TRACKING FORMATION CHANGES AND ITS EFFECTS ON SOCCER USING POSITION DATA

Jinho Kim
Marquette University

Follow this and additional works at: https://epublications.marquette.edu/dissertations_mu

Recommended Citation
Kim, Jinho, "TRACKING FORMATION CHANGES AND ITS EFFECTS ON SOCCER USING POSITION DATA" (2023). Dissertations (1934 -). 3181.
https://epublications.marquette.edu/dissertations_mu/3181
TRACKING FORMATION CHANGES AND
ITS EFFECTS ON SOCCER USING
POSITION DATA

By

Jinho Kim, B.S.

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
In Partial Fulfillment of the Requirements for
The Degree of Master of Science

Milwaukee, Wisconsin

December 2023
ABSTRACT

Jinho Kim, B.S.

This study investigated the application of advanced machine learning methods, specifically k-means clustering, k-Nearest Neighbors (kNN), and Support Vector Machines (SVM), to analyze player tracking data in soccer. The primary hypothesis posits that such data can yield a standalone, in-depth understanding of soccer matches. The study revealed that while k-means and spatial analysis are promising in analyzing player positions, kNN and SVM show limitations without additional variables.

Spatial analysis examined each team’s convex hull and studied the correlation between team length, width, and surface area. Results showed team length and surface area have a strong positive correlation with a value of 0.8954. This suggested that teams with longer team length have a more direct style of play with players more spread out which led to larger surface areas. k-means clustering was performed with different k values derived from different approaches. The silhouette method recommended a k value of 2 and the elbow recommended a k value of 4. The context of the sport suggested additional analysis with a k value of 11. The results from k-means suggested natural data partitions, highlighting distinct player roles and field positions. kNN was performed to find similar players with the model of k = 19 showing the highest accuracy of 8.61%. The SVM model returned a classification of 55 classes which indicated a highly granular level of categorization for player roles. The results from kNN and SVM indicated the necessity of further contextual data for more effective analysis and emphasized the need for balanced datasets and careful model evaluation to avoid biases and ensure practical application in real-world scenarios.

In conclusion, each algorithm offers unique perspectives and interpretations on player positioning and team formations. These algorithms, when combined with expert knowledge and additional contextual data, can significantly enrich the scope of analysis in soccer. Future work should consider incorporating event data and additional variables to enhance the depth of analytical insights, enabling a more comprehensive understanding of how formations evolve in response to various in-game situations.
ACKNOWLEDGMENTS

Jinho Kim, B.S

I would like to thank Dr. Paula Papanek for allowing me to further my education. I would like to thank Dr. Matthew Hawkey for his advice and for pushing me to think outside the box. I want to thank my parents for always believing in me.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... i

LIST OF TABLES ....................................................................................................................... iv

LIST OF FIGURES ...................................................................................................................... v

I. INTRODUCTION ................................................................................................................... 1

II. LITERATURE REVIEW ......................................................................................................... 6
   A. Overview and History of Machine Learning in Sports Analytics .................. 6
   B. The Use of k-means and Its Application in Sports ........................................ 8
   C. The Use of k-Nearest Neighbors and Its Application in Sports ................... 10
   D. The Use of Support Vector Machines and Its Application in Sports .......... 11

III. METHODOLOGY .................................................................................................................. 14
   A. R Programming ............................................................................................................ 14
   B. Data Processing .......................................................................................................... 14
   C. k-means ....................................................................................................................... 15
   D. k Nearest Neighbors (kNN) ...................................................................................... 17
   E. Support Vector Machines (SVM) ............................................................................. 19

IV. RESULTS ............................................................................................................................. 20
A. Spatial Analysis ................................................................. 20
B. k-means .............................................................................. 21
C. k Nearest Neighbors (kNN) .................................................. 26
D. Support Vector Machines (SVM) .......................................... 27

V. DISCUSSION ........................................................................ 28

VI. CONCLUSION AND FUTURE WORKS ................................. 35

BIBLIOGRAPHY ...................................................................... 37
LIST OF TABLES

Table 1. Team length, width, and surface area of each team .......................... 20
Table 2. Correlation matrix between spatial components .............................. 20
Table 3. Average total distance run per cluster of players ............................ 25
Table 4. Accuracy, Precision, Recall, F1 scores for each k value (kNN) ............. 27
LIST OF FIGURES

Figure 1. Visualization of correlation matrix between spatial components .................. 22

Figure 2. Silhouette method for the optimal value of k clusters .............................. 23

Figure 3. Elbow method for optimal value of k clusters ...................................... 23

Figure 4. k-means clustering with 2 clusters (k = 2) .......................................... 24

Figure 5. k-means clustering with 4 clusters (k = 4) ........................................... 25

Figure 6. k-means clustering with 11 clusters (k = 11) ....................................... 26

Figure 7. kNN accuracy vs k values .................................................................... 27
I. Introduction

Sports analytics is the investigation and modeling of professional sports performance and sports management using scientific techniques (Morgulev et al., 2018). This discipline frequently employs principles and techniques from statistics, data mining, game theory, biomechanics, and kinesiology. This data-driven discipline encompasses a variety of sports, from football and basketball to tennis and cricket, each with its unique set of variables and metrics. Analytics for sports management is also versatile and addresses areas from economic assessment of the impact of mega-sport events to the allocation of scarce resources while building professional teams’ rosters, to optimal ticket pricing via evaluation of the fans’ level of interest (Morgulev et al., 2018).

Data are more readily available, presenting unique opportunities to explore this resource with the use of suitable techniques to gain insight into the evaluation of player and team performances is now possible. In soccer, teams in the English Premier League (EPL) have become relatively advanced in terms of performance analytics, and some of them have even made performance data available to fans for open-source analysis (Davenport, 2016). Sports analysts scrutinize players' physical and tactical performance data, aiming to optimize both individual capabilities and team dynamics. They assess diverse elements such as a player's speed, agility, endurance, or shooting accuracy in real-time scenarios, enabling coaches to make tactical decisions mid-game.

