Artifical Intelligence & Data Engineering

Design Project Report

Title

Building A Large Language Model Based Artificial General Intelligence (AGI) System

Prepared By

15010000 Elon MUSK

15010001 Bill GATES

Supervisor

Assoc. Prof. Dr. A. Cüneyd TANTUĞ

May, 2024

CONTENTS

[1 INTRODUCTION 1](#_Toc166838321)

[2 BACKGROUND 2](#_Toc166838322)

[3 SYSTEM REQUIREMENTS 3](#_Toc166838323)

[3.1 Design Constraints and Relevant Engineering Standards 3](#_Toc166838324)

[3.2 Functional Requirements 3](#_Toc166838325)

[3.3 Non-Functional Requirements 3](#_Toc166838326)

[3.4 Evaluation Criteria 3](#_Toc166838327)

[4 SYSTEM ARCHITECTURE 4](#_Toc166838328)

[5 RESULTS AND EVALUATION 5](#_Toc166838329)

[6 CONCLUSIONS AND FUTURE WORKS 6](#_Toc166838330)

[7 REFERENCES 7](#_Toc166838331)

SUMMARY

A one-page summary of the project must be provided in this section.

# INTRODUCTION

Tactical aspects of football, the most popular sport worldwide, has gained interest with the increased availability of novel data. Event data, that consist of location and time information about notable actions such as simple passes, shots, air duels, etc., have made the examination of football tactics with a statistical approach, and football analytics in general, a lot easier. Grouping the playing styles of football teams has become a popular problem in this process. Moreover, the evolution in learning from data techniques made unsupervised learning algorithms the preferred approach, rather than conventional statistical inference methods which were formerly the common solution.

As mentioned above, clustering the playing styles of football teams is the aim of this project. The significance of this project comes from its ability to help the tactical staff in football like coaches about which playing style to choose in a specific game or as a main plan for the whole football season. To achieve this goal, the evaluation of clustering results has been held by analyzing the match outcomes. This analysis, revealed the effectiveness of each playing style against others. Coaches can choose the optimal playing style for their teams by exploring their rivals’ playing styles and which playing style works best against them with the results of this project.

As a Machine Learning project about football, this project intersects with multiple disciplines. Sports analytics, AI, statistics, football, and data science are among the disciplines that are related to this project.

To briefly summarize the previous research on the area, it must be highlighted that while some of the earlier studies about grouping football teams’ playing styles utilized statistical inference methods such as Principal Component Analysis, Chi-square, or Linear Regression; others utilized unsupervised learning techniques, mostly K-means algorithm. Event data, similar to the main source of this project, is one of the three most common data types, the others are basic match statistics, and tracking data. Literature survey about the area will be discussed in detail in the next section.

As previously mentioned, a event data which consists more than 3 million data points from 1826 distinct matches has been utilized for this project. All matches played in England, Germany, Italy, Spain, and France in the 2016/2017 football season are in the scope of the data. The source of the data is wyscout[1], and it is open for commercial or academic use.

For this project, a customized multi-step K-Means clustering algorithm was used, building on previous research that primarily uses K-means[2]. This unique method has been developed to tackle the problem by firstly dividing the dataset into 4, because as Smith(?) stated[3], here are 4 distinct phases in a football game. After training each 4 subparts of the dataset with K-means where k is equal to 3, 4xm seperate labels have been acquired for each team where m is equal to the number of matches that team has played (18 or 20 depending on the league). By using a majority voting technique, in other words by simply counting, probabilities of each cluster for a team have been obtained for each phase. In this way, a new dataset with n rows (number of teams) and 12 columns (4 dataset x 3 clusters) have been created. Each cell in the data shows the probability of a team to be in that cluster for that phase in the game. Finally, training a K-means model 1 more time with the final dataset, successfully divides the teams into 3 clusters.

