OPTIMIZING FOOTBALL LINEUP SELECTION USING MACHINE LEARNING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

YILMAZ TAYLAN GÖLTAŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

SEPTEMBER 2023

Approval of the thesis:

OPTIMIZING FOOTBALL LINEUP SELECTION USING MACHINE LEARNING

Submitted by YILMAZ TAYLAN GÖLTAŞ in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
Dean, **Graduate School of Informatics** ______________________

Prof. Dr. Altan Koçyiğit
Head of Department, **Information Systems Dept., METU** ______________________

Prof. Dr. Sevgi Özkan Yıldırım
Supervisor, **Information Systems Dept., METU** ______________________

Assoc. Prof. Dr. Mustafa Söğüt
Co-Supervisor, **Physical Education and Sports Dept., METU** ______________________

Examining Committee Members:

Prof. Dr. İbrahim Soner Yıldırım
Computer Education and Instructional Technology Dept., METU ______________________

Prof. Dr. Sevgi Özkan Yıldırım
Information Systems Dept., METU ______________________

Assist. Prof. Banu Yüksel Özkaya
Industrial Engineering Dept., Hacettepe University ______________________

Date: **05.09.2023**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Yılmaz Taylan Göltaş

Signature : __________________

ABSTRACT

OPTIMIZING FOOTBALL LINEUP SELECTION USING MACHINE LEARNING

Göltaş, Yılmaz Taylan
MSc., Department of Information Systems
Supervisor: Prof. Dr. Sevgi Özkan Yıldırım
Co-Supervisor: Assoc. Prof. Dr. Mustafa Söğüt

September 2023, 96 pages

Football has both the biggest economy and the largest audience in the sports world. Billions of dollars change hands every year in line with the decisions made in the sports economy. With the growth of the economic reflections of data decisions, decision systems have become more open to analytical approaches as in other sports. Thanks to increasing data types and developing semi- and fully automated data collection systems, data about both teams and players have become diverse and accessible. Increasing data opportunities have paved the way for on-field and off-field decisions in football to be solved with data-centred approaches. Team selection is one of these decisions. Traditionally, football coaches make this decision by analyzing players' match and training performances and by analyzing the data of the opposing team. In this thesis, a new solution to the team selection problem is proposed with a data-driven approach by using the match data of the players and teams, grouping the players based on their positions and roles, considering the opposing team, tactical formation and environmental factors.

Keywords: Football, Line-up, Decision Support, Optimization, Player Roles

ÖZ

MAKİNE ÖĞRENİMİ KULLANARAK FUTBOLDA KADRO SEÇİMİNİ OPTİMİZE ETME

Göltaş, Yılmaz Taylan

Yüksek Lisans, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Prof. Dr. Sevgi Özkan Yıldırım

Tez Ez Danışmanı: Doç. Dr. Mustafa Söğüt

Eylül 2023, 96 sayfa

Futbol spor dünyası içerisinde hem en büyük ekonomiye hem de en geniş izleyici kitlesine sahip spor. Her yıl milyarlarca dolar spor ekonomisinde verilen kararlar doğrultusunda el değiştirmekte. Verilerin kararların ekonomik yansımalarının büyümesiyle birlikte karar sistemlerinin diğer sporlarda olduğu gibi analitik yaklaşımlara daha açık olmaya başladı. Artan veri çeşitleri ve gelişen yarı ve tam otomatik veri toplama sistemleri sayesinde hem takımlar hem de oyuncular hakkında veriler çeşitli ve ulaşılabilir hale geldi. Artan veri olanakları futbol içerisinde saha içi ve saha dışı kararların veri merkezli yaklaşımlarla çözülmesinin önünü açtı. Takım seçimi de bu kararlardan biri. Geleneksel olarak futbol antrenörlerinin oyuncuların maç ve antrenman performanslarını inceleyerek ve rakip takımın verilerini inceleyerek verdiği bir karar. Bu tez çalışmasında takım seçimi problemi oyuncuların ve takımların maç verileri kullanılarak, oyuncuları pozisyon ve rolleri üzerinden gruplandırarak, rakip takım, taktiksel diziliş ve çevresel faktörleri göz önünde bulundurarak veri temelli bir yaklaşım ile yeni bir çözüm önerilmiştir.