Soccer, also known as football in many parts of the world, stands apart from many other popular sports due to several key characteristics. One of soccer's distinguishing features is its continuous play nature. Unlike American football or basketball which use
countdown clocks with regular stoppages, the soccer clock counts up to 45 minutes in
each half and does not stop for breaks in play. This contrasts with the regular
interruptions in sports like American football, basketball, or baseball, combined with the
fact that soccer has more players on the field (22 players), a larger pitch (around 105
meters long and 68 meters wide) makes soccer arguably the most challenging to analyze
of all the major team sports (Bornn et al., 2018). Perhaps the most defining characteristic
of soccer is its emphasis on team coordination and defined player roles. Each player has
specific responsibilities depending on their position, be it defender, midfielder, forward,
or goalkeeper. Unlike in basketball, where players frequently rotate positions, or in
baseball, where roles are highly specialized, soccer players usually stick to their assigned
roles throughout the game (Hughes & Frank, 2005). Notably, Bialkowski et al. (2014)
used player and ball tracking data accumulated over an entire season to identify the role
of a player and to find team-based features to predict the identity of soccer teams.

Machine learning is the study of algorithms that are used to teach machines to
extract, learn, and make decisions from relevant data (Mahesh, 2020). The algorithms
adjust and learn from the data to improve their performance over time. Key techniques in
machine learning include regression, classification, clustering, and dimensionality
reduction, and these can be implemented using various methods. k-means is a clustering
method that divides a dataset into k distinct, non-overlapping clusters. The objective is to
minimize the sum of the distances between the points and the corresponding cluster
centroid (MacQueen, 1967). k Nearest Neighbors (kNN) classify a data point based on
how its neighbors are classified. The k value is the number of neighbors considered, and
the classification of a point is made by a majority vote of its neighbors (Cover & Hart,
Support Vector Machines (SVM) are supervised learning models, used for classification or regression. They work by finding a hyperplane that best divides a dataset into classes by maximizing the margin between the classes (Cortes & Vapnik, 1995). Together, these methods represent key approaches within machine learning and embody a range of techniques from unsupervised learning (k-means) to supervised learning (kNN and SVM).

Many machine learning techniques have become vital tools in sports analytics and there have been many studies that have employed various machine learning methods to analyze player movement and formations in soccer. Ribeiro et al. (2021) used deep learning, a subset of machine learning, to analyze player movements in soccer. They introduced a deep learning model capable of analyzing sequences of player positions, predicting future actions, and recognizing strategic patterns. Kerr (2015) was able to apply various classification methods such as support vector machines to predict which teams produced the sequence of ball events to identify possible styles of winning teams. Research by Gudmundsson and Wolle (2012) applied cluster analysis to detect correlations between the most frequent movements of the players. The goal was to identify how the team moves in reaction to either gaining or losing possession of the ball.

In many studies, game event data is an essential component for analyzing soccer games, particularly when applying machine learning techniques. Event data refers to the detailed data for each ball event and is recorded anytime a player makes a play on the ball, apart from dribbling, and includes additional information for each ball event such as the location, the player involved, and the outcome (Brooks et al., 2016). These datasets provide an in-depth perspective on the performance of teams and players. For instance,
data on successful passes, shooting accuracy, or tackles made can reveal a player's effectiveness in their role. Memmert et al. (2013) showed players and coaches preparing goalkeepers by studying the probable directions of penalty shots, based on the shooters’ previous statistics before critical matches. This information is invaluable for coaches when trying to devise tactics to prepare for matches against different teams. In the context of machine learning, event data serves as the training input that enables algorithms to learn, predict, and evaluate complex patterns and strategies. Such analyses allow the models to highlight weaknesses in opposition teams, identify promising strategies, and enhance decision-making processes in training and during matches.

However, while the recent developments in player tracking systems in professional sports have enabled teams to track player movement automatically, collecting game event data is still a manual process performed by operators present at each match, one per team and another operator acting as responsible supervisor of the output of the whole match (Pappalardo, Cintia, Rossi, et al., 2019). This can lead to human error, and data missingness, and can even lower the objectivity of match analysis. The absence of event data, such as passes, shots, or tackles, challenges us to derive meaningful insights solely based on their locations on the field throughout the match. By tracking player movements, I aim to identify patterns, highlight areas of the field they frequent most, and discern their role within the team’s strategy.

The main hypothesis driving this thesis is that player tracking data when analyzed with advanced machine learning methods, can provide a standalone, in-depth understanding of soccer matches. If successful, this approach could streamline data collection efforts, focusing solely on player movement, while still generating valuable
insights for teams, coaches, and analysts. The thesis explored these questions using various machine learning techniques such as k-means, k-nearest neighbors, and support vector machines. The findings could have significant implications for the future of sports analytics, potentially reshaping how data is collected and analyzed in soccer and other team sports.
II. Literature Review

A. Overview and History of Machine Learning in Sports Analytics

Sports analytics is the investigation and modeling of professional sports performance using scientific techniques. This discipline frequently employs principles and techniques from statistics, data mining, game theory, biomechanics, kinesiology, etc. Recent advancements in machine learning have become an increasingly integral part of sports analytics, revolutionizing the way data is analyzed, enhancing decision-making processes, and transforming data analysis within the sports industry (Alamar, 2013).

Machine learning techniques have been applied in sports analytics for several decades, continually evolving to meet the growing demands of the industry (Minelli et al., 2013). One of the earliest works discussing the potential application of machine learning in sports performance to improve decision-making was done by Lapham and Bartlett (1995). The paper discusses how expert systems, a knowledge-based database used for reasoning, can be used to investigate sports biomechanics. Over time, advancements in computing power and data availability enabled the development of more sophisticated algorithms and predictive models.

Machine learning algorithms have enabled comprehensive player performance analysis by considering various metrics such as player tracking data, biometric data, and historical performance records (Minelli et al., 2013). These algorithms identify key performance indicators, aiding in player selection, tactics, and training programs. They analyze player movements, positioning, and interactions to provide insights into running patterns, positioning effectiveness, passing accuracy, and more. In baseball, machine learning algorithms have found extensive application in pitch analysis, where they can
classify and predict pitch types based on movement and velocity data (Albert & Bennett, 2006). Additionally, machine learning has been used to evaluate batter-pitcher matchups and optimize strategies for pitchers and batters. Machine learning has made significant contributions to basketball analytics, particularly in player tracking analysis (Geng & Wang, 2018). Algorithms can process player tracking data to understand movement patterns, identify offensive and defensive trends, and assess individual player contributions. These insights inform lineup optimizations, defensive strategies, and player development (Sarlis & Tjortjis, 2020). In football, machine learning has been employed for various applications, including player performance evaluation and play prediction (Sarma & Su, 2016). Algorithms can analyze player tracking data, game situations, and historical records to assess player performance, identify tactical patterns, and optimize game strategies. In soccer, machine learning algorithms have been used to analyze player tracking data, match statistics, and historical records (Lin & Zheng, 2019). They provide insights into player performance, playing styles, and team dynamics, aiding in player evaluation, tactical adjustments, and game strategy optimization.

