Position Prediction and Talent Discovery in Football Leagues Using Performance Data

Hasnain Hissam, Syed Bilal Majid, Amirita Dewani*, Memoona Sami, Mariam Memon
Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro.
***Correspondence**: amirita@faculty.muet.edu.pk

Citation | Hissam. H, Majid. S. B, Dewani. A, Sami. M, Memon. M, "Position Prediction and Talent Discovery in Football Leagues Using Performance Data", IJIST, Vol. 07, Issue. 04 pp 3169-3185, December 2025
Received | November 14, 2025 Revised | December 04, 2025 Accepted | December 09, 2025 Published | December 15 2025.

Football has always been dependent on the subjective evaluation of scouts and coaches to find and hire players. Although these methods work to some extent, they usually have restrictions due to human biases, irregularity, and the huge volume of football data. As more data on player performance is made available, data analytics and machine learning represent a chance to introduce objectivity, consistency, and scalability in the recruitment process. This research study suggests a machine learning-based classification model along with a clustering model to classify football players in their main positional roles using statistical performance features. The research is based on the development of models that would help to differentiate among defenders, midfielders, and attackers based on their passing efficiency, contributions to defense, won duels, and attacking indicators. For data extraction, Fbref has been used as the source of data. The player-level data of the 2023-24 season of the Top 5 European Leagues has been extracted using the Python programming language. The data involved various statistical categories addressing all the areas of performance. Position labels were merged with the scraped tables to ensure accurate role mapping. This combination resulted in the creation of an entire dataset with both performance and position features. The dataset was cleaned and prepared using data preprocessing techniques, and selected features were then used in the training process. K-Means Clustering was applied to the PCA-transformed data to cluster similar players based on their playing profile. Different supervised learning algorithms have also been applied, such as Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, and Voting Classifier. The standard evaluation parameters are used to provide a detailed evaluation of the predictive performance. It was found that ensemble algorithms, in particular Random Forest and the Voting Classifier, performed better than single baseline models and were stronger and more reliable in positional classification. The results suggest the potential of machine learning models when recruiting players in football teams and to facilitate and aid expert judgment. This research sets up a systematic, data-driven framework that helps clubs to screen the enormous number of players effectively in a non-subjective manner.

Keywords: Machine Learning; Player Recruitment Analytics; Football Position Prediction Model; Similarity-Based Player Search; Football Talent Discovery.

Introduction:

Data analytics have become a vital part of every field, including sports, and their integration has thus transformed the traditional mechanisms of player evaluation and recruitment. This transition gained extreme global attention because of the popular "Moneyball" approach in Major League Baseball, which showcased the capabilities of statistical models in identifying and uncovering undervalued talent, providing sports teams with a strong competitive advantage [1]. This inspired the adoption of similar frameworks in football, also known as soccer in some regions. A sport that is historically dominated by traditional scouting and subjective judgments. In recent years, the increase in availability of performance statistics has accelerated this shift. Websites like FBref [2] contain comprehensive event-based data, which includes metrics for passing accuracy, defensive actions, goal contributions, and even positional tendencies. Applying machine learning (ML) algorithms to these datasets allows analysts to gain a deeper understanding of a player's style, strengths, and shortcomings, which is hardly possible with the conventional "eye test. One of the biggest and most important financial markets in international sports is the football transfer market. The uncertainty surrounding player adaptability, potential, and fit within a team's tactical system makes recruitment decisions extremely risky, even though European clubs collectively spend billions of euros annually in an attempt to gain a competitive advantage [1]. As a result, the application of Artificial Intelligence (AI) and machine learning (ML) in recruitment has become an area of growing research and practical interest. Recent studies have applied clustering methods to group players with similar characteristics. Some other studies have incorporated classification algorithms to predict suitable positions or roles [2], demonstrating that such tools can support scouts and managers in making more informed decisions. Many recruitment methods are still dependent on subjective assessment and on using a few performance metrics. This way, emerging talent is usually missed by these traditional methods, and it affects their decision-making. To tackle these challenges, there is a need for a scalable and objective framework to identify players' positions and similarities by embedding machine learning models into its publicly available data. This strategy lowers the risks connected with expensive transfers while improving talent discovery [3]. By using clustering and classification techniques to examine player performance metrics, this study expands on the principles of data-driven football analysis. This framework will identify similar patterns for players and predict positions according to their profile and style. It will contribute to the growing need for applying modern scouting techniques over traditional methods, while building the foundation of data-driven football analysis.

Objective of The Study:

The major objectives of this work are given below:

To gather player performance data from multiple open-source online football databases through web scraping and combine them to form a dataset.

To preprocess the data and perform feature engineering to capture appropriate features that represent significant player behaviors.

To apply clustering techniques for the identification of player groups with similar characteristics.

To implement and train various models on the compiled dataset for position prediction.

To evaluate and compare the results using state-of-the-art metrics.

The rest of the paper is organized as follows:

In Section 2, we carry out an in-depth analysis of available literature, where we study the related work done in the field of football analytics with the blend of AI and machine learning in sports. The limitations and the research gap identified from the survey of existing studies and the problem addressed by this research work are discussed in Section 3. In Section 4, we highlight the scope of the current work, clearly stating the parameters that will be

considered. We present the research methodology and discuss the phases adopted in the research framework in Section 5. In Section 6, we elaborate on the research results in detail using state-of-the-art metrics. It shows experimental results, such as classification results and confusion matrices, to demonstrate the performance of the implemented models. And finally, in Section 7, we conclude the findings of this study, showing the shortcomings of the present work, and suggesting improvements for future work.

Literature Review:

Sports analytics is the investigation and modeling of sports performance data, implementing scientific techniques. More specifically, sports analytics refers to the management of structured historical data, the application of predictive analytics models that use this data, and the utilization of information systems to inform decision makers and enable them to assist their organizations in gaining a competitive advantage on the field of play [4].

The use of data analytics in sports can be traced back to the revolutionary concept of Moneyball, which demonstrated how statistical insights could outperform traditional scouting and subjective assessments [5]. While initially applied in baseball, the principles of data-driven recruitment have since been extended to football, where the complexity of performance measurement has driven a need for objective and quantifiable approaches. Past studies have successfully demonstrated specific roles in the field of sports. For example, Carey Analytics [6] investigated the evolution of the modern full-back, emphasizing how player profiles can be quantified through a combination of offensive and defensive contributions. Similarly, further research explored European midfielders' playing styles [7], offering a framework to distinguish between different tactical roles using performance statistics. These studies highlight the potential of statistical profiling in understanding positional dynamics. The authors in [1] demonstrated a deep evaluation of data collection, modeling, and position analysis, which helped them evaluate how better data analytics facilitates both tactical evaluations and long-term recruitment strategies. FBref [8] has become one of the most comprehensive and accessible sources, providing player-level statistics across Europe's "Big 5" leagues. The accessibility of these datasets ensures that analytical approaches are not confined to proprietary data providers, thereby democratizing football analytics. In this context, different researchers have put their efforts into analyzing data and deducing useful information that can be applied practically.