Even though a method previously unapplied to this problem has been introduced in this project, it is not truly novel. It is simply a multi-step version of the most popular unsupervised learning algorithm. The novelty that this project aims to achieve, comes from the feature engineering steps rather than the modeling phase. Literature survey, which will be discussed in the next section, revealed that previous research lacks the extensiveness of the feature engineering that is constructed for this project. To generate new features, four distinct methods from four papers were combined. The idea of pass motifs from Cho's paper [5], the concept of dividing the pitch into different zones from Zhou's paper [6], the approach of treating passes between football players as a graph and calculating the average connectivity score from Tuvhyko's paper [6], and the categorization of events into four different game phases—possession, out of possession, positive transition (events after stealing the ball), and negative transition (events after losing the ball) from Smith's paper [3]—were utilized. In addition to integrating these four ideas to create a more inclusive dataset, new features were developed by calculating the ratios of each event's different subtypes. In addition to integrating these four ideas to create a more inclusive dataset, new features were developed by calculating the ratios of each event's different subtypes. For example, passes were classified as side, backward, or forward based on pass angle information, and as short, middle distance, or long based on pass distance information. Ratios of each pass type were then calculated relative to the total number of passes after that. A similar process was applied to different event types like shots, runs, and duels. Consequently, 30 features for the in-possession dataset and 45 features for the out-of-possession dataset were acquired. This process of feature engineering, which broadens the dataset beyond previous projects, adds novelty to this project.

Moreover, this project proposes an original evaluation method for the problem by comparing the performance of each playing style against each other with respect to match results. It is expected to acquire reasonable evaluation results which can be commented on. The evaluation, which is considered a complication for all unsupervised learning projects due to the absence of labeled data, is aimed to be solved in this way. The comparison of the resulting clusters, as well as the comparison with chosen teams with the resulting clusters are visualized by utilizing dendrogram, scatter plot, heatmap, bar plot, stacked bar plot, and pie chart which will all be shown in the results and evaluation section of this report.

In order to provide an Interface for football coaches to use, Flask web framework have beeun utilized. All the visuals mentioned above are available for coaches in the web page, as well as the win probability for each cluster against each other. Consequently, coaches can recieve an advice on on their team’s playing style by only choosing their rival team or the team they play in. They can also analyze their rivals by courtesy of the visualizations showing the comparison between each cluster’s average score for each event and the probability of their rivals to be in each cluster as well as the closest teams to their rivals in terms of the playing style.

As the project results, the following are expected are reached: different playing styles acquired with multi-step K-means method, characteristics of each of these playing styles analyzed with respect to features of the final datasets for different phases of the game, performance analysis of each playing style against each other using the match outcomes between the clusters, visual representation of clusters, playing characteristics, performance comparisons, and finally a User Interface to obtain recommend playing style for a specific competition or against a certain opponent, while observing the related visuals that supports the recommendation.

On the next pages of this report, the following will be discussed: the previous research on clustering the playing styles of football teams problem, descriptions of the utilized techniques such as K-means, Bisecting K-Means, and Principal Component Analysis, design constraints and relative engineering standards concerning the problem, functional and non-functional requirements of the projects regarding the design constraints and relative engineering standards, evaluation methodology, system architecture showing the feature engineering and modeling processes, visual and numeric results of the project, and finally the conclusion text which also discusses what can be added to move this project forward.

# BACKGROUND

Plakias et. al. [1], reviewed 40 papers about Identifying Soccer Teams’ Styles of Play. Since that’s exactly my problem, this paper is essential in my related work research. This survey reveals the fact that most of the research in the area uses conventional statistical methods such as Factor-PCA, Chi-square, and Linear Regression. There are also several pieces of research that use AI techniques like I will in my project. While most of them use older clustering methods like K-means, there is research where Deep Neural Networks based on Multi-Layer Perceptron and feature engineering are being used. Since I also to use K-Means algortihm and PCA, it was beneficial to explore initial papers for the methods they use, and it was beneficial to explore the latter paper for feature engineering which is an essential step in my project too. Apart from making my job easier in literature survey, a concept from this paper made a huge impact to my project: game phases. The writers divide a single football match to 4, with respect to ball possesion. When a team controls to ball, when their opponents control it, when they recieve the ball recently, and when they lose it recently. Since it is visible to everyone watching the game that both teams have separate plans for this 4 phases, their playing style for each phase should be clustered separately. This idea that I inherited from this paper, resulted in the methodolgy of my project, multi-step K-means.

According to the same paper, all the research in the area focuses on at least 1 of these 3 targets: game style recognition, contextual variables, and game style effectiveness. First of all, recognition, which is the most common research target, only focuses on distinguishing different playing styles, and in some of these papers, identifying the characteristics of those playing styles. This target area will also be the main focus of my project. Secondly, contextual variables, focus on match details like the location of the game, rankings of the teams playing, and competition of the match, instead of event data like I use. These projects aim to reveal relations such as “Home teams are more likely to build up from the back.”. Contextual variables are out of scope in my project. Lastly, the least researched area, effectiveness, compare how effective are these playing styles. Different papers use various metrics for comparison. To compare the effectiveness of the playing styles, I investigated the results between the games of each cluster of teams to figure out which playing style works better again which ones. According to both my research and the papers examined in this survey, this comparison has never been done before.

