure 4. Shot position distribution on the football field [35]

Table 2. Category division of feature data

It	Category	Featured
1	Shooting ability	Shots, xG
2.	Bait ability	xA, crosses, total passes, total short passes, total long passes, passes in attacking thirds, penalty area entries
3	Ball-carrying ability	Take-ons
4	Survivability	Defensive actions in own third, defensive actions in middle third, defensive actions in opposition thirds

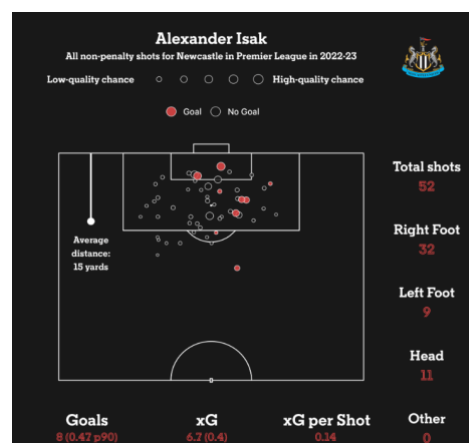


Figure 5. The profile of one of the champions league football players [36]

3.3. Prediction system results based on player talent

After the development and evaluation of the recommendation model, the final step involved deploying the trained model to an interactive and user-friendly platform. This was accomplished using Streamlit, a Python-based framework ideal for building web applications for data science projects. The deployment was carried out within a Google Colab notebook, integrating the trained model, player dataset, and interface logic.

Before deployment, dimensionality reduction using PCA was applied to the dataset. This step was essential for improving computational efficiency and eliminating noise from irrelevant features. The number of retained principal components was determined based on the cumulative explained variance criterion, ensuring that at least 95% of the dataset's variance was preserved. For outfield players, this resulted in the retention of 11 principal components, while for goalkeepers, 3 components were sufficient due to their more specialized and fewer performance metrics. This selection balances dimensionality reduction with information preservation, optimizing the performance of the KNN algorithm while minimizing overfitting and computational overhead. The deployed system enables users, such as analysts, scouts, or coaches to identify players with similar statistical profiles based on historical match data. The key features of the platform include:

- Player type selection: users can specify whether the query targets an outfield player or a goalkeeper, enabling position-specific recommendations.
- Search functionality: a text input allows users to search for a player by name. Upon selection, the player's club and other basic information are displayed.
- Customizable result count: by default, the system returns five similar players, but users can modify this to display between three and ten recommendations.
- League filter: a dropdown menu allows filtering results by specific leagues (e.g., Premier league and La liga), with the default set to include all leagues.
- Position matching filter: users can choose whether to display only players from the same position or from any position. By default, all positions are included.
- Age filter: a slider provides filtering based on player age, ranging from 15 to 45 years, with a default range of 15 to 41 years to match common professional career spans.

The output is presented in a sortable table that displays: the name of each recommended player, their similarity percentage (based on cosine similarity), position, league, age, and total number of matches played (expressed in 90-minute equivalents). An example of the system's interface and recommendation results is presented in Figure 6, where users can interactively explore and evaluate players who most closely resemble the statistical profile of a selected individual. The system's design emphasizes transparency and flexibility, making it an effective decision-support tool in the context of talent identification and recruitment.

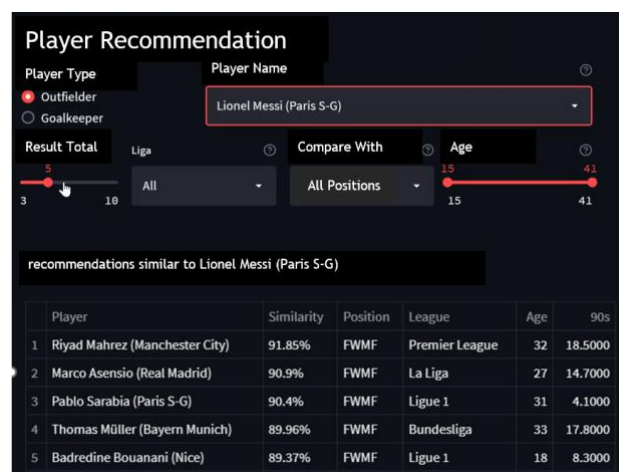


Figure 6. Player recommendation system

3.4. Comparison with previous studies and performance evaluation

Table 3 provides a summary of relevant prior research in the domain of football player prediction and recommendation systems, highlighting their methods, limitations, and the comparative strengths of the current study. While previous studies have addressed player classification and performance prediction, they

often overlook two critical aspects: effective dimensionality reduction and the use of appropriate similarity metrics. Many existing models struggle with high-dimensional data, which can hinder scalability and accuracy. In contrast, this study employs PCA to retain only the most relevant features, preserving up to 98.7% of the original information while reducing model complexity. Additionally, the adoption of cosine similarity provides a more robust measure of similarity by capturing the directional pattern of performance, rather than relying solely on magnitude-based differences, thereby offering a more meaningful comparison between players with varying play intensities but similar styles.