Work contributed by authors in [9] introduces a working pipeline to predict the playing positions on the basis of the statistics of match events. They carry out a feature selection (ANOVA) to discover a small number of discriminative measures (e.g., passing accuracy, duel success rates) and a back propagation neural network (a BP multi-layer perceptron). Their findings indicate that the competitive position classification accuracy can be achieved by features of the event, if carefully chosen, and they explain the trade-offs between the dimensionality of features and the complexity of the model. The paper highlights preprocessing and feature selection as two important processes to achieve stable predictions with match data. Research carried out in [10] introduced a supervised learning strategy designed to be used in in-game position recognition, which is a combination of spatial (field zones/heatmap aggregates) and classic event (traditional) metrics. They test tree-based models and ensemble learners, stating that the integration of the spatial features with the statistics of events consequently helps to make the classification more robust, particularly in cases of players of hybrid roles. Research conducted in [11] provides a comprehensive approach to data collection, data cleanup, data analysis, feature engineering, various classifiers, and a performance evaluation framework. It makes a comparison of classical learners (logistic regression, decision trees, KNN) and documents lessons on feature correlation, class imbalance, and model interpretability. Experimental appendices (code, parameter grids) and suggestions on applied deployment are included as well in this work for reproducibility.

The review paper in [1] is a survey of positional data sources, data collection techniques (tracking systems), and modeling techniques in football analytics. It also combines the use of positional (tracking) data and event data and describes their uses in tactical analysis for player evaluation. The authors highlight methodological issues (data standardization, privacy, and reproducibility) and indicate that improved combinations of positional and event data are required to show the role of players dynamically. In one of the recent works in [12], the authors discuss the use of ML in professional football, including classification, regression (value/performance), and time-series prediction. They compare the typical modeling options (ensemble methods, gradient boosting, neural nets) and provide recommendations on feature engineering and evaluation protocols. The review also includes examples of successful deployments and provides recommendations on best practices to apply to robust model validation and explain ability. A study performed by the author in [3] discusses similarity search through clustering and SVM, along with a few other methods, is used to predict roles. It describes an applied recruitment workflow that uses aggregated statistics to match players by style and predict roles based on their style. The work demonstrates how unsupervised grouping is effective in identifying replacement candidates and shows the superiority of SVM against simple baselines. The author also addresses the practical limitations in combining models with scouting processes. The research study in [13] analyzes zone-based and spatial classification techniques, which combine aggregated heatmap traits with the number of events. It evaluates several classifiers on an open dataset. It finds that adding spatial encodings (which zone a player spends the most time in) makes position prediction on wide and wing positions more accurate. In [14], the authors proposed a new method for assessing a football player's value, utilizing machine learning models to evaluate the relationship between their salaries and specific features.

Using different football websites, data for football players were gathered to perform this research. The research proposed in [15] used a quantitative method to evaluate the football player's market value. This method applied machine learning algorithms to the performance data of football players. The data used by the authors in the experimentation is FIFA 20 video game data. The work estimates players' market values using four regression models that were tested on the full set of features, including linear regression, multiple linear regression, decision trees, and random forests. A research study in [15] suggested a machine learning architecture to forecast the performance of football players based on the match statistics (passes, tackles, and shots). It considers the skill set values of the football player and predicts the performance value. The system proposed by the authors is based on a data-driven approach, and they trained models to generate an appropriate holistic relationship between the players' attribute values, market value, and performance value to be predicted. These values are dependent on the position that the football player plays in and the skills they possess. This study in [16] proposed a machine learning-based classification approach to group soccer players according to their in-game running performance rather than predefined playing positions. Using GPS-derived match data, the authors identified performance clusters that better reflected actual player roles and workloads. The research demonstrated that traditional position labels often fail to capture performance diversity and that data-driven categorization offers deeper tactical insights for coaches and analysts. Work contributed by the authors in [17] aimed to characterize the playing styles of each playing position in the Chinese Football Super League (CSL) matches, integrating a recently adopted Player Vectors framework. The work suggests a multi-dimensional playing style representation expressed as a player (player vectors) based on a vector-based representation of players. Clustering and similarity measures are adopted to discover archetype styles in positions and to confirm the method on CSL data. Their approach emphasizes the usefulness of learned embeddings/vectors to generalize the player behavior beyond individual measurements.

Summarizing the systematic review, it can be concluded that the existing body of research has contributed to using machine learning in football analytics, player valuation, positional analysis, and predicting player performance. Most works focus on classification based on structured sets of players. Some studies are also concerned with regression or descriptive analytics. Nevertheless, existing work is limited, and not many studies employed a combination of various algorithms, including Logistic Regression, Decision Tree, KNN, Random Forest, XGBoost, and Voting Classifier to predict positions, along with using the full set of player performance features and clustering methods. Additionally, the second shortcoming that has been noted in previous research is that it is hard to separate the overlap of roles, especially in midfielders and hybrid attackers, since they have similar statistical profiles.

Research Gap and Problem Statement:

Football scouting often relies on subjective judgments, which can lead to inconsistent and biased decisions. Traditional methods fail to fully capture the complexity of player performance and do not provide a standardized way to compare players across leagues or positions. This creates a gap between available data and its effective use in recruitment.

As discussed earlier, despite the abundance of existing research studies and the development of football analytics systems, several limitations are still not addressed. Most current studies in the literature are based on a minimal number of performance characteristics or a particular league, and their models cannot be generalized. Although previous studies such as [17][12], and [18] have shown that machine learning can be used to classify or cluster players according to their similar playing styles, few studies have been conducted at the point of complete supervision, positional prediction with publicly available datasets such as Fbref. The next shortcoming that has been noted in previous research is that it is hard to separate the overlap of roles, especially in midfielders and hybrid attackers, since they have similar statistical profiles [15]. Models are not always able to provide the same level of accuracy in all positional groups, and midfield roles have historically lower levels of accuracy and recall. There is a dire need for a more reliable system that accurately predicts football player positions using performance statistics, while addressing class imbalance, feature variability, and role overlap.

Research Scope:

In this research work, the main positional types of footballers, i.e., Defender, Midfielder, and Attacker, are considered only in terms of the statistical performance data obtained in the largest European leagues. The current scope is limited to the use of publicly available datasets (FBref) [8]. Also, the analysis is limited to performance-based numerical attributes. We have implemented supervised machine learning classifiers trained on historical season data and employed clustering methods. Currently, sub-positions (e.g., right-back vs. left-back) or tactical positions (e.g., deep-lying playmaker) are outside the scope of this work, although the proposed methodology can be expanded upon in future extensions.

Research Methodology:

The major objective of this research study is to develop and evaluate machine learning models capable of accurately clustering and classifying football players into positional roles, i.e., defenders, midfielders, and attackers, based on performance attributes. To achieve this, a systematic methodology was followed. The methodological framework includes data acquisition, data preprocessing, feature engineering, exploratory data analysis, model training, and evaluation. All the steps were carried out rigorously so that the final model would be accurate, generalizable, and interpretable. Figure 1 illustrates the general workflow of the research and presents the main stages and relationships between them, beginning with the raw data acquisition phase and ending with the system output and model evaluation.