Diquigiovanni and Scarpa [2], form networks for teams using pass, dribble, tackle, and shot locations. Then it hierarchically clusters teams by comparing the similarity between the networks. It was beneficial to study this project because it inspired me to divide the pitch into various areas, and representing the team as a graph. I used graph representation for only pass event while calculating the average connectivity score of the team. Even though I found their event dataset a lot more limited than mine overall, it influenced my project with two different ideas.

Similar to the paper mentioned above, Peña and Touchette[3] discusses the network representation of football teams. Unlike the other mentioned papers, it is not directly about clustering football teams’ playing styles. They analyze the football matches from 2010 World Cup by utilizing graph theory, primarily focusing on player performances. It was influential on my work with the same reason I mentioned above, average connectivity calculation of the teams’ pass graphs.

In another research on the Spanish League, Gyarmati et. al. [4], form networks with event data while focusing on only passes. Their algorithm examines the pass sequences, which they name as “pass motifs”, to search for teams’ passing styles. After that, they use hierarchical clustering to analyze the similarities and differences of teams’ passing styles. I directly used their pass motifs idea as a part of my feature generation process. It was quite effective on clustering result in terms of feature importance. For that reason, I can safely say that, this work is one of the most important references in this report in a way that it is amongst the few papers that have direct influence on my project. Like I declared in the previous section, I have managed the took their work forward by combining it with a far more inclusive list of features as well as using various types of events like shots, duels, runs etc. instead of using only pass data.

From now on, the papers that will be mentioned did not directly influenced my project. However they must be discussed in order to display the state of the art solutions on clustering football teams’ playing style problem. Lopez-Valenciano et. al. [5], explore the relationship between the variables they created which represents the attacking and defensive playing styles of teams, and the rankings of football teams in the Spanish League. They use Principal Component Analysis (PCA) for this task rather than ML methods, and they are not trying to cluster or classify teams. The similarity between the problem in this research and mine is representing playing style with feature engineering. Even though I generated many more features than them, it’s still important to see their approach too. Their dataset is also limited to only 1 league, and conventional statistics of games such as the number of shots, number of recoveries, and ball possession percentage. This data is a lot more available and a lot less significant than the one I use, event data. I also to utilized PCA, but it was a part of my data preprocessing steps, with the goal of avoiding the curse of dimensionality, since the number of columns are huge in my data. Besides from that, PCA was also needed for visualizing the clustering results by squeezing the data to only 2 features.

In an older paper from 1988, Pollard et. al. [6], try to examine playing styles with only 6 variables. They use Principal Component Analysis (PCA) like most of the projects in this area. Their data handling and statistical inference methodologies are not significant, to be honest, however, this research shows the continuing interest of researchers in grouping playing styles.

Bialkowski et. al. [7], utilize spatiotemporal tracking data, which is more detailed and less available than event data. They examine the formations of football teams which shows the distributions of football players. They discover that formation is an essential attribute to reveal the playing style. Since I do not have access to tracking data, I had to find a way to represent the distribution of football players. I used the coordinates of passes for this. Even though it is not possible to represent the locations of the football players as comprehensive as tracking data, the information of which zones are more occupied, which is generated from the event data, still benefited the clustering process.

Ruan et. al. [8], use a similar approach to Pollard et. al. [5], by using variables from simple match statistics and applying Principal Component Analysis (PCA) to them, in order to analyze defensive playing styles. They take the previous research forward by applying also regression models with the aim of measuring the effectiveness of the defensive playing styles they identified with PCA. They measure the effectiveness with a popular statistic called Expected Goals (xG). With a different approach, by comparing the match outcomes in the games that each cluster played against each other, I measured the relative effectiveness of playing styles against each other.

Now that previous research on the problem have been discussed, the methods that have been used must also be adressed, starting with the main algorithm behind the project. K-Means is a popular machine learning technique for dividing a dataset into K unique, non-overlapping clusters. The approach was first introduced in 1967 by MacQueen [9]. It works by first randomly allocating a number of centroids equal to K, which stand for the initial centers of clusters. Each data point are allocated to the closest centroid, and the centroid of each cluster is recalculated using the average of the points assigned to it. Until the centroids stabilize, this assignment and recalculation process is repeated iteratively. Convergence suggests that the clusters are heterogeneous amongst themselves and rather homogeneous inside. K-means is very useful for cluster analysis in a variety of application domains, such as computer vision and market segmentation, due to its straightforwardness and effectiveness. Nevertheless, the approach has drawbacks as well, such as its sensitivity to the original centroid positions and its inability to handle non-spherical clusters or outliers. K-means is still an essential technique in the field of unsupervised learning, regardless of these challenges.