Anahtar Sözcükler: Futbol, Takım Seçimi, Karar Destek Mekanizmaları, Optimizasyon, Oyuncu Rolleri

To My Family and My Lovely Wife

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor, Prof. Dr. Sevgi Özkan Yıldırım and co-advisor Assoc. Prof Dr Mustafa Söğüt for their guidance and valuable contributions throughout this thesis. I am also grateful for Head Coach Ömür Altunsöz valuable support and suggestions that helped me get deepen to this field and find the opportunity to improve myself.

Besides my supervisor, I would also like to thank my dear friends Orkun Temiz, Volkan Ertaş, Berkan Cengiz, Deniz Göktaş and Talha Mehmetoğlu for being my companions and for the encouragement they gave me in this difficult journey.

I'd like to thank my dearest mother, father, and my dear sister for always being there for me, trying for my happiness, and picking me up when I was down.

Last but not least, I would like to thank my biggest source of support in this life, my beloved wife Burcu Koç Göltaş. It was not possible for me to complete this process without her. I am excited and impatient for the new roads we will take, the new adventures we will embark on and the steps we will climb in our career together with my wife, who has been with me in every difficult moment and has always provided me with the necessary strength and support to go further. I hope she will always be with me today and in the future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AHP	Analytical hierarchy process
CBR	CatBoost Regressor
CD	Central Defender
CM	Central Midfielder
DM	Defensive Midfielder
FW	Forward
FW-NMF	Feature Weighted Non-Negative Matrix Factorization
GBM	Gradient Boosting Machine
GBR	Gradient Boosting Regressor
GPS	Global Positioning System
ICC	Intraclass Correlation Coefficient
KPI	Key Performance Indicator
LD	Left Defender
LGBM	Light Gradient Boosting Machines
LM	Left Midfielder
LSTM	Long-Short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCDA	Multi-Criteria Decision Aid
MLP	Multi-Layer Perceptron
NMF	Non-Negative Matrix Factorization
NMT	Neural Machine Translation
PCA	Principal Component Analysis
PL	Premier League
RD	Right Defender
RM	Right Midfielder
RMSE	Root Mean Square Error
SGD	Scholastic Gradient Decent
SVM	Support Vector Machine
UEFA	Union of European Football Associations
WandB	Weights & Biases
xA	Expected Assist
xG	Expected Goal

CHAPTER 1

INTRODUCTION

1.1 Research Background

Football, also known as soccer, is frequently cited as having the largest economic impact of all sports. There are billions of fans of the sport around the world, which has a huge global following. Football is a big actor in the sports economy as a result of its popularity, which generates enormous economic activity. The annual review of football finance published by Deloitte [1] indicates that the football sector has seen significant income increase in recent years. Several causes, including television rights, sponsorship agreements, ticket revenues, and item sales, are responsible for this rise. The stakes rise along with the scale of the football economy, and decision-making procedures become increasingly vulnerable to external factors. With the increase in the economic impact of every decision, the approach in decision-making mechanisms has started to move away from the traditional form and transform into data-based forms with the effect of developing technology. In this field, the Moneyball concept [2] published by Lewis in 2003 and its success in the American National Baseball League caused a crossroads in decision making systems in football and all other sports economies. With the rapid development of technology and the automation of data collection systems [3], data-centered systems have become the basis of decision-making mechanisms for both football and other sports. The use of scientific decision-making in conjunction with real-time sports data monitoring networks has improved the management of sports in terms of policy and training decisions. This helps players to effectively observe opponents' pregame information during games and to make logical decisions to counter the opponent's offense [4].