Machine learning algorithms optimize game strategies by analyzing opponent data, game situations, and historical records (Bialkowski & Lucey, 2014). They identify patterns in player movement, predict opponent tactics, and suggest optimal lineups or tactical adjustments. By leveraging real-time data streams during games, machine learning models provide recommendations for tactical adjustments, allowing teams to adapt quickly. In baseball, machine learning models can analyze historical data to optimize defensive shifts, positioning, and pitch sequencing based on opponent tendencies (Albert & Bennett, 2006). This helps teams gain a competitive advantage by
making data-driven decisions during games. Machine learning algorithms have been used in basketball to analyze opponent tendencies, identify defensive weak points, and optimize offensive strategies (Scudellari & Napolitano, 2018). By leveraging advanced analytics, teams can exploit opponent weaknesses and make informed decisions regarding shot selection, offensive spacing, and defensive assignments. In football, machine learning algorithms assist in-game strategy optimization by predicting opponent play-calling tendencies, identifying optimal play calls in specific game situations, and optimizing player matchups (Gudmundsson & Horton, 2016). These insights enable teams to develop effective game plans, improve in-game decision-making, and maximize their chances of success. In soccer, machine learning algorithms can analyze opponent data to identify tactical patterns, predict playing styles, and optimize team formations and strategies (Dufour et al., 2021). By leveraging historical data, teams can make informed decisions regarding pressing intensity, attacking strategies, and defensive organization.

B. The Use of k-means and Its Application in Sports

k-means clustering is a popular unsupervised machine learning algorithm used for data clustering and pattern recognition. It is an iterative algorithm that partitions data into k clusters based on similarity measures (Jain, 2010). The algorithm aims to minimize the within-cluster sum of squares, assigning each data point to the nearest centroid. k-means is known for its simplicity, scalability, and efficiency, making it suitable for large-scale sports datasets.

k-means clustering has been applied to analyze player performance in various sports, including soccer. By clustering players based on their statistical attributes, k-means help identify groups of players with similar playing styles, strengths, or positions
It has notable uses in player classification for various sports from baseball (Gerlica et al., 2020) to badminton (Sinadia & Murwantara, 2022). This information aids in player evaluation, scouting, and even team composition.

In soccer, k-means clustering has been used to analyze player positions and movements on the field. By clustering player trajectories, the algorithm can identify distinct playing patterns and tactical formations (Liu et al., 2021). This information is valuable for understanding team strategies, assessing opponent tactics, and optimizing game plans. k-means clustering can be used to optimize game strategies in different sports. By clustering opponents' historical data, the algorithm helps identify recurring patterns, tendencies, and weaknesses (Sampaio et al., 2018). This information enables teams to tailor their strategies, make data-driven decisions, and identify opponent weak positions.

Lastly, k-means clustering has been used to assess injury risks in sports. By clustering athletes based on factors such as workload, training patterns, and injury history, the algorithm helps identify high-risk groups (O’Donoghue et al., 2019). This information supports injury prevention strategies, workload management, and player health monitoring.

The application of k-means clustering in sports analytics has proven to be valuable in various domains. From player performance analysis to tactical assessment and fan engagement, the algorithm provides insights that assist decision-making processes in sports organizations. By leveraging k-means clustering, teams can optimize game strategies, improve player evaluations, and enhance the fan experience. Additionally, the algorithm helps identify injury risk factors and aids in injury prevention strategies. As the
field of sports analytics continues to advance, k-means clustering will likely remain a valuable tool for data-driven decision-making in the sports industry.

C. The Use of k-Nearest Neighbors and Its Application in Sports

k-Nearest Neighbors (kNN) is a versatile machine learning algorithm used for classification and regression tasks. k-Nearest Neighbors is a non-parametric algorithm that uses a similarity measure to classify data points based on their proximity to neighboring points. In the context of sports analytics, kNN leverages historical data and player attributes to make predictions or classify players (Cover & Hart, 1967). The algorithm selects k nearest neighbors and assigns the class label based on the majority vote or average of the neighbors.

kNN has been applied to analyze player performance in various sports, including soccer. By utilizing kNN, analysts can classify players based on their statistical attributes, playing positions, or roles on the field (Cover & Hart, 1967). This information aids in player evaluation, talent identification, and team composition. kNN has used features of players such as physical, mental, and technical qualities to find talented players that fit a team’s style (Bazmara & Jafari, 2013). By utilizing kNN, analysts can identify similar game situations or playing patterns based on historical data and player positions (Rao et al., 2016). This analysis provides insights into team strategies, and opponent behavior, and assists in optimizing game plans. kNN has been used to assess player similarity and identify players with comparable characteristics. By comparing player attributes and performance metrics, kNN can identify similar players, providing valuable information for scouting, transfer decisions, and player development (Fernández-Caballero et al.,
This technique helps teams find players who possess desired traits and fit into specific team strategies.

The application of k-Nearest Neighbors in sports analytics has proven to be valuable across various domains. From player performance analysis to tactical assessment, player similarity assessment, and injury prediction, kNN provides insights that assist decision-making processes in sports organizations. By leveraging kNN, teams can optimize game strategies, improve player evaluations, enhance injury prevention strategies, and gain a competitive edge.

D. The Use of Support Vector Machines and Its Application in Sports

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm known for its ability to classify and predict data thus they are popularly used for classification and regression tasks. SVM aims to find an optimal hyperplane that separates data points into different classes, maximizing the margin between the classes (Cortes & Vapnik, 1995). SVM's ability to handle high-dimensional data and nonlinear relationships makes it suitable for analyzing complex sports datasets.