A variation of the classic K-means algorithm known as Bisecting K-means, uses a hierachical clustering method in place of the more common agglomerative one. As Di and Gou explained [10], Bisecting K-means starts with all points in a single cluster rather than a fixed number of clusters and splits the clusters iteratively until the desired number of clusters is obtained. The algorithm chooses a cluster to divide in each iteration according to a predetermined criterion, either the cluster's size or a measure of cluster variation. The fundamental K-means method is then used to partition the selected cluster into two, and this procedure is repeated. For several types of data distributions, the Bisecting K-means method is well-known for generating more balanced cluster sizes and frequently results in superior clustering quality. Even though regular K-Means method have been used for simplicity in my project, the same datasets have also been trained with Bisecting K-Means, with the purpose of visualizing the clustering hierarchically with dendrograms.

A statistical method for dimensionality reduction that keeps as much of the data's variance as possible is principal component analysis, or PCA, first presented by Karl Pearson in 1901 [11], divides a group of potentially related variables into a smaller number of uncorrelated variables known as principal components. These components, which aid in determining the directions of maximum variance in high-dimensional data, are obtained by computing the eigenvalues and eigenvectors of the covariance matrix of the data. With each subsequent component having the maximum variance allowed by the requirement that it be orthogonal to the preceding components, the first principal component provides the maximum amount of variance. This technique is especially helpful for processing and visualizing genetic data, as well as for improving the interpretability of predictive models, which is why it is being used in this project. Since there are many columns in my data, PCA is needed to overcome the curse of dimensionality. Moreover, it is essential in visualization of the clustering results by squeezing the data to only 2 variables, making it possible to display the data in two dimensional space.

When clustering datasets without prior knowledge of the group allocations, a common difficulty is figuring out the ideal number of clusters for the K-means clustering algorithm. This is where the Elbow Method comes in. Plotting the sum of squared distances from each point to its designated cluster center (SSE) versus the number of clusters is how the Elbow Method is implemented, as described by Humaira and Rasyidah [12]. When the number of clusters grows, this plot usually exhibits a rapid drop in SSE, which is followed by a decrease gradually or "elbow" where more clusters result in lesser SSE reductions. Then, at this "elbow" point—where incorporating further clusters does not yield appreciable increases in variance explained—the ideal number of clusters is chosen, thereby striking a balance between a simple model and effective clustering performance. This approach is used because of its empirical efficacy and visual simplicity, particularly in scenarios involving exploratory data analysis where computational efficiency is crucial.

Developed in 1987 by Peter J. Rousseeuw, the silhouette score is a powerful graphical tool for evaluating the integrity of clustering results. As explained by Rousseeuw [13], the silhouette score computes an object's similarity to its own cluster in relation to other clusters, hence aiding in the interpretation and validation of cluster analysis. For researchers and analysts engaged in cluster analysis, Rousseeuw's method highlights the significance of the silhouette score in assessing the overall clustering arrangement as well as in estimating the coherence inside a cluster.

In summary, there have been multiple researches with the purpose of grouping football teams’ playing styles. While most of them only work on identifying playing styles, some researchers take the next step to discover the effectiveness of playing styles. Even though most of the papers used statistical inference methods like PCA, there is also a significant amount of research that uses unsupervised learning methods like my project. K-means is the most popular choice, followed by hierarchical clustering. While most of the projects utilize event data like my project or simple match statistics, some of them make use of tracking data which is the hardest one to get access to, and therefore, out of my project’s scope.

In order to move the current state of the art forward, my project carries out a detailed feature engineering process, involving some of the previous papers’ methods as well as original ones. Since I saw an improvable area with the used features in the previous researches, I believe the original features I generate and merging them with previous work, made a difference. Moreover, using a multi-step version of K-means algorithm instead of directly using it, like most of the previous researches have done, was a step forward to acknowledge the different phases of the game. Clustering the playing styles for each game phase successfully captured the fact that teams have different tactical schemas for different phases. Finally, examining the effectiveness of each playing style against each other with the investigation of match outcomes, also contributed to the uniqueness of my project.