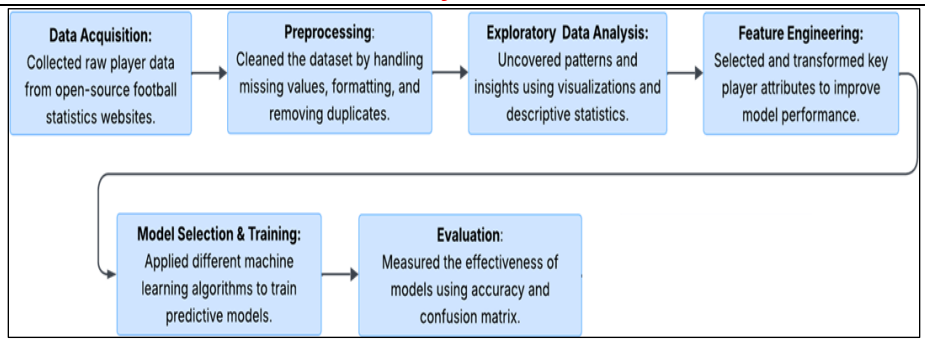
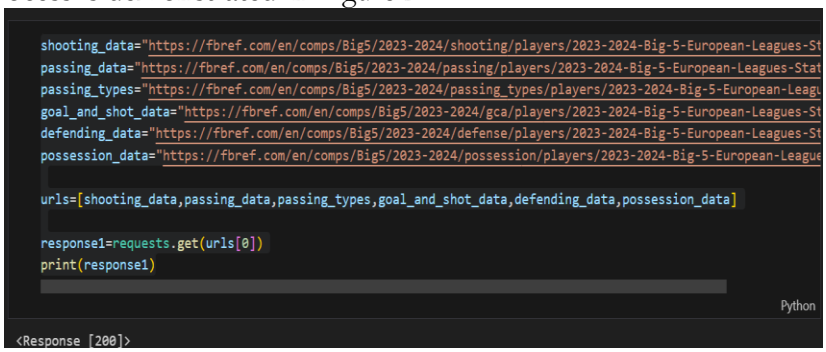


Figure 1. Research Methodology Framework

Data Acquisition:

Source Data Description:

This research paper used the FBref [8] as the source of its data, a reliable open-source football statistics site that gathers player-level data of the 2023-24 season of the Top 5 European Leagues (Premier League, La Liga, Bundesliga, Serie A, and Ligue 1) in detail. The data involved various statistical categories addressing all the areas of performance, including accuracy on passing, shots, defensive duels, and goal contributions. To evaluate and train machine learning models to classify positions, these measures were utilized. The data acquisition process is demonstrated in Figure 2.



```
shooting_data="https://fbref.com/en/comps/Big5/2023-2024/shooting/players/2023-2024-Big-5-European-Leagues-St
passing_data="https://fbref.com/en/comps/Big5/2023-2024/passing/players/2023-2024-Big-5-European-Leagues-Stat
passing_types="https://fbref.com/en/comps/Big5/2023-2024/passing_types/players/2023-2024-Big-5-European-Leagu
goal_and_shot_data="https://fbref.com/en/comps/Big5/2023-2024/gca/players/2023-2024-Big-5-European-Leagues-St
defending_data="https://fbref.com/en/comps/Big5/2023-2024/defense/players/2023-2024-Big-5-European-Leagues-St
possession_data="https://fbref.com/en/comps/Big5/2023-2024/possession/players/2023-2024-Big-5-European-League

urls=[shooting_data,passing_data,passing_types,goal_and_shot_data,defending_data,possession_data]

response1=requests.get(urls[0])
print(response1)
```

Python

```
<Response [200]>
```

Figure 2. Data Acquisition Process

Data was collected programmatically using Python language and several core libraries, including requests to access HTML contents of web pages, BeautifulSoup to extract and break down data within the tables of the HTML, Pandas to clean, combine, and store data in tabular form, and lxml to increase the speed of parsing and work with complicated elements of HTML. Tables were scraped from multiple FBref statistical categories, including Standard Player Stats (goals, assists, appearances), Shooting Stats (shots, shots on target, expected goals (xG)), Passing Stats (key passes, accuracy, expected assists (xA)), Defensive Stats (tackles, interceptions, and clearances) and Possession and Progression (carries, dribbles, and progressive passes). A secondary dataset, which contained player position labels, was also extracted so that it could be merged with the scraped tables to make sure role mapping was accurate. This combination resulted in the creation of an entire dataset with both performance and position features (Defender, Midfielder, Attacker). The screenshot of the raw dataset is shown in Figure 3. Even though the collected data set represented the Top European leagues along with the significant features and players' information, it was currently limited to one season, i.e., 2023-24.

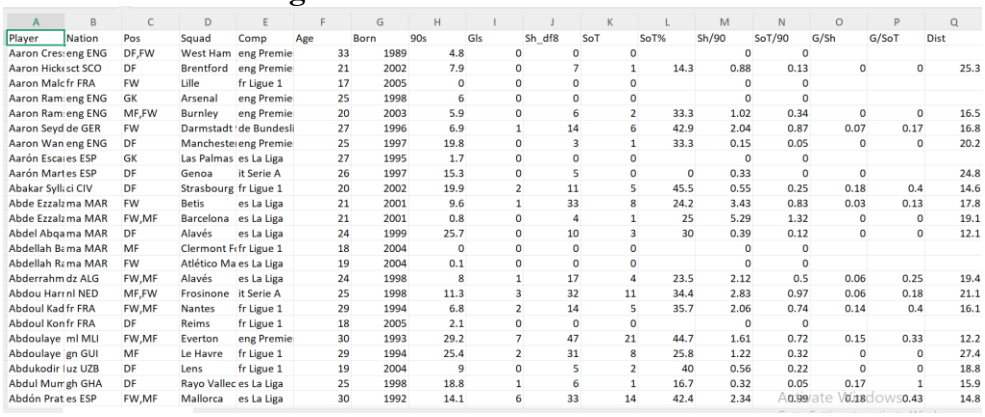
Data Preprocessing:

Data Integration Process:

After web scraping, the fetched tables were amalgamated together using unique player identifiers, and as a result, a single, comprehensive table of player performance was formed, as shown in Figure 4.

	Player	Nation	Pos	Squad	Comp	Age	Born	90s	Gls	Sh	...	Dist	FK	PK	PKatt	xG	np
0	Max Aarons	eng ENG	DF	Bournemouth	eng Premier League	23	2000	13.7	0	2	...	23.9	0	0	0	0.0	
1	Brenden Aaronson	us USA	MF,FW	Union Berlin	de Bundesliga	22	2000	14.1	2	18	...	18.4	0	0	0	2.0	
2	Paxten Aaronson	us USA	MF	Eint Frankfurt	de Bundesliga	19	2003	1.1	0	2	...	15.1	0	0	0	0.1	
3	Keyliane Abdallah	fr FRA	FW	Marseille	fr Ligue 1	17	2006	0.0	0	0	...		0	0	0	0.0	
4	Yunis Abdelhamid	ma MAR	DF	Reims	fr Ligue 1	35	1987	30.9	4	21	...	15.0	0	1	1	3.4	
5	Salis Abdul Samed	gh GHA	MF	Lens	fr Ligue 1	23	2000	16.9	0	5	...	21.6	0	0	0	0.8	
6	Nabil Aberdin	fr FRA	DF	Getafe	es La Liga	20	2002	2.0	0	0	...		0	0	0	0.0	
7	Laurent Abergel	fr FRA	MF	Lorient	fr Ligue 1	30	1993	31.8	2	26	...	26.5	0	0	0	1.1	