Support Vector Machines have been applied to analyze player performance in various sports, including soccer. By using SVM, analysts can classify players based on their statistical attributes, playing positions, or roles on the field (Joo & Chung, 2014). This information aids in player evaluation, scouting, and team composition. Support Vector Machines have been employed to predict match outcomes in sports, including soccer. By training SVM models on historical data, analysts can predict the likelihood of different match outcomes (Asimakopoulos & Asimakopoulos, 2017) This information helps coaches, analysts, and sports bettors make informed decisions regarding strategies,
game plans, and wagers. In addition, Support Vector Machines have been used to assess injury risks in sports. By training SVM models on player data, including workload, physical attributes, and injury history, analysts can predict the likelihood of future injuries (O’Donoghue et al., 2019). While there is dogma that suggests that higher training loads cause higher injury rates, there is also evidence that training has a protective effect against injury (Gabbett, 2016). To accurately assess this data, SVMs can be used to identify critical variables and thus assist in injury prevention strategies, workload management, and player health monitoring by appropriately prescribing high training loads to improve player fitness.

Support Vector Machines have been prominently used to analyze soccer tactics and player movement patterns. By utilizing SVM, analysts can classify and predict tactical formations based on player positioning and movement data (Brooks et al., 2016). This analysis provides insights into team strategies, and opponent behavior, and assists in optimizing game plans. Support Vector Machines have also been applied to analyze player positioning and movement patterns. SVM models can predict player movements and trajectories based on historical data, aiding in tactical analysis and strategic decision-making (Ramos & Ribeiro, 2020). This analysis helps coaches and analysts understand player behavior, identify playing patterns, and optimize team formations.

The application of Support Vector Machines in sports analytics has demonstrated significant value across multiple domains. From player performance analysis to tactical assessment, injury risk assessment, and match outcome prediction, SVM models provide insights that assist decision-making processes in sports organizations. By leveraging Support Vector Machines, teams can optimize game strategies, improve player
evaluations, enhance injury prevention strategies, and gain a competitive edge. As sports analytics continues to evolve, Support Vector Machines will likely remain a valuable tool for extracting meaningful information from complex sports datasets.

Continued advancements in machine learning techniques hold promise for further enhancing player performance analysis and game strategy optimization in sports analytics. Research focuses on integrating real-time data streams, computer vision for enhanced player tracking, and applying deep learning-based component models to capture the complex intricacies of spatiotemporal tactics (Fernández et al., 2019). These advancements have the potential to revolutionize the field, providing even more precise insights and recommendations.
III. Methodology

A. R Programming

The R programming language, complemented by its integrated development environment RStudio, provides a powerful and versatile platform for implementing a wide array of machine learning algorithms. I decided to use R to perform the various machine learning tasks required for this study. The syntactical simplicity of R, alongside its extensive collection of packages, allows for quick implementation and experimentation with different models.

B. Data Processing

The data was obtained from 14 games of 11 different teams in the Major League Soccer (MLS) from 2014. The data was publicly available through a US Soccer-sponsored hackathon. The data contains half, time, and the x, and y coordinates of every player who participated in each match. Half refers to the first or second half of the game and time was collected in seconds.

To run the algorithms mentioned data were reformatted. The original data contained half, time, and each player’s x, and y position in columns. The team name and player name as well as the x or y position are contained in a single column. For better ease of analysis, the data was converted from a wide format to a long format with the player name, team, and x, y position separated into different columns. From the x, and y positions I created lead/lag metrics to see the difference from one data point to the next which was then used to calculate the total distance traveled, average, and maximum velocity for each player. Velocity is in meters per second and capped at 10 m/s. The
fastest soccer players can sprint at speeds of up to 10 m/s, but such speeds are not sustainable for long durations (Delaney et al., 2018). Thus, capping the velocity at this 10 m/s reflected realistic human capabilities and also allowed for data smoothing and error reduction (Delaney et al., 2018).

Exploratory data analysis focused on creating additional spatial features for the dataset. Moura et al. calculated the convex hull i.e., the area covered by the contour of the player’s position to quantify players’ ability to spread depending on the phases of the game which was linked to the team’s playing style (Moura et al., 2013). Similarly, I calculated convex hulls for each separate team from the x, and y coordinates. The ‘chull’ function was used to calculate the convex hull indices from which team length, width, and surface area were derived. Team length reflects the vertical stretch of the team across the field while the width indicates how much the team stretched to the side of the field. The surface area enclosed by the convex hull represented the total area occupied by the team. From the spatial metrics, a correlation matrix was created using the ‘corrplot’ package, to find possible correlations between various spatial variables.

C. k-means

Match data was combined into a single data frame for analysis. Experimentation was done with various k values to find different ways to cluster the players. By performing k-means on the data, I divided the players into k different groups based on their positions on the field which is represented by their x, y coordinates. To determine the optimal number for k I performed two different methods: the silhouette method and the elbow method.
The silhouette method is a state-of-the-art method used to identify the correct cluster number in the dataset (Masud et al., 2018). It provides a succinct graphical representation of how well each object lies within its cluster (Rousseeuw, 1987). The silhouette value is a measure of how similar an object is to its cluster compared to other clusters. The silhouette value ranges from -1 to 1, where a high value indicates that the object is well-matched to its cluster and poorly matched to neighboring clusters. A silhouette close to 1 implies that the object is in an appropriate cluster, while a silhouette close to -1 implies that the object is in the wrong cluster. A silhouette score near 0 suggests overlapping clusters. When used in conjunction with k-means the method involves calculating the average silhouette of observations for different numbers of clusters (k). The optimal number of clusters is then taken to be the one that maximizes the average silhouette over a range of possible values for k. The ‘cluster’ and ‘factoextra’ packages are used in R to compute and visualize the silhouette widths for different numbers of clusters.

The elbow method consists of plotting the explained variance as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. This "elbow" point is theoretically where the marginal gain in explained variance drops, signifying that additional clusters do not explain much more variance in the data (Ketchen & Shook, 1996). Essentially, the idea is to choose a small value of k that still has a low sum of squared distances (SSD). The method's name derives from the shape of the plot; the optimal number of clusters is assumed to be at the "elbow" of the curve, representing the point after which the inertia or within-cluster sum of squares (WSS) starts decreasing at a slower rate. The ‘cluster’ and ‘factoextra’ packages are used again in R to compute
and visualize the optimal number of k clusters for the elbow method. I looked for the point where the rate of change starts to decrease. This point represented the "elbow", suggesting that adding more clusters beyond this point might not capture much more structure in the data, making it the optimal number of clusters.