# SYSTEM REQUIREMENTS

## Design Constraints and Relevant Engineering Standards

Design constraints:

1. Extensibility and Reproducibility

If incoming event data becomes accessible, the system shall be available to benefit from the new data. The system be able to accommodate the inclusion of additional league teams or match data in the future. Moreover, the same results will be shown in the upcoming sections of this report, must be able to be reached with simply running the provided code.

1. Cost Efficiency

The system must minimize the usage of the computational resources. The algorithm should be carefully implemented to work well with limited hardware.

1. Interpretability and Usability

As the project outcome, clearly visualized and interpretable results must be reached. As well as, a user interface can be used by football coaches to get tactical advice.

Relevant Engineering Standards:

1. ISO/IEC 25012:2008 Data Quality Assurance Criteria:

Data quality in this project is critically dependent on compliance with ISO/IEC 25012:2008 [8]. Accuracy and broadness are crucial, so an in-depth analysis of the Wyscout event dataset, is required to look for any inconsistencies, errors, or missing data. Correlating position and timestamp information with recorded events such as passes, shots, and duels, as well as verifying the correctness of the data are all part of this process.

1. ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML):

For the AI and machine learning components of the project, compliance with ISO/IEC 23053:2022 [9] is essential. The large-scale datasets must be handled by an efficient and effective unsupervised learning algorithm in order to cluster team playing styles. This involves developing algorithms that can handle enormous volumes of data and add new data sources as they become available. This flexibility is essential to the project's sustainability over time because it allows the inclusion of teams from leagues outside of its original scope and takes evolving team tactics into account. In order to provide coaching staff and tactical analysts with precise, useful insights, the AI model must be powerful as well as adaptable to these criteria.

## Functional Requirements

1. The system shall create new features from existing features.
2. The system shall group the data by teams.
3. The system shall be extensible with the introduction of new data, by seamlessly build on the model with recently added data.
4. The system shall be reproducible by simply running the provided code in a different computer.
5. The system shall explore the feature importances and average values for the features with respect to found clusters, in order to discover the characteristics of playing styles.
6. The system shall provide visualized results.
7. The system shall provide an interface to choose the competition or team.
8. The system shall provide an interface to return the optimal playing style for the given team.

## Non-Functional Requirements

1. Clustering results must be visualized in a clearly understandable way.
2. When the user interacts with the interface by providing a team or competition name, the recommended playing style a without a significant waiting time, 2 seconds at most.
3. At least 3 types of plots shall be created to visualize different aspects of clustering results.
4. When the user interacts with the interface by choosing the requested plot type and filtering teams or competitions, visualizations shall be shown without a significant waiting time, 5 seconds at most.
5. Silhoutte score for all 5 clustering results, shall be greater than 0.25.
6. The interface shall be as simple as possible for accessibility purposes.

## Evaluation Methodology

Evaluation is considered an issue for unsupervised learning, due to the absence of labeled data. Unlike supervised learning, the success of the algorithm can not be measured by the difference between the predicted labels and the actual labels, because the actual labels does not exist. Nevertheless, evaluating the success of an unsupervised learning task is still possible by examining the clustering results, as well as using techniques that compares Intra-cluster and Inter-cluster distances, such as silhouette score. Furthermore, the project must ensure all functional and non-functional requirements listed above. These requirements contains criteria about the clustering success, extensiveness of data preparation and visualization steps, interpretability of the clustering results, along with usability and performance of the user interface.

First of all, to validate the clustering success, clustering results must be checked to make sure that multi-step K-Means model has successfully grouped the data. If there is only one cluster even though predetermined number of clusters is greater than one, it means the data is not suitable for K-Means algorithm, and possibly, it can not be clustered. Moreover, when the overall outcomes of the matches between different clusters are examined, if the win, draw, and lose percentages do not significantly differ from 1/3, it might reveal the failure in clustering. However, if they considerably differ from 1/3, which is the outcome of randomly chosen matches, it means that multi-step K-Means method successfully grouped the data, and its results might reveal notable information about different playing styles of football teams. Furthermore, it must be checked if the silhouette scores of all five clustering tasks are greater than the chosen threshold, 0.25, indicating a satisfying clustering.

Secondly, extensiveness of data preparation and visualization steps should be evaluated. The new datasets grouped by teams, must include information about all the major event types available on the original data, such as pass, shot, acceleration, duel. Moreover, there must be many new generated features which are non-trivial, in other words, they must contain information which are not directly understandable from the original variables. Preferred pass motifs, more used areas in the football pitch, average connectivity of the pass graph are some examples of newly generated non-trivial features. The created visuals should also be comprehensive enough. As the non-functional requirements dictated, there must be at least 3 separate types of plots to visualize the clustering results and analyze the distinct playing styles.