Figure 3. Screenshot of Raw Dataset



Player	Nation	Pos	Squad	Comp	Age	Born	90s	Gls	Sh_df8	SoT	SoT%	Sh/90	SoT/90	G/Sh	G/SoT	Dist	
Aaron Cres	eng ENG	DF,FW	West Ham	eng Premie	33	1989	4.8	0	0	0		0	0				
Aaron Hicks	sct SCO	DF	Brentford	eng Premie	21	2002	7.9	0	7	1	14.3	0.88	0.13	0	0	25.3	
Aaron Malc	fr FRA	FW	Lille	fr Ligue 1	17	2005	0	0	0	0		0	0				
Aaron Ram	eng ENG	GK	Arsenal	eng Premie	25	1998	6	0	0	0		0	0				
Aaron Ram	eng ENG	MF,FW	Burnley	eng Premie	20	2003	5.9	0	6	2	33.3	1.02	0.34	0	0	16.5	
Aaron Seyd	de GER	FW	Darmstadt	de Bundesli	27	1996	6.9	1	14	6	42.9	2.04	0.87	0.07	0.17	16.8	
Aaron Wan	eng ENG	DF	Manchester	eng Premie	25	1997	19.8	0	3	1	33.3	0.15	0.05	0	0	20.2	
Aarón Escal	es ESP	GK	Las Palmas	es La Liga	27	1995	1.7	0	0	0		0	0				
Aarón Mart	es ESP	DF	Genoa	it Serie A	26	1997	15.3	0	5	0	0	0.33	0	0		24.8	
Abakar Syll	ci CIV	DF	Strasbourg	fr Ligue 1	20	2002	19.9	2	11	5	45.5	0.55	0.25	0.18	0.4	14.6	
Abde Ezzal	ma MAR	FW	Betis	es La Liga	21	2001	9.6	1	33	8	24.2	3.43	0.83	0.03	0.13	17.8	
Abde Ezzal	ma MAR	FW,MF	Barcelona	es La Liga	21	2001	0.8	0	4	1	25	5.29	1.32	0	0	19.1	
Abdel Abq	ma MAR	DF	Alavés	es La Liga	24	1999	25.7	0	10	3	30	0.39	0.12	0	0	12.1	
Abdellah B	ma MAR	MF	Clermont F	fr Ligue 1	18	2004	0	0	0	0		0	0				
Abdellah R	ma MAR	FW	Atlético Ma	es La Liga	19	2004	0.1	0	0	0		0	0				
Abderrahm	dz ALG	FW,MF	Alavés	es La Liga	24	1998	8	1	17	4	23.5	2.12	0.5	0.06	0.25	19.4	
Abdou Harr	nl NED	MF,FW	Frosinone	it Serie A	25	1998	11.3	3	32	11	34.4	2.83	0.97	0.06	0.18	21.1	
Abdoul Kad	fr FRA	FW,MF	Nantes	fr Ligue 1	29	1994	6.8	2	14	5	35.7	2.06	0.74	0.14	0.4	16.1	
Abdoul Kon	fr FRA	DF	Reims	fr Ligue 1	18	2005	2.1	0	0	0		0	0				
Abdoulaye	ml MLI	FW,MF	Everton	eng Premie	30	1993	29.2	7	47	21	44.7	1.61	0.72	0.15	0.33	12.2	
Abdoulaye	gn GUI	MF	Le Havre	fr Ligue 1	29	1994	25.4	2	31	8	25.8	1.22	0.32	0	0	27.4	
Abdukodir	uz UZB	DF	Lens	fr Ligue 1	19	2004	9	0	5	2	40	0.56	0.22	0	0	18.8	
Abdul Mur	gh GHA	DF	Rayo Vallec	es La Liga	25	1998	18.8	1	6	1	16.7	0.32	0.05	0.17	1	15.9	
Abdón Prat	es ESP	FW,MF	Mallorca	es La Liga	30	1992	14.1	6	33	14	42.4	2.34	0.99	0.18	0.43	14.8	

Figure 4. Screenshot of Merged Dataset

Adding Position Labels:

In this step, the secondary dataset was used to map each player's name and unique identifier with a position label. The following tasks were performed on the data to achieve this:

Original FBref positions (e.g., "DF", "MF", "FW") were mapped to generalized classes.

Role names were standardized across all tables.

Ambiguous cases were resolved, such as players with multiple secondary positions.

Duplicate data was eliminated.

This process helped establish the final target variable, which was subsequently used in supervised model training.

Age	Born	90s	Gls	Sh_df8	...	PrgDist_df6	PrgC	1/3_df6	CPA	Mis	Dis	Rec	PrgR	position	sub_position
21.0	2002.0	7.9	0.0	7.0	...	505.0	9.0	3.0	1.0	14.0	5.0	225.0	13.0	Defender	Left-Back
20.0	2003.0	5.9	0.0	6.0	...	168.0	8.0	5.0	4.0	11.0	10.0	142.0	17.0	Midfield	Central Midfield
27.0	1996.0	6.9	1.0	14.0	...	117.0	7.0	3.0	3.0	26.0	12.0	214.0	35.0	Attack	Centre-Forward
20.0	2002.0	19.9	2.0	11.0	...	3215.0	8.0	11.0	0.0	14.0	7.0	962.0	2.0	Defender	Centre-Back
24.0	1999.0	25.7	0.0	10.0	...	1418.0	7.0	6.0	0.0	15.0	1.0	499.0	1.0	Defender	Centre-Back
25.0	1998.0	11.3	3.0	32.0	...	1221.0	43.0	31.0	8.0	28.0	7.0	378.0	56.0	Midfield	Central Midfield
29.0	1994.0	6.8	2.0	14.0	...	1179.0	41.0	25.0	13.0	17.0	13.0	327.0	73.0	Attack	Right Winger

Figure 5. Preprocessed Dataset (After Adding Position Labels)

Exploratory Data Analysis (EDA):

For better data visualization and understanding, we performed an exploratory data analysis. The EDA process was carried out to capture and analyze the data patterns and determine trends related to leagues and playing positions of players.

League Distribution Analysis:

Figure 6 helps assess the player's balance and depicts the player's distribution across the top five leagues. The representation helps identify which leagues are most represented in the dataset. It gives a visual insight into the dataset's breadth and focus across global competitions.

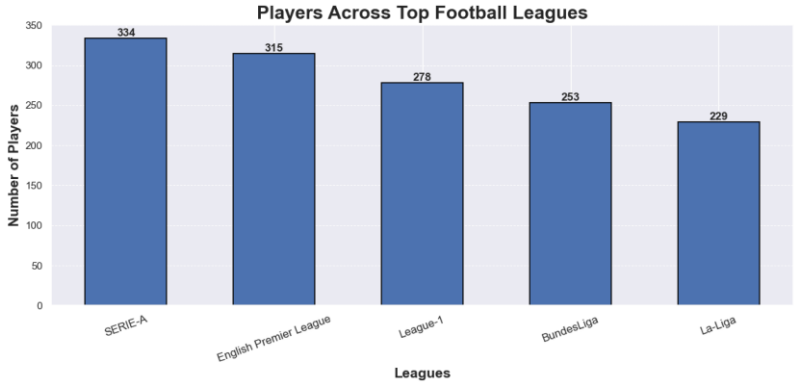


Figure 6. League Distribution Analysis for Players

Match Minutes and Age Distribution:

Secondly, to determine the levels of experience of players, we analyzed the data for the matches played by them. The histogram in Figure 7 represents the number of full match equivalents (90s) played by each player. Most players have played between 10 and 25 full matches, suggesting consistent participation across the dataset, with fewer players at the extremes.