In addition to the two possible k values found through the silhouette and elbow methods, a different model was also run with k = 11. This was to take into account the fact that soccer is played with 11 players. I hoped the model would be able to show 11 distinct clusters that can represent each position on the field. After finding potential k values for the algorithm, k-means was run using the ‘dplyr’ package. The ‘ggplot2’ package was used to visualize the results. The results created a scatterplot with the soccer players grouped by cluster, colored differently for each cluster.

D. k Nearest Neighbors (kNN)

To run kNN in RStudio, I first have to install the ‘class’ library package in R which contains the ‘knn’ function. Since kNN is a supervised algorithm, the data is split into training and testing sets. In an ideal situation, the dataset would include additional variables or labels such as player position. However, since the data did not contain player position, the model was run based on each player. By running kNN on individual players the focus becomes identifying players with similar attributes or performances, regardless of their positions. The elbow method was used again to determine the optimal k value. Through experimentation and cross-validation, an optimal k value can be found. Post-classification, the effectiveness of the model was evaluated using a confusion matrix, which provided a detailed breakdown of the true positives, false positives, false negatives, and true negatives. Metrics such as accuracy, precision, recall, and F1 score
were derived from the confusion matrix. Each metric provides different perspectives on the performance of the classification model. First, some basic terms are defined:

- True Positives ($TP$): The number of correct positive predictions.

- True Negatives ($TN$): The number of correct negative predictions.

- False Positives ($FP$): The number of incorrect positive predictions.

- False Negatives ($FN$): The number of incorrect negative predictions.

Metrics derived:

1. Accuracy measures the overall correctness of the model. It is calculated as the proportion of total correct predictions (both true positives and true negatives) to all predictions made as shown by the equation below. In the soccer context, high accuracy would indicate that the model is generally good at matching players to their correct playing positions based on their similarities in the dataset (Bazmara & Jafari, 2013).

\[
Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]

2. Precision measures the correctness of positive predictions. It is the number of true positives divided by the total number of positive predictions shown in the equation below. For soccer player classification, high precision in a specific position means that when the model predicts a player is a forward, it is likely correct.

\[
Precision = \frac{(TP)}{(TP + FP)}
\]

3. Recall measures the model's ability to correctly identify all relevant instances (true positives) out of all actual positives. For example, high recall for a goalkeeper means that
the model is good at identifying all players who are goalkeepers. Recall is calculated using this equation.

\[
Recall = \frac{(TP)}{(TP + FN)}
\]

4. The F1 score is the harmonic mean of precision and recall. It provides a single metric that balances both the concerns of precision and recall into one number. The F1 score is calculated using this equation.

\[
F1\ Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}
\]

The algorithm was run with incrementally increasing k values until the accuracy of the model started to level off indicating an elbow. A table was compiled and the ‘ggplot’ package was used to visualize the plot of the accuracy of each model for each k value.

E. Support Vector Machines (SVM)

The SVM analysis was implemented in RStudio, employing the ‘e1071’ package, which provides a convenient interface to the LIBSVM library for SVM classification and regression. SVMs are a versatile set of algorithms that can be used for both classification and regression tasks. However, with the absence of labels such as player position, I determined the best use of SVM would be to find player similarities based on the categorical variables available. SVM was run based on each individual player during each half of the match.
IV. Results

A. Spatial Analysis

The results of the spatial analysis are shown in Table 1. Team length reflects the vertical stretch of the team across the field while the width is indicative of how much the team is using the sides of the field. The surface area enclosed by the convex hull represents the total area occupied by the team. These metrics are used for a post-match analysis to understand team performance and can allow teams to make tactical adjustments during live games.

### Table 1. Team length, width, and surface area of each team

<table>
<thead>
<tr>
<th>Team</th>
<th>Team Length (m)</th>
<th>Team Width (m)</th>
<th>Team Surface Area (m²)</th>
<th>Average Distance run (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team 1</td>
<td>117.04</td>
<td>75.34</td>
<td>8223.59</td>
<td>8709</td>
</tr>
<tr>
<td>Team 2</td>
<td>115.40</td>
<td>72.82</td>
<td>7847.00</td>
<td>9507</td>
</tr>
<tr>
<td>Team 3</td>
<td>121.40</td>
<td>74.18</td>
<td>8539.83</td>
<td>9383</td>
</tr>
<tr>
<td>Team 4</td>
<td>122.73</td>
<td>74.79</td>
<td>8852.99</td>
<td>10380</td>
</tr>
<tr>
<td>Team 5</td>
<td>111.00</td>
<td>76.35</td>
<td>8029.40</td>
<td>9069</td>
</tr>
<tr>
<td>Team 6</td>
<td>120.64</td>
<td>74.17</td>
<td>8173.20</td>
<td>8905</td>
</tr>
<tr>
<td>Team 7</td>
<td>120.18</td>
<td>75.07</td>
<td>8487.69</td>
<td>9064</td>
</tr>
<tr>
<td>Team 8</td>
<td>115.07</td>
<td>77.90</td>
<td>8314.21</td>
<td>8845</td>
</tr>
<tr>
<td>Team 9</td>
<td>114.82</td>
<td>75.57</td>
<td>8185.75</td>
<td>9647</td>
</tr>
<tr>
<td>Team 10</td>
<td>105.14</td>
<td>74.73</td>
<td>7524.37</td>
<td>9268</td>
</tr>
</tbody>
</table>

### Table 2. Correlation matrix between spatial components

<table>
<thead>
<tr>
<th></th>
<th>Team length</th>
<th>Team width</th>
<th>Team surface area</th>
<th>Average distance run</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team length</td>
<td>1.0000</td>
<td>-0.3159</td>
<td>0.8954</td>
<td>0.1772</td>
</tr>
<tr>
<td>Team width</td>
<td>-0.3159</td>
<td>1.0000</td>
<td>0.0404</td>
<td>-0.2716</td>
</tr>
<tr>
<td>Team length</td>
<td>0.8954</td>
<td>0.0404</td>
<td>1.0000</td>
<td>0.1693</td>
</tr>
<tr>
<td>Average distance run</td>
<td>0.1772</td>
<td>-0.2716</td>
<td>0.1693</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

From Table 1, a correlation matrix is created to see the relationship between the metrics.

Table 2 shows the correlation matrix and Figure 1 shows a visualization of the correlation
matrix. Table 2 shows a strong positive correlation between team length and surface area with a value of 0.8954. There is a slight negative correlation between team length and width with a value of -0.3159. Team width and average distance run show a very weak correlation of 0.0404 indicating that there isn’t a clear relationship between the width of the team's field and the average distance run by the team.