Football line-up decisions are crucial since they can affect a team's performance, player morale, and even legal implications. One of the toughest challenges facing football coaches is choosing the starting lineup, one of the most important requirements for success [5]. When selecting the starting lineup, there are several things to consider, such as team strategies, opposition teams, and environmental circumstances. The fundamental responsibility of football coaches is to assess each of these factors and choose the best players for their squads. Although line-up selection is of great importance for football and for the success of the football economy, the fact that it depends on too many variables and that it is not easy to parametrize these variables has not yet made it popular for data-driven decision-making. Studies in this field are quite limited both in literature and in the football industry. In this thesis, I propose a novel approach to this field from the perspective of player roles and a data centric approach to optimize line-up selection based on opponent, tactical and environmental factors.

1.2 Research Aim and Objectives

This research aims to fulfil the following objectives while approaching the line-up decision problem in the football industry with a data centric approach.

1. Determination of the ideal data type and method for the line-up selection problem.

2. Data centric determination of the ideal team selection specific to each match.

3. Determining the relationship between player performance and team performance.

4. Determining the interaction of player roles with each other and with team performance.

5. Determining the contribution of expert opinion to data-driven sports analytics when determining the defining factors of players and teams.

6. Determining the effect of tactical formation, tactical formation of the opposing team and line-up preferences on team performance.

7. Determining the effects of parameters such as environmental factors, referee, and match score on team performance.

1.3 Significance of the study

The proposed approach for optimal line-up selection is a completely new and original approach for the literature. In literature, studies in this field consist of approaches that evaluate teams independently of players [5],[6], focus on a single formation for a single team [7], or analyze teams independently of opponents [8]. The approach proposed in this study determines the ideal team selection by considering the effects of the opponent team effect, tactical formation effect, referee, score and environmental conditions, which are not present in any of the studies in the literature.

1.4 Structure of the Thesis

The thesis consists of 5 chapters. The following chapter provides an overview of data-oriented studies in the field of sport in literature. Chapter 3 describes the research methodology. In Chapter 4 the proposed methodology is tested with real word data and in Chapter 5 the contribution of the outputs to the literature, the limitations of the study and how it can be improved are discussed.

CHAPTER 2

LITERATURE REVIEW

Recently, technology and statistics have had an increasing impact on sports. In the past, sports-related data primarily appeared as a way to create in-depth and intelligent commentary. Data are now one of the most valuable resources in sports, offering excellent information for the development of the entire sports business. Moneyball: The Art of Winning an Unfair Game by Michael Lewis [2] demonstrates the increasing popularity of statistical analysis in sports. With the development of data analysis methods, it has become easier to handle tasks that require attention as optimization tasks in sports analytics, such as player selection, optimal player lineup, player ranking, and player responsibilities [9]. Data-based solutions have been developed in many sports disciplines due to the success of data-driven methodologies in sports. The effect of the data on the results grew substantially as the methodologies and approaches advanced.

Although football is the most popular sport in the world [10], it is not as popular as other sports in data-driven studies because of its complex and low-scoring structure compared to other sports. The ability to conduct data-driven research in the field of football has been made possible by recently developed data collection methods (wearable technologies such as TRACKTICS [11] and the diversity of data available from data providers (WyScout, InStat, and Opta).

Football has become increasingly popular as a subject of study in sports analytics research, and investments in this area have risen significantly due to the high return potential of machine learning models with rich data sources [12]. Due to its popularity, research on football in sports analytics is broken down into sub-specialties, which are explored in more detail in the subsequent sections. This review focuses on football-related studies, but practices and approaches in other sports have also been covered.

2.1 Moneyball Concept and Sports Analytics

Once the concept of "Moneyball" emerged and became widespread, it was viewed as an opportunity for inefficient markets, such as football [13]. The Moneyball concept can be broadly classified as a MCDA(multi-criteria decision aid) problem. The MCDA approaches offer satisfactory solutions when more than two alternatives are evaluated against more than two performance criteria [14]. The Moneyball model benefits significantly from the probabilistic approach. Sampling introduces uncertainty into quantitative models as human intervention is still necessary for data gathering [14]. Because the numbers in a football game are random, it gives more credibility to the probabilistic approach than to deterministic approaches.