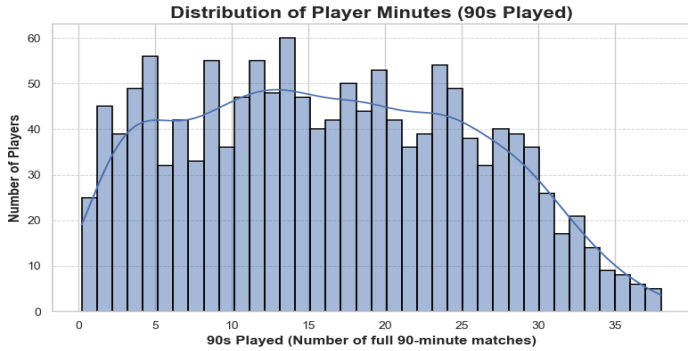


Figure 7. Player Distribution Data

Age Distribution Across Leagues:

The diagram in Figure 8 displays a box plot of player age ranges across different leagues. This was an analysis of the average range of the ages of the players in various leagues. The purpose was twofold, i.e., to identify whether younger or senior players dominate certain leagues and to examine role-specific age trends, such as peak ages for midfielders or attackers.

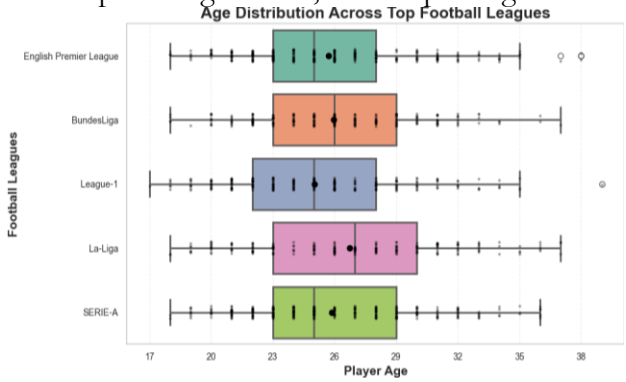


Figure 8. Age Distribution Across Leagues

Average Playing Time by League:

The bar chart in Figure 9 estimates the mean minutes of matches per league. It helped identify the leagues that had higher rotation (decreased average playing time), competitions that the players from the two teams play full games, and whether playing time affects performance measures.

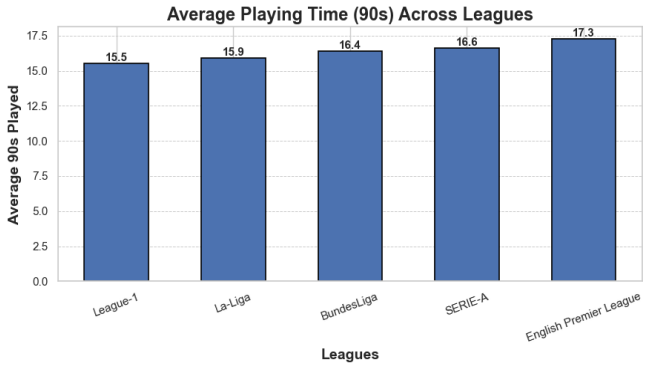


Figure 9. Average Playing Time Plot Across Top Leagues

Player Position Count:

Next, a distribution plot was generated (as shown in Figure 10) representing the count of defenders, midfielders, and attackers. This chart categorizes players by their primary positions on the field. It helps analyze team structure and positional trends in the data.

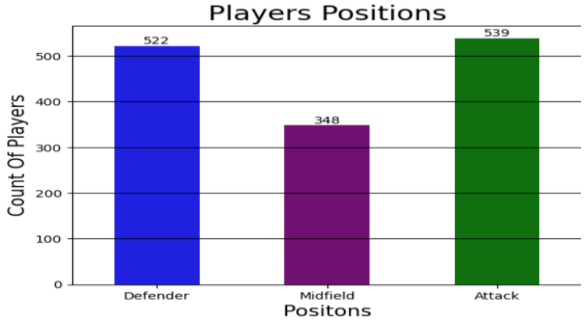


Figure 10. Player Position Count Diagram

Player's Sub-Position Distribution:

Sub-position categories (e.g., CB, RB, CAM, LW, etc.) were studied and plotted (as represented in Figure 11) in the process of exploratory data analysis in order to enrich the knowledge further. This clarified the tactical diversity within each main position and potential overlaps that could challenge model predictions.

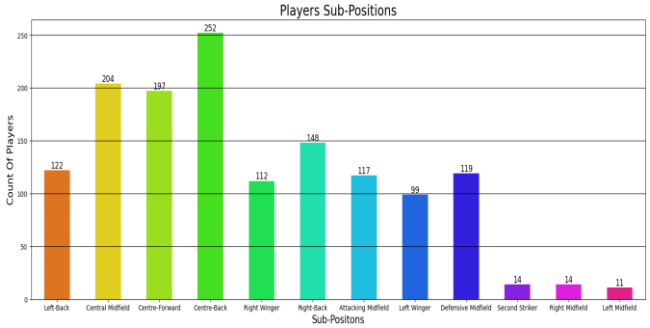


Figure 11. Sub-Position Categories of Players

Feature Engineering:

Numerical Feature Selection:

Non-numeric data (e.g., player names, clubs) were excluded. Numerical measures like passes, tackles, interceptions, and progressive carries were the only measures that were retained with a custom function select_numerical_columns ().

Standardization:

StandardScaler from sklearn in Python was used to standardize all the features (mean = 0, standard deviation = 1). Scaler object was stored as scaler.pkl to make it reproducible.

Clustering-Based Feature Understanding:

The optimum number of clusters (the value of k) was calculated by means of the Within-Cluster-Sum-of-Squares (WCSS) and the Elbow Method. The Principal Component Analysis (PCA) helped to bring down the data from high-dimensions into fewer components so that around 95 percent of the total variance is maintained. The plot is represented in Figure 12. The diagram highlights an increase in variance for initial components, whereas the curve gets flattened after 37 components, indicating that most of the variation in the data is captured before this point. The optimal components are also highlighted in the graph.

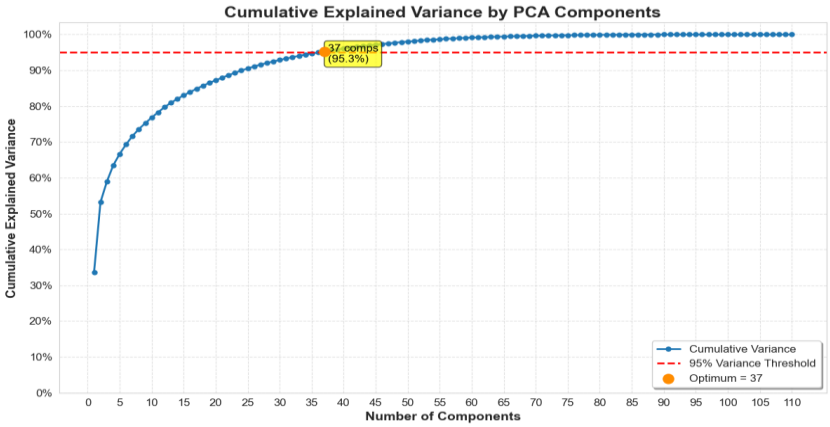


Figure 12. Elbow Curve & PCA Explained Variance Plot

PCA Transformation and K-Means Clustering:

K-Means Clustering was applied to PCA-transformed data to cluster the similar players based on their playing profiles. The scatter plot of player clusters using the K-Means algorithm is shown in Figure 13. The findings confirmed that clusters were similar to the real-life football positions.