![Figure 1](image)

**Figure 1. Visualization of correlation matrix between spatial components**

**B. k-means**

Figure 2 shows the plot generated by the silhouette method to find the optimal value of k. Since the plot shows that k = 2 has the highest silhouette score, k = 2 was the first k value chosen for modeling the k-means algorithm. This suggests that the data is most naturally partitioned into two distinct groups. In the context of the k-means algorithm, this means that the algorithm has found two centroids such that the average silhouette score across all data points is maximized when the data is divided into two clusters.
Figure 2. Silhouette method for the optimal value of k clusters

Figure 3 shows the plot generated when the elbow method was used to determine the optimal k value. k = 4 was observed as the point in which the rate of change decreases thus indicating the elbow of the plot. Therefore k = 4 was chosen as the second value to run the k-means algorithm.

Figure 3. Elbow method for the optimal value of k clusters
The k-means algorithm was run with values of 2, 4, and 11 as mentioned before. Figures 4 through 6 show the visualizations of the k-means results with each k value respectively. With each figure, it is possible to derive different interpretations based on the number of clusters within the plot.

Figure 4. k-means clustering with 2 clusters (k = 2)

As the clustering was done on soccer player positions based on their locations on the field, Figure 4 shows that two clusters might suggest a division between offensive and defensive players.
Figure 5. k-means clustering with 4 clusters (k = 4)

With 4 clusters, Figure 5 suggests that there may be a clear division between 4 groups of players. Cluster 1 is positioned on the right side of the field and may represent players who are primarily active in that area on the field. Given the linear distribution along the Y-axis, this could correspond to players who operate along the right flank, such as right defenders, right midfielders, or right-wingers. Players in Cluster 2 are centrally located in the lower half of the pitch. This could suggest that there is a concentration of defensive players. Cluster 3 occupies the upper central part of the plot; this cluster might indicate attacking players such as strikers or attacking midfielders. Lastly, players in Cluster 4 could represent left defenders, left midfielders, or left-wingers. They would have responsibilities mirroring those of the right-sided players but focused on the left side of the field.
Figure 6. k-means clustering with 11 clusters (k = 11)

With k = 11, it is possible to distinguish 11 clusters of players on the field. Since these clusters correlate to known specific soccer positions, the average total distance run per cluster was derived for additional context. Table 3 shows the average total distance run per player cluster.

Table 3. Average total distance run per cluster of players

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Average Total Distance (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11208</td>
</tr>
<tr>
<td>2</td>
<td>7522</td>
</tr>
<tr>
<td>3</td>
<td>10519</td>
</tr>
<tr>
<td>4</td>
<td>10902</td>
</tr>
<tr>
<td>5</td>
<td>11461</td>
</tr>
<tr>
<td>6</td>
<td>11003</td>
</tr>
<tr>
<td>7</td>
<td>10043</td>
</tr>
<tr>
<td>8</td>
<td>11102</td>
</tr>
<tr>
<td>9</td>
<td>10889</td>
</tr>
<tr>
<td>10</td>
<td>10937</td>
</tr>
<tr>
<td>11</td>
<td>10533</td>
</tr>
</tbody>
</table>
C. k Nearest Neighbors

The algorithm was run with the k value incrementally increasing. Figure 7 shows the accuracy of the kNN model increases until k = 19 which the results were accuracy = 8.73%, precision = 7.89%, recall = 8.63%, and F1 score = 8.03%. Table 3 shows the accuracy, precision, recall, and F1 scores of each model with different k values. The four metrics all followed the same trend of increasing until k = 19.

Figure 7. kNN accuracy vs k values
Table 4. Accuracy, Precision, Recall, F1 scores for each k value (kNN)

<table>
<thead>
<tr>
<th>k</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0586</td>
<td>0.0545</td>
<td>0.0547</td>
<td>0.0553</td>
</tr>
<tr>
<td>3</td>
<td>0.0625</td>
<td>0.0583</td>
<td>0.0582</td>
<td>0.0590</td>
</tr>
<tr>
<td>4</td>
<td>0.0662</td>
<td>0.0611</td>
<td>0.0609</td>
<td>0.0618</td>
</tr>
<tr>
<td>5</td>
<td>0.0697</td>
<td>0.0644</td>
<td>0.0642</td>
<td>0.0650</td>
</tr>
<tr>
<td>6</td>
<td>0.0729</td>
<td>0.0671</td>
<td>0.0669</td>
<td>0.0681</td>
</tr>
<tr>
<td>7</td>
<td>0.0754</td>
<td>0.0690</td>
<td>0.0690</td>
<td>0.0701</td>
</tr>
<tr>
<td>8</td>
<td>0.0774</td>
<td>0.0706</td>
<td>0.0711</td>
<td>0.0713</td>
</tr>
<tr>
<td>9</td>
<td>0.0795</td>
<td>0.0725</td>
<td>0.0730</td>
<td>0.0731</td>
</tr>
<tr>
<td>10</td>
<td>0.0808</td>
<td>0.0735</td>
<td>0.0750</td>
<td>0.0744</td>
</tr>
<tr>
<td>11</td>
<td>0.0815</td>
<td>0.0740</td>
<td>0.0759</td>
<td>0.0749</td>
</tr>
<tr>
<td>12</td>
<td>0.0820</td>
<td>0.0745</td>
<td>0.0768</td>
<td>0.0754</td>
</tr>
<tr>
<td>13</td>
<td>0.0827</td>
<td>0.0749</td>
<td>0.0778</td>
<td>0.0760</td>
</tr>
<tr>
<td>14</td>
<td>0.0833</td>
<td>0.0754</td>
<td>0.0792</td>
<td>0.0766</td>
</tr>
<tr>
<td>15</td>
<td>0.0842</td>
<td>0.0761</td>
<td>0.0804</td>
<td>0.0774</td>
</tr>
<tr>
<td>16</td>
<td>0.0847</td>
<td>0.0765</td>
<td>0.0813</td>
<td>0.0778</td>
</tr>
<tr>
<td>17</td>
<td>0.0852</td>
<td>0.0768</td>
<td>0.0822</td>
<td>0.0782</td>
</tr>
<tr>
<td>18</td>
<td>0.0860</td>
<td>0.0776</td>
<td>0.0834</td>
<td>0.0791</td>
</tr>
<tr>
<td>19</td>
<td>0.0873</td>
<td>0.0789</td>
<td>0.0863</td>
<td>0.0803</td>
</tr>
<tr>
<td>20</td>
<td>0.0861</td>
<td>0.0773</td>
<td>0.0835</td>
<td>0.0787</td>
</tr>
<tr>
<td>21</td>
<td>0.0859</td>
<td>0.0771</td>
<td>0.0842</td>
<td>0.0787</td>
</tr>
</tbody>
</table>