After the Moneyball concept became widespread and it was mathematically demonstrated that it could deliver meaningful results in sports such as baseball, almost all industries have given substantial attention to sports analytics, multidimensional data analysis, big data, and predictive business analytics to improve services for their stakeholders. Sports organizations, websites, broadcasters, and online platforms increasingly use statistical and predictive analytics to discover player insights, scoring patterns, and comparison-based professional player selection. Such methods have attracted the interest of many scholars for player performance evaluations and the prediction of optimal solutions because of their persistence in real-time dynamic applications and their complexity [15].

2.1.1 Data Sources

Advances in technology have led to more sophisticated data collection techniques. In addition to well-known sports data vendors [16],[17],[18], crowdsourced data are now available [19],[20]. While semi-automated and automated data-gathering systems have significantly increased the quantity of data collected, wearable technologies and movement analysis systems have significantly widened the diversity of the data collected [3]. Additionally, as a result of both the employment of more sophisticated technology by data providers and the exponential growth of the football industry, the quality of data on the football field is continually improving [21]. Such a growing dataset in every aspect makes it possible to investigate the link between performance and success. A team's success can be defined as its performance in a tournament, and the success of an individual player can be defined as his popularity level or market value [21].

Football-related data can be divided into three categories. These include performance data from wearable devices, in-game statistics, and spatiotemporal event data. Spatiotemporal data is a type of data that is created by capturing all the events in the game by tagging software and processing them together with GPS (Global Positioning System) and video surveillance data. It also records the players' movements in a match with location and time information [21]. This makes discretizing the match and examining its sub-events possible [10]. In-game statistics show the sum of certain events in a match grouped by in-match segments. On the other hand, physical data is data about players' physical conditions, such as sprinting and running speed, obtained directly from players using wearable sensing devices or video surveillance data. In addition to these data types, advanced statistics, such as expected goals (xG) and expected assists (xA) derived from in-game data, are also available. These datasets can be used together or separately, depending on the scope and objectives of the study.

2.1.1.1 Expected Goals

As football is one of the lowest-scoring games compared to other sports, it is crucial to take advantage of scoring opportunities. Teams have a significant competitive edge when they have predictions and statistics for every position that can result in a goal. As football is a low-scoring sport, evaluating a player's true worth might be misleading based on how many goals they score [22]. For this reason, Sam Green [23] created the expected goal metric in 2012 to measure the probability of a shot resulting in a goal. However, the expected goal concept first appeared in Vic Barnett and Sarah Hilditch's [24] research, published in 1993.

The expected goal is a metric that better reflects match performance than the score. In football, victories and losses are occasionally decided by a single goal; thus, randomness disproportionately influences match results. As a result, match results sometimes do not accurately reflect the level of play between the two teams on the pitch, and it is questionable whether match results are a reliable indicator of performance, especially when considering a limited range of matches in the context of a single season [25]. If result-based performance assessment is used in situations where random factors are the main determinants of the results of sports events, systematic misjudgment should be considered to occur [26]. The expected goal metric provides several features for addressing this problem. First, goal chances are far more frequent than goals, making them less vulnerable to game-related chance effects. Second, it considers several types of scoring opportunities. Any football game plan should include maximizing your opportunities and minimizing those of the opponent [26]. According to Green [27], the expected goal is the metric most suited to sports analytics based on these characteristics. The expected goal value represents a probability ranging from 0 to 1 and indicates the chance that each shooting opportunity will result in a goal [22]. This metric allows us to determine whether a player or team's performance in a match is above or below expectations [28].

The expected goal parameter demonstrated its explanatory value in the field of football in 2019 with Tippet's study "The Expected Goals Philosophy" [29], and it started to be employed in other sports branches [22]. Subsequently, Herold criticized the concept of expected goals and stated that the opposing team's movements should also be included in calculations [30]. In 2020, Brechot and Flepp [26] proposed a new expected goal metric by adding the distance, angle, rule setting, and body part of the shot to existing parameters. In 2022, Cavus and Biecek [31] created an "explainable expected goals" metric using an "explainable" AI technique to generate a precise expected goals metric for monitoring team and player performance. Expected goal excludes movements that do not result in a shot because it is a shot-based metric. In his study "Beyond Expected Goals," Spearman [32] presented a model that quantifies the impact of non-shooting positions and off-ball movements on goals.