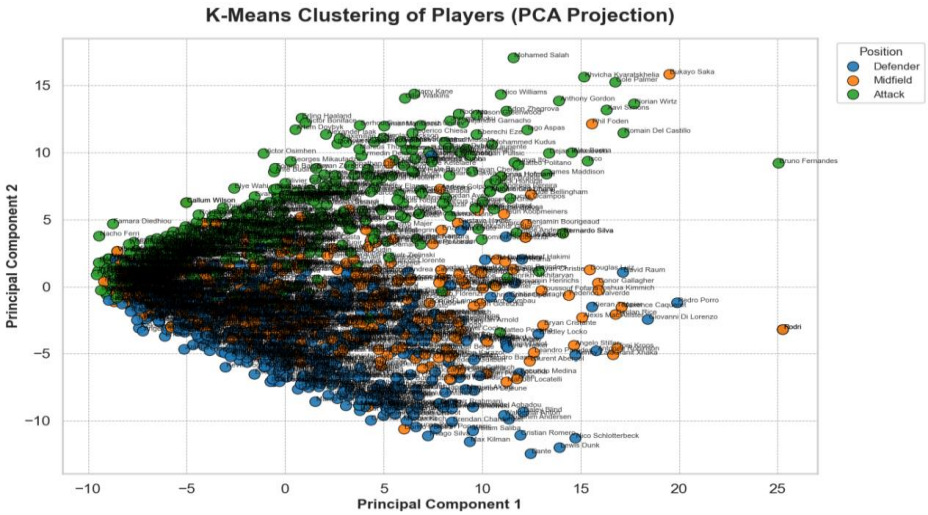


Figure 13. Scatter Plot of Player Clusters

Model Development and Configurations:

For model development and configuration, the Python programming language was used because it has a vast array of data analysis and machine learning packages. Jupyter Notebook was used as the primary development environment to execute, test, and interactively visualize the code. Six supervised learning algorithms, including machine learning and ensembles, were implemented for comparative analysis, i.e.

Logistic Regression: A linear baseline classifier.
Decision Tree: A non-linear classifier with recursive partitioning.
K-Nearest Neighbors (KNN): A distance-based classifier relying on neighborhood voting.
Random Forest: An ensemble of decision trees, reducing variance through bagging.
XGBoost: A gradient boosting algorithm optimized for speed and performance.
Voting Classifier: A hybrid ensemble combining Logistic Regression, Random Forest, and XGBoost using majority voting.

Default hyperparameters from the scikit-learn package and XGBoost were used initially, with light tuning applied for ensemble models to prevent overfitting. The models were optimized using Grid Search CV (for Logistic Regression, Decision Tree, KNN, and XGBoost) and Randomized Search CV for Random Forest due to its larger parameter space. This tuning enhanced classification performance and stability. Specifically, for Random Forest, we kept n_estimators= '200' to get a balance between computation cost and model performance. Even though the study applied PCA for dimensionality reduction to a good extent, we used max_features= 'sqrt' to select the most optimal features without causing the model to overfit. For XGBoost, the learning rate was equal to 0.01 so that the model can learn essential patterns at a feasible pace, and n_estimators= '200'. The rest of the features were tuned lightly, using Randomized Search CV and grid search CV as stated earlier.

Model Evaluation:

Evaluation metrics have been utilized to evaluate and carry out a comparative analysis of the models. In this work, we used Accuracy to estimate the overall correct predictions, Precision to determine the number of correct positive predictions, F1 score to analyze the proportion of actual positives correctly identified, and recall to determine the proportion of all actual positives that were classified correctly as positives. Furthermore, for each classifier, various confusion matrices were generated to visualize correct and incorrect classification of player roles. The heatmaps provided a clear evaluation of each model's performance and its classification of player positions. These metrics were used to evaluate the model's robustness to make effective and fair decisions across different player categories.

Results and Discussion:

This section presents the empirical findings from the application of six supervised learning classifiers, i.e., Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, XGBoost, and a Voting Classifier, for the task of football player role classification using the dataset compiled for this work. As illustrated earlier, each model was evaluated using accuracy, precision, recall, and F1-score. The results are summarized in classification reports and visualized through confusion matrices.

Logistic Regression:

Logistic Regression provided a solid baseline performance with an overall accuracy of 0.79. It demonstrated high precision and recall for defenders (0.79 and 0.88, respectively) and attackers (0.79 and 0.87), but midfielders were less accurately classified (recall = 0.58). The classification report and confusion matrix for this model are given in Figures 14 and 15, respectively. This reflects the overlapping characteristics of midfielders with both defensive and attacking roles.

	precision	recall	f1-score	support
Defender	0.79	0.88	0.83	111
Midfield	0.80	0.58	0.67	103
Attacker	0.79	0.87	0.83	139
accuracy			0.79	353
macro avg	0.79	0.78	0.78	353
weighted avg	0.79	0.79	0.78	353

Figure 14. Classification Report for Logistic Regression

Decision Tree:

The Decision Tree achieved an accuracy of 0.76, with strong classification for defenders, demonstrating a precision of 0.87, and attackers with a recall of 0.87. However, its performance dropped for midfielders with precision at 0.67 and recall at 0.57. This inconsistency is indicative of the overfitting tendencies common in single-tree models. Even with high performance for defenders and attackers, this algorithm showed little generalization capabilities as compared to ensemble approaches. The classification metrics and confusion matrix are presented in Figure 16 and Figure 17, respectively.

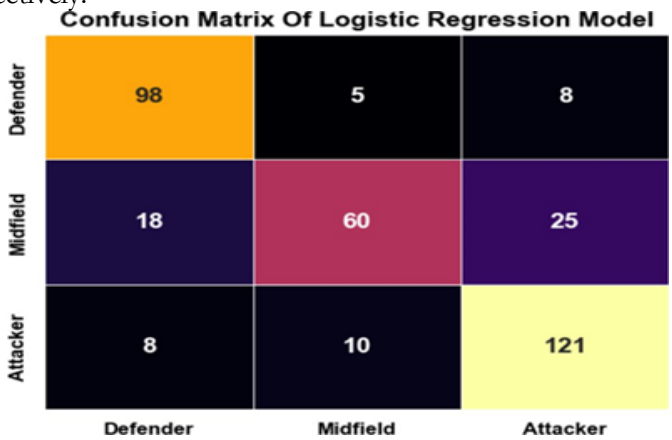


Figure 15. Confusion Matrix for Logistic Regression

	precision	recall	f1-score	support
Defender	0.87	0.81	0.84	111
Midfield	0.67	0.57	0.62	103
Attacker	0.75	0.87	0.81	139
accuracy			0.76	353
macro avg	0.76	0.75	0.75	353
weighted avg	0.76	0.76	0.76	353

Figure 16. Classification Report for Decision Tree

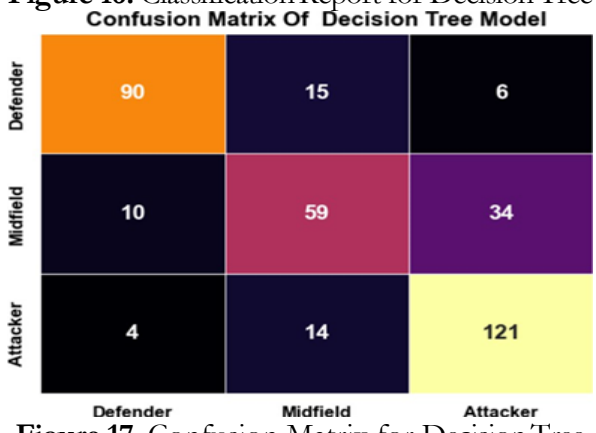


Figure 17. Confusion Matrix for Decision Tree

K-Nearest Neighbors (KNN):

KNN produced a similar set of results, working exceptionally well for attackers and defenders but considerably less for midfielders. KNN classification results and confusion matrix are displayed in Figures 18 and 19, respectively. The confusion matrix uncovered consistent defender classification, but major overlap between midfielders and attackers. These results revealed KNN's sensitivity to feature scaling and local neighborhood decisions in complex football statistics.