Table 3. Comparison of current study with previous research

Study	Method used	Feature selection	Distance metric	Main limitations	Performance highlight	Comparison advantage
[13]	Regression model for transfer prediction	Manual selection (13 features)	N/A	Focused only on post-transfer performance	Predictive accuracy for transfers	Limited to transfer outcomes, not talent matching
[14]	SRP-CRISP-DM framework	Filter-based	Not specified	General prediction framework	Structured approach	Does not provide similarity-based recommendation
[15]	Multiple ML algorithms (SVM, LDA, and KNN)	Not emphasized	Euclidean distance	Inconsistent feature preprocessing	LDA accuracy: 83.77%	No dimensionality reduction; lower interpretability
[16]	NMF + clustering	Manual, position-based	N/A	Focused on CSL players only	Found 18 player types	Context-limited and unsuited for similarity ranking
[17]	KNN (with multiple distance metrics)	None	City block	Focused on clinical EEG data	City block performed best	Irrelevant domain; insights not directly transferable
Current Study	PCA + KNN with cosine similarity	PCA-based dimensionality reduction (Outfield: 13→10, GK: 4→3)	Cosine similarity	Only linear transformation considered (PCA)	Maintained ≥98% variance, Mahrez 91.85% similarity to Messi	Superior balance of dimensionality reduction, performance, and interpretability

4. CONCLUSION

This study proposes a structured and scalable recommendation model for football player recruitment by integrating PCA for dimensionality reduction with KNN and cosine similarity for performance-based comparison. The PCA effectively reduces feature complexity while preserving up to 98.7% of the original data variance, ensuring that essential performance characteristics are retained. Cosine similarity further enhances the model by capturing the directional alignment of player performance patterns, enabling meaningful comparisons between individuals with different play intensities but similar styles. Key preprocessing steps, such as per-90-minute normalization, consistent player identification across seasons, and position-based categorization, support fair and robust comparisons within a dynamic dataset. The model demonstrated its effectiveness by identifying Riyad Mahrez as the most similar player to Lionel Messi with a 91.85% similarity score, highlighting its practical relevance for scouting and talent identification. This work contributes to the advancement of sports analytics by offering an interpretable, data-driven, and application ready solution for modern football recruitment strategies.

ACKNOWLEDGMENTS

The authors would like to express our appreciation to the Research Institutions of Universitas Gunadarma for their continued facilitation and support throughout the research activities.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Maukar	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Rodiah	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

C : **C**onceptualizationM : **M**ethodologySo : **S**oftwareVa : **V**alidationFo : **F**ormal analysisI : **I**nvestigationR : **R**esourcesD : **D**ata CurationO : Writing - **O**riginal DraftE : Writing - Review & **E**ditingVi : **V**isualizationSu : **S**upervisionP : **P**roject administrationFu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

This study did not involve human participants; therefore, informed consent was not required.

ETHICAL APPROVAL

This study utilized pest dataset and did not involve human or vertebrate animal subjects. Therefore, ethical approval was not required.

DATA AVAILABILITY

This study utilizes player performance data obtained from FBref, a well-established and reputable football statistics platform that compiles comprehensive and advanced player metrics across major professional leagues. The dataset employed in this research is publicly accessible at <https://fbref.com>, thereby allowing for potential replication of the data collection process by other researchers. Detailed descriptions of the data preprocessing procedures, feature selection criteria, and variable transformations applied in this study are available upon request from the corresponding author.