2.1.1.2 Determination of Player Positions in Football

The number of players and use of predetermined playing patterns have encouraged greater specialization of player positions in football [33]. When specialized roles emerge in a sport, the concept of player position becomes even more crucial. As player positions become more specific, greater attention must be paid during selection to maintain the essential balance between player roles [34]. If the proposed combination of players fails, poor selection in a crucial match can cost significant money. The player selection problem can be made more complex by adding a position-decision problem for each player in the squad. It is the responsibility of the coach to build the team or find the right person for each position in the team [35]. For this reason, formation and position definitions should be clarified before the player selection phase. In the literature, studies to determine the position of players on the field have been formed to include the concepts of formation and position definitions as they affect each other. The on-field placement of players has been studied in the literature using two approaches. The former uses spatiotemporal data generated by motion detection and video tracking applications to calculate players' positions relative to their standard position, whereas the latter groups player positions according to the formation and player responsibilities on the pitch.

2.1.1.3 Determination of Player Positions Using Spatiotemporal Data

Methodologies for determining player positions have shifted to data-centric approaches after spatiotemporal data became available. Frey [11] combined the TRACTICS tracking system with GPS and motion data to determine player positions in the field and examined player movements for 28 football matches. He divided the players into five positions using machine learning tools based on their typical locations in the field: center back, wing back, wing, midfield, and forward.

In a more comprehensive study, Pappalardo et al.[36] proposed a data-driven system called PlayRank to rank players. They used data from matches played over four seasons in five major European football leagues. As a result of the study, players were divided into eight groups according to their typical field locations. Figure 3 shows the distribution of the player positions and clusters. In light of the symmetrical distribution of players in the field, results consistent with those in Frey's study can be obtained by combining–C1-C4, C5-C7, and C6-C8 clusters [11]. Another study [37] used spatiotemporal data of 9300 matches from various European leagues and applied particle swarm optimization, a genetic clustering algorithm, to reproduce Pappalardo's findings.

Excluding the goalkeeper, team players are usually classified into one of the following five positions: central defender, wing-back, center-midfielder, wide midfielder, or forward [38],[39],[40],[41],[42],[5].

2.1.1.4 Determination of Player Positions Using Football Terminology

Another method for identifying player positions is to use football terminology. This method considers players' on-field responsibilities and the locations of the fields in which they fulfill these responsibilities. The player's position in the field also provides information regarding the criteria by which the player's performance will be evaluated [5]. In this method, excluding the goalkeeper, players are typically divided into three positions: forward, midfield, and defense [33],[35],[43]. In numerous studies in literature, it has also been preferred to use positions defined by well-known video game series such as Football Manager and FIFA [44],[45].

2.1.2 Player Roles in Football

Depending on the above characteristics, each team has a specific composition, style of play, and profile of players needed for each position [45]. In football, player roles and positions are two ideas that are sometimes used interchangeably. However, while the concept of player role refers to a player's on-field tasks, player position refers to the location in the field where these tasks are performed. The roles of two players in the same position may vary. Although players are often categorized into attacking, defending, wing-back, or other traditional positions, in practice, player types have much more complex subdivisions than these groupings. In professional sports, players' roles are influenced by team composition, playing style, and coaches' expectations [46].

The identification of player roles has been the subject of numerous studies. Based on player roles in the Football Manager game series, Aalbers [44] identified 21 roles with key identification parameters and categorized real players according to these 21 roles. In another study, Ghar [45] used player positions specified in the FIFA video game series to assess team cohesion and performance using a data-driven methodology. García-Aliaga et al. [47] proposed a position classifier model using 52 in-game statistics and taking the nine on-field positions defined by OPTA as a basis. Their research revealed that traditional positions did not provide sufficient specificity for player differentiation. Kalenderoğlu [48] suggested the use of hierarchical clustering methods to classify players according to their roles in their positions. Aydemir [28] grouped the players according to 16 roles for the player ranking model she proposed in her thesis.