	precision	recall	f1-score	support
Defender	0.73	0.87	0.80	111
Midfield	0.74	0.51	0.61	103
Attacker	0.81	0.86	0.84	139
accuracy			0.76	353
macro avg	0.76	0.75	0.75	353
weighted avg	0.76	0.76	0.76	353

Figure 18. Classification Report for KNN

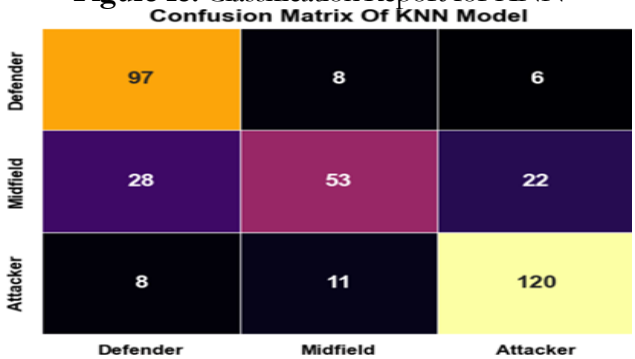


Figure 19. Confusion Matrix for KNN

Random Forest:

Random Forest proved to be one of the best among all other models, with an accuracy of 0.82. It created a balance relation between precision and recall across all roles, with particularly robust classification of defenders (with precision 0.82 and recall = 0.87) and attackers (with precision= 0.82 and recall = 0.88). Random forest reduced variance for single-tree models to make it more reliable for player classification due to its ensemble learning capability. These results are shown in Figures 20 and 21.

	precision	recall	f1-score	support
Defender	0.82	0.87	0.85	111
Midfield	0.82	0.68	0.74	103
Attacker	0.82	0.88	0.85	139
accuracy			0.82	353
macro avg	0.82	0.81	0.81	353
weighted avg	0.82	0.82	0.82	353

Figure 20. Classification Report for Random Forest

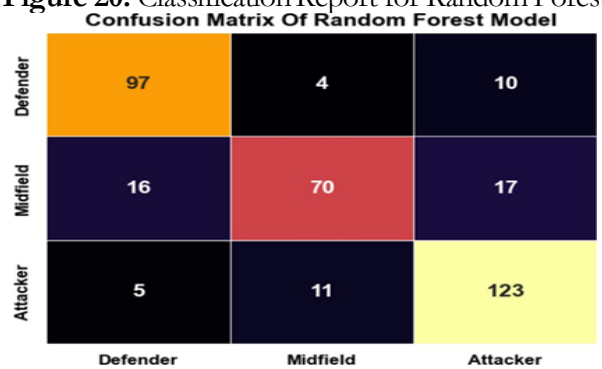


Figure 21. Confusion Matrix for Random Forest

XGBoost:

The XGBoost ensemble model demonstrated the highest accuracy among all individual models, with an accuracy of 0.83. It illustrated excellent performance for defenders (precision= 0.84, recall = 0.92) and attackers (precision=0.83, recall = 0.87), though midfielders again exhibited comparatively lower recall (0.68). Non-linear relationships were effectively captured

by their gradient boosting framework, leading to robust generalization. The performance evaluation of XGBoost is shown in Figures 22 and 23.

Voting Classifier:

This method used a technique to combine Random Forest, logistic Regression, and XGBoost together, achieving an accuracy of 0.81. While not surpassing the overall performance results of XGBoost, this approach maintained a balanced relationship across all roles, with defenders achieving an F1 score of 0.85, attackers achieving an F1 score of 0.84, and midfielders achieving an F1 score of 0.72. The classification report and confusion matrix of this classifier are shown in Figures 24 and 25, respectively. For consistent classification, the ensemble model reduced the variance present in individual models.

	precision	recall	f1-score	support
Defender	0.84	0.92	0.88	111
Midfield	0.81	0.68	0.74	103
Attacker	0.83	0.87	0.85	139
accuracy			0.83	353
macro avg	0.83	0.82	0.82	353
weighted avg	0.83	0.83	0.83	353

Figure 22. Classification Report for XGBoost

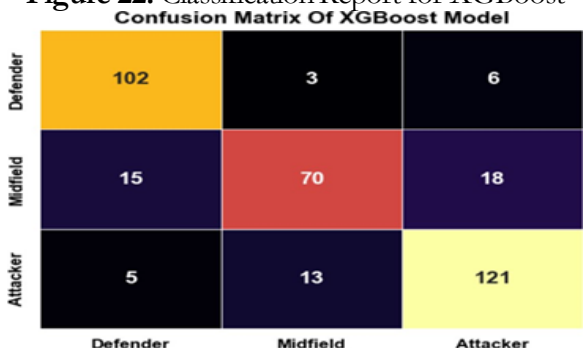


Figure 23. Confusion Matrix for XGBoost

	precision	recall	f1-score	support
Defender	0.82	0.88	0.85	111
Midfield	0.81	0.65	0.72	103
Attacker	0.81	0.88	0.84	139
accuracy			0.81	353
macro avg	0.81	0.80	0.80	353
weighted avg	0.81	0.81	0.81	353

Figure 24. Classification Report for Voting Classifier

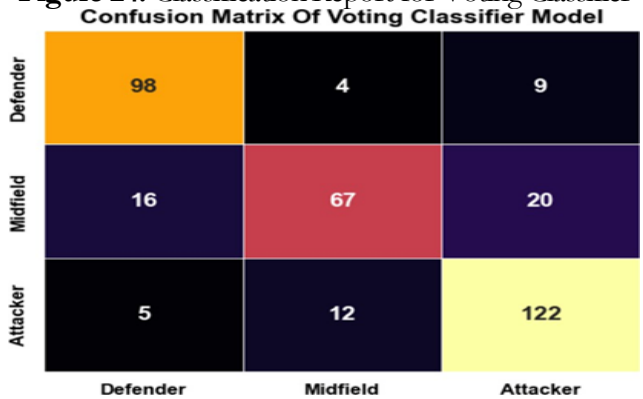


Figure 25. Confusion Matrix for Voting Classifier

The comparative analysis of all the applied models presents several key findings.

Firstly, it demonstrates that precision and recall metrics had consistently high values for both attackers and defenders, suggesting that these two classes have more unique statistical profiles.

Midfielders were the most challenging role to classify across all models, with lower recall values (typically ranging from 0.51–0.68). This supports the observation that midfielders often share hybrid attributes with both defenders and attackers.

Among single models, XGBoost achieved the best performance (accuracy=0.83), demonstrating the effectiveness of gradient boosting for structured sports data.

Random Forest (accuracy = 0.82) also provided reliable results, confirming the strength of ensemble methods.

The Voting Classifier offered balanced generalization and stability across player categories.

Taken together, the results confirm that the performance of ensemble ML techniques such as Random Forest, XGBoost, and Voting Classifier surpassed simpler classifiers like Logistic Regression, Decision Tree, and KNN in football player role classification.

Conclusion and Future Work:

This research examined the use of machine learning and clustering algorithms to classify football players based on their performance metrics and other statistics. By applying a range of algorithms such as Random Forests, Logistic Regression, k-Nearest Neighbors, Decision Trees, and XGBoost on a dataset gathered via automated web scraping and subsequent pre-processing steps. Based on the results, we conclude that integrating multiple models worked better in comparison to single models. Ensemble methods consistently showed higher accuracies. It proves that sports analytics could benefit from ensemble approaches, further confirming their applicability in situations where data is of a diverse nature.