D. Support Vector Machines

The SVM model outputs a classification into 55 classes with 258053 support vectors. The model used the classes to represent the different categories that the model is trying to predict. Upon clarification, these classes indicated a highly granular level of categorization for player roles or strategies.
V. Discussion

The main hypothesis driving this thesis was that player tracking data when analyzed with advanced machine learning methods, can provide a standalone, in-depth understanding of soccer matches. This study explored this possibility using various machine learning techniques such as k-means, k-Nearest Neighbors (kNN), and Support Vector Machines (SVM). Out of the analysis performed, k-means and spatial analysis components showed the most promise when applying only x, and y position data whereas kNN and SVM showed more limitations with the same data.

In the realm of sports analytics, particularly in soccer, the application of machine learning algorithms such as k-means clustering, kNN, and SVM can offer valuable insights into player behavior and team formations. While having more data in terms of additional game data or more data on single teams may impact the results of supervised learning algorithms, the lack of additional variables outside of x, y position data severely limits the scope of possibilities for each model. Despite these limitations, each approach showed promise and offered interesting possibilities that could impact decision-making processes for coaches and managers.

Exploratory data analysis provides some interesting insights without the use of any machine learning models. In the context of soccer player data, team length, width, and surface area derived from convex hull analysis can provide significant insights into a team's operational and tactical behavior on the field. Team length reflects the vertical stretch of the team across the field. A longer length indicates a team playing a more direct style of soccer, utilizing long passes and a more spread-out formation to cover more
ground (Castellano et al., 2013; Duarte et al., 2013). It can also suggest that the team is utilizing the depth of the field, potentially to exploit spaces behind the opposition's defense. Team width is indicative of how much the team is using the sides of the field. Teams with greater width can indicate tactical strategies that leverage wide players, such as wingers or fullbacks, to create crosses into the penalty area or to stretch the opponent's formation horizontally, thereby creating spaces in the central areas of the field. The combination of these factors is seen in the surface area of each team shown in Table 1. A large surface area could suggest a team that is more spread out, possibly playing a possession-based game with players providing options across the field (Castellano et al., 2013). Conversely, a smaller area might indicate a compact team that focuses on tight spaces to maintain defensive solidity (Castellano et al., 2013) and not allow the opposition to shoot the ball as teams are more likely to concede shots when they are not compact enough (Moura et al., 2013). The correlation matrix in Figure 1 shows that team length and average distance run have a moderately negative correlation, indicating that as the team length increases, the average distance run tends to decrease slightly. This might seem counterintuitive but could suggest that on larger fields, play is more spread out and may involve more passing than running. Larger playing areas might lead to different styles of play that don't require as much running, or it could reflect a strategic approach where possession and position take precedence over covering distances. It's important to note that correlation does not imply causation. The relationships observed here can be due to various factors and would require further investigation to understand the underlying causes which as of now I can only draw limited interpretations in the absence of additional data.
To build robust and reliable models it is important to choose the right k value for k-means. A silhouette score greater than 0 indicates that the data fits the best within the value k for clusters (Wijngaard, 2020). Figure 2 shows that the silhouette method indicates that the optimal number of clusters for the dataset is 2, suggesting that the data is most naturally partitioned into two distinct groups. In the context of the k-means algorithm, this means that the algorithm has found two centroids such that the average silhouette score across all data points is maximized when the data is divided into two clusters. As the clustering was done on soccer player positions based on their locations on the field, two clusters might suggest a division between offensive and defensive players. When examining the scatterplot of the clusters, Figure 4 suggests that there is a distinct separation along the vertical y-axis for both clusters. This suggests that while there is a clear division between the players, there isn't a significant distinction in terms of players' forward or backward positioning on the field within each cluster. In other words, players in Cluster 1 are spread out across the forward positions towards the opponent's goal. This can hint at a strategic approach where players on each side cover either the forward or back part of the pitch, with not much overlap between offensive and defensive players.

The elbow method suggests that k = 4 may be more suitable compared to k = 2 suggested by the silhouette method. With 4 clusters, Figure 5 suggests that there may be a clear division between 4 groups of players. These clusters could represent a group of positions that have similar movements. Cluster 1 is positioned on the right side of the field and may represent players who are primarily active in that area on the field. Given the linear distribution along the Y-axis, this could correspond to players who operate along the right flank, such as right defenders, right midfielders, or right-wingers. Their
role typically involves covering the right side of the pitch, both in defensive duties and in supporting the attack. Players in Cluster 2 are centrally located in the lower half of the pitch. This could suggest that there is a concentration of defensive players such as center backs, goalkeepers, or defensive midfielders. These players usually have a role that includes distributing the ball, breaking up opposition play, and linking the defense with the offense. Cluster 3 occupies the upper central part of the plot; this cluster might indicate attacking players such as strikers or attacking midfielders. Their positioning further up the field suggests a primary focus on creating scoring opportunities and attempting to score goals. Lastly, players in Cluster 4 could represent left defenders, left midfielders, or left-wingers. They would have responsibilities mirroring those of the right-sided players but focused on the left side of the field. Having a lower number of clusters allows for coaches to identify positions that play similarly. This could impact how they adjust the team’s lineup depending on the players available.

Apart from the two k values found from the silhouette method and elbow method, the context of the sport must also be considered. Soccer is a game played with 11 players on the field. When performing position analysis, I must consider that having 11 clusters to essentially divide the players into 11 different groups based on their positions on the field may yield optimal results. When running k-means with 11 clusters, each cluster represents a group of players who are frequently found in specific regions of the field. Figure 5 shows a clear division of different positions that operate in different parts of the field. It is possible to make out various positions such as goalkeepers, centerbacks, wingbacks, defensive midfielders, attacking midfielders, wingers, and strikers. This is also supported by the total distance run by the cluster in Table 3. The average total
distance run by cluster ranges approximately between 10000 to 12000 meters which corresponds to previous findings by Di Salvo et al. (2006) regarding total distance run by position per match. Cluster 2 was observed as the exception and by conferring with the location of each cluster in Figure 6, I inferred that Cluster 2 is most likely referring to goalkeepers which explains the low average distance run compared to other clusters.