REFERENCES

- [1] L. Lolli *et al.*, "Data analytics in the football industry: a survey investigating operational frameworks and practices in professional clubs and national federations from around the world," *Science and Medicine in Football*, vol. 9, no. 2, pp. 189–198, 2025, doi: 10.1080/24733938.2024.2341837.
- [2] Z. Bai and X. Bai, "Sports big data: management, analysis, applications, and challenges," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/6676297.
- [3] N. Chinthamuri and M. Karukuri, "Data science and applications," *Journal of Data Science and Intelligent Systems*, vol. 1, no. 2, pp. 83–91, 2023, doi: 10.47852/bonviewjdsis3202837.
- [4] W. Bull and M. Faure, "Agents in the sporting field: a law and economics perspective," *International Sports Law Journal*, vol. 22, no. 1, pp. 17–32, 2022, doi: 10.1007/s40318-021-00195-x.
- [5] J. H. Hewitt and O. Karakuş, "A machine learning approach for player and position adjusted expected goals in football (soccer)," *Franklin Open*, vol. 4, 2023, doi: 10.1016/j.fraope.2023.100034.
- [6] V. C. Pantzalis and C. Tjortjis, "Sports analytics for football league table and player performance prediction," in *2020 11th International Conference on Information, Intelligence, Systems and Applications*, 2020, pp. 1–8, doi: 10.1109/IISA50023.2020.9284352.
- [7] T. G. Rumsey, "A statistical look into how common soccer metrics influence expected goal measures in the professional game," *B.S. Thesis*, Department of Mathematical Sciences, Butler University, Indianapolis, United States, 2024.
- [8] J. Mead, A. O'Hare, and P. McMenemy, "Expected goals in football: Improving model performance and demonstrating value," *PLoS ONE*, vol. 18, no. 4 April, 2023, doi: 10.1371/journal.pone.0282295.
- [9] M. Rocchetti, F. Berveglieri, and G. Cappiello, "Football data analysis: the predictive power of expected goals (xG)," in *25th International Conference on Intelligent Games and Simulation, GAME-ON 2024*, 2024, pp. 20–24.
- [10] G. Haddad and D. O'Connor, "Developing players for athlete leadership groups in professional football teams: Qualitative insights from head coaches and athlete leaders," *PLoS ONE*, vol. 17, 2022, doi: 10.1371/journal.pone.0271093.
- [11] Secretary of State for Culture Media and Sport, "A sustainable future - reforming club football governance," *United Kingdom Government*. 2023. [Online]. Available: <https://www.gov.uk/government/publications/a-sustainable-future-reforming-club-football-governance/a-sustainable-future-reforming-club-football-governance>
- [12] M. Musaiqwa, "The role of leadership in managing change," *International Review of Management and Marketing*, vol. 13, no. 6, pp. 1–9, 2023, doi: 10.32479/irmm.13526.
- [13] D. Dinsdale and J. Gallagher, "Transfer portal: accurately forecasting the impact of a player transfer in soccer," *SciSpace*, pp. 1–25, 2020.
- [14] R. P. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Applied Computing and Informatics*, vol. 15, no. 1, pp. 27–33, 2019, doi: 10.1016/j.aci.2017.09.005.