Third of all, the visuals created should provide interpretable information about the playing styles of football teams. Comparing the average value of a feature for each cluster should reveal patterns about the characteristics of found playing styles. If a pattern such as “The teams that prefer ABCD pass motif more are also tend to prefer high passes rather than low passes.”, it means that the aim of the project is reached. In the results and evaluation section of this report, the found patterns will be discussed and visualized. Furthermore, the success of the project will be evaluated based on the quality and quantity of the found patterns.

Finally, the usability and performance of the user interface will be evaluated.

# SYSTEM ARCHITECTURE

You should describe your system architecture in this section.

# RESULTS AND EVALUATION

You should present and evaluate your experimental results in this section.

# CONCLUSIONS AND FUTURE WORKS

You should summarize your project and present final remarks. It is also crucial that you include future work and possibilities to extend your design in the future.

**TAKIMLARIN FARKLI SKORLARA VE DAKİKALARA GÖRE OYUN PLANI DEĞİŞİYOR. MAÇ BOYUNCA AYNI STİLLE OYNAMAK ZORUNDA DEĞİLLER. FUTURE WORK BUNUN ÜZERİNE YOĞUNLAŞABİLİR.**

# REFERENCES

[1] S. Plakias et. al, “Identifying Soccer Teams’ Styles of Play: A Scoping and Critical Review”, JFMK, vol. 8, no. 2, p. 39, Mar. 2023. [Online]. Available: doi: 10.3390/jfmk8020039

[2] J. Diquigiovanni and B. Scarpa, “Analysis of association football playing styles: An innovative method to cluster networks”, Statistical Modelling, vol. 19, no 1, pp. 28-54, Feb. 2019. [Online]. Available: doi: 10.1177/1471082X18808628.

[3] J. L. Peña and H. Touchette, “A network theory analysis of football strategies”, 2012, doi: 10.48550/ARXIV.1206.6904.

[4] L. Gyarmati, H. Kwak, and P. Rodriguez, “Searching for a Unique Style in Soccer”, 2014. [Online]. Available: doi: 10.48550/ARXIV.1409.0308.

[5] A. Lopez-Valenciano et. al., “Association between offensive and defensive playing style variables and ranking position in a national football league”, Journal of Sports Sciences, vol. 40, no 1, pp. 50-58, Jan. 2022. [Online]. Available: doi: 10.1080/02640414.2021.1976488.

[6] R. P. Hartley Charles Reep, Simon, “The Quantitative Comparison of Playing Styles in Soccer”, in Science and Football (Routledge Revivals), Routledge, 1988.

[7] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, “Identifying Team Style in Soccer Using Formations Learned from Spatiotemporal Tracking Data”, in 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, China: IEEE, Dec. 2014, pp. 9-14. [Online]. Available: doi: 10.1109/ICDMW.2014.167.

[8] L. Ruan, H. Ge, Y. Shen, Z. Pu, S. Zong, and Y. Cui, “Quantifying the Effectiveness of Defensive Playing Styles in the Chinese Football Super League”, Front. Psychol., vol. 13, no. 899199, Jun. 2022. [Online]. Available: doi: 10.3389/fpsyg.2022.899199.

[9] J. MacQueen, “Some methods for classification and analysis of multivariate observations”, 1967. [Online]. Available: https://www.semanticscholar.org/paper/Some-methods-for-classification-and-analysis-of-MacQueen/ac8ab51a86f1a9ae74dd0e4576d1a019f5e654ed

[10] J. Di ve X. Gou, “Bisecting K-means Algorithm Based on K-valued Selfdetermining and Clustering Center Optimization”, JCP, pp. 588-595, 2018, doi: 10.17706/jcp.13.6.588-595.

[11] K. Pearson, “LIII. On lines and planes of closest fit to systems of points in space”, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no 11, pp. 559-572, Oct. 1901, doi: 10.1080/14786440109462720.

[12] H. Humaira and R. Rasyidah, “Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm”, in Proceedings of the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia: EAI, 2020. doi: 10.4108/eai.24-1-2018.2292388.

[13] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”, Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, Oct. 1987, doi: 10.1016/0377-0427(87)90125-7.

[14] ISO/IEC 25012:2008, "Data Quality Assurance Criteria," International Organization for Standardization/International Electrotechnical Commission, Standard, 2008.

[15] ISO/IEC 23053:2022, "Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)," International Organization for Standardization/International Electrotechnical Commission, Standard, 2022.