The ensemble model predictions suggested by this study can support scouting decisions, reduce individual model bias, and improve robustness in player evaluation and recruitment strategies. The results of this study will also have a profound impact on various stakeholders of the football and sports analytics ecosystem, such as; Football Clubs and Scouts to make more informed recruitment decisions based on data-based insights as opposed to subjective judgments and Coaches, and Technical Staff members to have a greater insight into the strengths and weaknesses of players for making strategies and assigning roles, so that the available talent is used optimally.

Although this research produced encouraging results, a few limitations still persist. The major issue is the limitation of the dataset, which comprises a single season of football statistics. Hence, it may limit the generalizability of the models across multiple seasons or different leagues. Furthermore, it did not entertain contextual or unstructured data sources such as match commentary, video analysis, or sensor-based tracking, focusing only on structured numerical data. Using these modalities, the model can enhance its predictive nature and provide richer player profiling. Therefore, we can extend this research in multiple directions. First, we suggest adding multi-season and cross-league datasets to enhance the generalizability of the models. Second, more feature engineering techniques could be applied to further extract the usually missed performance indicators. Third, with a diverse dataset, the use of advanced deep learning models such as graph neural networks (GNNs) and recurrent neural networks (RNNs) can improve the overall results. The deployment of these pre-trained models could bridge the gap between football clubs and players' analytics. Additionally, the real-world sports data is huge in volume and inherently unbalanced in nature with respect to different target classes. Therefore, one of the other possible future extensions of this work can be the implementation and analysis of commonly used data balancing techniques, including over- and undersampling, for handling the class imbalance problem. Some of the significant data balancing techniques that the future research community can explore are SMOTE (Synthetic minority oversampling approach),

ADASYN (Adaptive Synthetic Sampling), Borderline-SMOTE, and Safe-Level SMOTE.

In summary, this work has established a strong baseline for the applicability of ensembled and machine learning techniques in football analytics on a curated dataset, demonstrating both its immediate benefits and its potential for future expansion. By integrating multiple datasets from various platforms and further refining them, these techniques can significantly change the dynamics of sports analytics.

Acknowledgement: We are extremely grateful to the Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, for providing all the technical support and resources to accomplish this project.

Funding Details: This research study received no external funding. All the resources consumed were internally available at the Institute of Information and Communication Technologies, Mehran University of Engineering and Technology, Jamshoro.

Data Availability Statement: Most of the data used by this research work are included as part of this manuscript. Complete data can be provided by the corresponding author on a valid request.

Author's Contribution: All the authors of this work contributed equally to study conception, methodology design, implementation, result reporting and validation, draft writing, editing, proofreading, and formatting.

Conflict of Interest: All the authors declare no conflict of interest for publishing this manuscript in IJIST.

References:

[1] D. R. Daniel Memmert, "Data Analytics in Football: Positional Data Collection, Modelling and," Routledge. Accessed: Jan. 06, 2026. [Online]. Available: https://www.routledge.com/Data-Analytics-in-Football-Positional-Data-Collection-Modelling-and-Analysis/Memmert-Raabe/p/book/9781032532479

[2] K. A. Morgan, R. Godasu, and E. S. Grant, "Player Position Binary Classification Model," *Proc. - 2024 3rd Int. Conf. Comput. Appl. Technol. CCAT 2024*, pp. 13–17, 2024, doi: 10.1109/CCAT64370.2024.00011.

[3] T. Cannon, "Beyond the Eye Test: Improving Football Recruitment Through The Use Of Clustering And Support Vector Machines: Data Science Report," *NORMA eResearch @NCI Libr.*, 2023, [Online]. Available: https://norma.ncirl.ie/6918/

[4] E. Morgulev, O. H. Azar, and R. Lidor, "Sports analytics and the big-data era," *Int. J. Data Sci. Anal. 2018 54*, vol. 5, no. 4, pp. 213–222, Jan. 2018, doi: 10.1007/S41060-017-0093-7.

[5] J. Scott Armstrong, "Predicting Job Performance: The Moneyball Factor," foresight . Accessed: Jan. 06, 2026. [Online]. Available: https://www.researchgate.net/publication/254416540_Predicting_Job_Performance_The_Moneyball_Factor

[6] "Quantifying Player Profiles: The Evolution of the Full-Back – Carey Analytics." Accessed: Jan. 06, 2026. [Online]. Available: https://careyanalytics.wordpress.com/2018/02/22/quantifying-player-profiles-the-evolution-of-the-full-back/

[7] "A closer look into European Midfielder Playing Styles – Carey Analytics." Accessed: Jan. 06, 2026. [Online]. Available: https://careyanalytics.wordpress.com/2020/04/30/a-closer-look-into-european-midfielder-playing-styles/

[8] "2023-2024 Big 5 European Leagues Stats | FBref.com." Accessed: Jan. 10, 2026. [Online]. Available: https://fbref.com/en/comps/Big5/2023-2024/2023-2024-Big-5-European-Leagues-Stats

[9] Z. P. B. Zeng, "A Machine Learning Model to Predict Player's Positions based on

Performance," *ic Sport.*, 2021, [Online]. Available: https://www.scitepress.org/Papers/2021/106533/106533.pdf

[10] Sam Brown & Abdulla Kerimov, "Identifying Current Position of a Player Using Machine Learning Approach - NHSJS." Accessed: Jan. 07, 2026. [Online]. Available: https://nhsjs.com/2025/identifying-current-position-of-a-player-using-machine-learning-approach/#google_vignette

[11] J. S. D. Chandra B, "Prediction of Football Player Performance Using Machine Learning Algorithm," *Researchgate*, 2024, doi: 10.21203/rs.3.rs-3995768/v1.

[12] C. T. Diego Moya, "Machine Learning Applied to Professional Football: Performance Improvement and Results Prediction," *Mach. Learn. Knowl. Extr*, vol. 7, no. 3, p. 85, 2025, doi: https://doi.org/10.3390/make7030085.

[13] U. Di Giacomo, F. Mercaldo, A. Santone, and G. Capobianco, "Machine Learning on Soccer Player Positions," *Int. J. Decis. Support Syst. Technol.*, vol. 14, no. 1, 2022, doi: https://doi.org/10.4018/IJDSST.286678.

[14] P. T. Chenyao Li, Stylianos Kampakis, "Machine Learning Modeling to Evaluate the Value of Football Players," *arXiv:2207.11361*, 2022, [Online]. Available: https://arxiv.org/abs/2207.11361

[15] R. Pariath, S. Shah, A. Surve, and J. Mittal, "Player Performance Prediction in Football Game," *Proc. 2nd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2018*, pp. 1148–1153, Sep. 2018, doi: 10.1109/ICECA.2018.8474750.

[16] S. van der Z. Michel de Haan, "Beyond Playing Positions: Categorizing Soccer Players Based on Match-Specific Running Performance Using Machine Learning," *J. Sport. Sci. Med.*, vol. 24, pp. 565–577, 2025, [Online]. Available: https://www.jssm.org/researchjssm-24-565.xml.xml

[17] 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," *J. Sports Sci.*, vol. 40, no. 14, pp. 1629–1640, Jul. 2022, doi: 10.1080/02640414.2022.2096771;SUBPAGE:STRING:ACCESS.

[18] P. S. Hashir Sayeed, "A Machine Learning Framework to Scout Football Players," *NORMA eResearch @NCI Libr.*, 2023, [Online]. Available: https://norma.ncirl.ie/7265/