- [15] A. Chavan, "Recruitment of suitable football player by using machine learning techniques," *M.Sc. thesis*, School of Computing, National College of Ireland, Dublin, Ireland, 2019.
- [16] Y. Li, S. Zong, Y. Shen, Z. Pu, M. Á. Gómez, and Y. Cui, "Characterizing player's playing styles based on player vectors for each playing position in the Chinese Football Super League," *Journal of Sports Sciences*, vol. 40, no. 14, pp. 1629–1640, 2022, doi: 10.1080/02640414.2022.2096771.
- [17] C. W. Yean *et al.*, "Analysis of the distance metrics of KNN classifier for EEG signal in stroke patients," in *2018 International Conference on Computational Approach in Smart Systems Design and Applications*, 2018, pp. 1–4, doi: 10.1109/ICASSDA.2018.8477601.
- [18] FBref, "2022-2023 big 5 European Leagues stats," *FBref*. 2022. [Online]. Available: <https://fbref.com/en/comps/Big5/Big-5-European-Leagues-Stats>
- [19] Opta Analyst, "Opta football stats definitions," *Opta Analyst*. 2024. [Online]. Available: <https://theanalyst.com/articles/opta-football-stats-definitions>
- [20] C. Yeung, R. Bunker, R. Umemoto, and K. Fujii, "Evaluating soccer match prediction models: a deep learning approach and feature optimization for gradient-boosted trees," *Machine Learning*, vol. 113, no. 10, pp. 7541–7564, 2024, doi: 10.1007/s10994-024-06608-w.
- [21] A. Ngandjui, "Betting just got easier: the power of machine learning and making predictions," *B.S. thesis*, Department of Computer Science, Salem State University, Salem, Massachusetts, 2021.
- [22] K. Paul and D. Basu, "Data expedition: travel through data preprocessing, EDA and PCA," *Educational Administration: Theory and Practice*, vol. 30, no. 6, pp. 2576–2590, 2024, doi: 10.53555/kuey.v30i6.5828.
- [23] I. Setiawan, R. Gernowo, and B. Warsito, "A systematic literature review on missing values: research trends, datasets, methods and frameworks," *E3S Web of Conferences*, vol. 448, 2023, doi: 10.1051/e3sconf/202344802020.
- [24] A. Palanivinaiyagam and R. Damaševičius, "Effective handling of missing values in datasets for classification using machine learning methods," *Information*, vol. 14, no. 2, 2023, doi: 10.3390/info14020092.
- [25] N. Kovač, K. Ratković, H. Farahani, and P. Watson, "A practical applications guide to machine learning regression models in psychology with Python," *Methods in Psychology*, vol. 11, 2024, doi: 10.1016/j.metip.2024.100156.
- [26] P. Srivastava and N. Kaur, "An overview on data cleaning on real world data," *TechRxiv*, no. 12, pp. 1–10, 2022, doi: 10.36227/techrxiv.21064039.v1.
- [27] R. Rimal, "Cheat sheet: the pandas dataframe object the conceptual model," *Scribd*, pp. 1–10, 2015.
- [28] N. Sahakyan, T. Avetisyan, H. Avetisyan, A. Khan-Aslanyan, and H. Madoyan, "Analyzing soccer 's transfers and predicting footballers ' transfer price," *Research Gate*, pp. 1–79, 2020, doi: 10.13140/RG.2.2.20121.24167.
- [29] Y. He, "Predicting market value of soccer players using linear modeling techniques," *UC Berkeley Statistics*, pp. 1–15, 2014.
- [30] A. P. M. S. Hamdard and A. P. H. Lodin, "Effect of feature selection on the accuracy of machine learning model," *International Journal of Multidisciplinary Research and Analysis*, vol. 6, no. 9, Sep. 2023, doi: 10.47191/ijmra/v6-i9-66.
- [31] Scikit-learn, "PCA," *Scikit Learn*. 2024. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
- [32] M. Gifford and T. Bayrak, "A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression," *Decision Analytics Journal*, vol. 8, 2023, doi: 10.1016/j.dajour.2023.100296.
- [33] S. Lang, R. Wild, A. Isenko, and D. Link, "Predicting the in-game status in soccer with machine learning using spatiotemporal player tracking data," *Scientific Reports*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-19948-1.
- [34] A. Rayhan and R. Kinzler, "Advancing scientific computing with Python's SciPy library," *Research Gate*, pp. 1–19, 2023, doi: RG.2.2.21131.87841.
- [35] S. Cao, "Passing path predicts shooting outcome in football," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-60183-7.
- [36] M. Carey, "Champions league 2023-24: ten players to keep an eye on in the group stage," *The New York Times*. 2023. [Online]. Available: <https://www.nytimes.com/athletic/4822705/2023/09/01/champions-league-10-players-to-watch/>

BIOGRAPHIES OF AUTHORS



Maukar    is a leading professional in the field of computer science. Armed with a bachelor's degree in computer science from the University of Indonesia which was completed in 1991, he continued his education by earning a master's degree in technology management from the Bandung Institute of Technology in 1994. His academic dedication culminated in obtaining a doctoral degree in information technology from Universitas Gunadarma in 2014. Currently, he has conducted a number of researches and published articles in various prestigious journals, making significant contributions to the development of the field of computer science. He can be contacted at email: maukar@staff.gunadarma.ac.id.



Rodiah    is currently Researcher, Lecturer and Vice Head of Postgraduate Academic System Development at Universitas Gunadarma. From 2012 until now, won 9 research grants from Indonesian Directorate General for Higher Education DIKTI (RISTEKDIKTI). Nowadays, authoring 2 books about medical image processing for retinal fundus image, 1 book about retinal biometric. In other hand, she has more than 60 publications within; journals, proceeding and book chapter. She also has more than 20 intellectual property rights (IPR) and 6 patent. She can be contacted at email: rodiah@staff.gunadarma.ac.id.