While it may make sense contextually to run the model with 11 clusters, with a high-value k, there's a risk that the clustering is too granular, capturing noise in the data as if it were a meaningful pattern. This can lead to overfitting, where the model is tuned to the idiosyncrasies of the dataset and may not generalize well to new, unseen data. From a practical standpoint, managing and utilizing the results of 11 different clusters can be cumbersome. It might not offer a significant advantage over a model with fewer clusters that still capture the main variations in the data. As the number of clusters increases, each additional cluster may contribute less to the overall explanation of variance in the data. Beyond a certain point, adding more clusters doesn't provide additional value. This is why I recommend the highest value of k to be tested to be no larger than 11 which corresponds to the number of players on each team during a soccer match.

With the dataset only having x, and y coordinate data, it is normally inadvisable to run kNN. However, running kNN on individual players rather than predefined player positions can lead to some interesting possibilities. Running kNN on individual players can focus on identifying players with similar attributes or performances, regardless of predefined positions. All players’ attributes would be treated equally in determining similarity, and position-specific attributes may be less emphasized. This would lead to the
results highlighting players with similar attributes or performance levels, which could be useful for scouting, player comparison, or team formation strategies. Coaches could be able to find players capable of running a specific tactic potentially impacting substitutions and real-time strategy adjustments.

When interpreting the results of the confusion matrices, low accuracy was common in all the metrics with the highest accuracy being 8.73%. This indicates that either the results suggest that no player is similar to another player or that the model is ineffective due to the lack of relative player position labels. While the percentages may be low, each metric is important for real-world scenarios. Precision is particularly important when the cost of a false positive is high, such as when scouting for a player in a specific position where misclassification could lead to wasted resources. This is particularly useful when coaches and scouts need to have a balanced view of both false positives and false negatives. In soccer, a high F1 score for a position would indicate that the model effectively identifies players of that position without falsely classifying other players into that position. Clubs might look at the F1 score to balance the precision and recall in the evaluation of players, ensuring a fair trade-off between the quality of player predictions and the coverage of potential candidates. This is crucial when it's important not to miss any player with a certain play style. While the kNN algorithm holds potential for classifying soccer player positions from spatial data; however, its performance is shown to be highly contingent upon proper hyperparameter tuning and thorough model evaluation.

The SVM model outputs a classification into 55 classes, which initially were misconstrued as clusters. Upon clarification, these classes indicated a highly granular
level of categorization for player roles or strategies. The large number of support vectors, 258053, hinted at a complex dataset with nuanced distinctions between classes. It suggested that players exhibited subtle variations in their playing style or role, which the model could discern. The interpretation of a 55-class output presented challenges.

Ensuring a balanced dataset for effective model training was critical, as was assessing the model's performance given the risk of overfitting. Given the large number of classes, it is possible to conclude that a potential imbalance in class representation needed to be addressed to prevent bias in the model's predictive capabilities further suggesting the need for additional variables. Another factor to consider is the time required to run SVM models. The models require a lot of time to run and from a practical standpoint for coaches and analysts, it may not be a feasible option to run during the season. However, it has possibilities for scouting and player recruitment strategies where teams can identify and recruit players whose playing style complements their strategic approach.

Machine learning has significantly impacted player performance analysis and game strategy optimization in sports analytics. By enabling advanced analysis of player tracking data, biometric data, and historical performance records, machine learning algorithms provide valuable insights for teams and coaches. Moreover, by analyzing opponent data and real-time game situations, machine learning models optimize game strategies and facilitate data-driven decision-making. Continued advancements in machine learning techniques and the integration of emerging technologies hold promise for further enhancing sports analytics.
VI. Conclusion and Future Works

Each algorithm brings a different approach to analyzing positional data. k-means can provide a high-level overview of formations, kNN can classify player roles within known tactical setups, and SVM can attempt to draw distinctions between different tactical structures. Practically, these algorithms can support match analysts and coaches by providing an overview of player distribution and potentially identifying key tactical structures. They can highlight the spatial organization of players and indicate which areas of the field are most occupied or left vulnerable. In training scenarios, these algorithms can assist in evaluating the effectiveness of specific tactical drills designed to maintain or break certain formations. Nonetheless, the fluid nature of soccer and the continuous movement of players mean that any insights derived from these algorithms must be interpreted cautiously and in conjunction with expert knowledge. Coaches and analysts should use the outputs of these algorithms as one of several tools for tactical analysis, not as definitive solutions. The lack of event data further complicates matters, as it strips away the context in which movements occur, making it more challenging to draw concrete conclusions about player roles, responsibilities, and team strategies. Therefore, while the application of k-means, kNN, and SVM to soccer player x, y position data can offer valuable insights, these should be viewed as complementary to broader tactical analysis rather than standalone verdicts.

The question this study aimed to answer was if x, y position data can track formation changes for the game of soccer. While it is possible to track player movement and gain insights from spatial analysis, it lacked the additional depth that could have been provided with the presence of event data. However, an interesting possibility can be seen
in which the results of k-means clustering may be used in the place of actual player positions. While there may be cases in which players are incorrectly classified into a certain cluster, it would be interesting to compare the results of running supervised algorithms with k-means clustering defined positions. In addition, incorporating event data into the analysis of soccer player movements using machine learning algorithms like k-means clustering, k-Nearest Neighbors (kNN), and Support Vector Machines (SVM) can significantly enrich the scope and depth of analytical insights. With event data, it’s possible to track how formations evolve in response to different in-game situations, such as attacking phases, defending scenarios, or counterattacks. Having access to event data can also provide context to the positioning, helping to understand whether a particular spatial arrangement is a defensive setup, an attacking formation, or a transitional phase. The incorporation of event data opens up a myriad of possibilities for deeper, more contextual, and actionable insights into soccer analytics. Future work can span from advanced tactical analysis to individual player development, injury prevention, and enhancing fan experiences. As data availability and computational power continue to grow, the potential applications in sports analytics are only limited by the creativity and domain knowledge of researchers and analysts in the field.
BIBLIOGRAPHY