In another study, Decroos and Davis [49] further suggested that instead of categorizing players into specific roles, it would be better for computer vision research and experts to express players as fixed-size vectors. A similar approach was adopted by Li [41] using the Non-Negative Matrix Factorization (NMF) method to investigate similarities between players' playing styles.

2.1.3 Usability of In-Game Stats

Game analysis is crucial for understanding and improving team sports performance by identifying elements contributing to success [50]. It is now possible to pinpoint the tactical and technical characteristics of teams related to their success or failure due to data's greater availability and reliability. Data from game analysis provide crucial key performance indicators for coaches and sports scientists that help assess, monitor, and prescribe ideal training programs [51],[52],[47]. In the literature, in-game statistics have been used to determine player performance, team performance, and the requirements of on-field positions.

2.1.4 Player Performance Representation Using In-Game Stats

Player transfers, a distinctive component of the financial dynamics of football teams, are used to increase revenue and improve team performance. As a result, accurate evaluations of player performance are crucial for teams' sporting and financial success when performing player transfers, and clubs spend a lot of money on these evaluations [28]. Player performance depends on athletic performance and technical skill. In-game statistics are reliable indicators of a player's technical ability [28],[53],[54],[5].

Numerous studies in literature use in-game statistics to assess player performance. Li [41] vectorized players according to their playing styles using in-game statistics such as shooting and dribbling passes. Brooks [10] used shot and pass statistics to rank players using the player ranking method proposed in his study. In his player performance prediction model, Toemen [55] employed in-game statistics from StatsBomb. Similarly, Pappalardo [36] proposed a player ranking methodology using the in-game statistics provided by Wyscout.

2.1.4.1 Stat Selection Based on Player Position

The question of which in-game statistics are more important for evaluating player performance has begun to be investigated because of the frequent use of in-game statistics in player performance evaluation studies. Due to the nature of football, the performances of a forward and a defender should not be evaluated using the same in-game statistics. Therefore, the in-game statistics used in data-driven studies should be customized according to player positions and roles.

Konefal [42] investigated the in-game statistics influencing the match results for five in-field positions (central defender, fullback, central midfielder, wide midfielder, and forward). According to the study's results, shot statistics for forwards are the parameters that most affect the match result. In contrast, the number of tackles for defenders and the number of passes for midfielders are the parameters with the highest correlation with the match result.

Using a similar approach, Ermidis [38] statistically analyzed the data provided by Opta for the 2015 Asian Cup to determine which in-field data were more descriptive for the same five field positions. The same approach was used to identify position specific KPIs using different datasets, such as Champions League matches between 2010 and 2016 [39] and the 2021 European Football Championship qualifying matches [40]. Similarly, Laasko [57] identified position-specific key performance indicators for attacking, midfielders, and defenders using similar statistical methods, focusing on young athletes.

2.1.5 Performance and Result Prediction

Football is the most popular sport in the world and the one with the largest economy. For example, The Premier League recently announced that it had extended its TV rights contract to £5.1 billion along with numerous other broadcasting companies [57]. In addition to teams and players, this economy rewards bookmakers, broadcasters, and companies serving fans. For clubs and other stakeholders in the football economy, the success of performance and result prediction efforts could result in significant financial benefits or losses [28]. Therefore, numerous studies in the literature have focused on predicting match results with betting and club success motivations. However, teams rely heavily on transfer revenues as sources of income. In 2019, the player transfer market for European clubs was worth €28.9 billion [58]. For all these reasons, it is of utmost importance for clubs to make the right player investments and comprehensively evaluate player performance [28].

Football matches can be unpredictable, and unexpected results often occur because of the game's low-scoring nature [25]. When predicting football match results, the first choice is to approach the problem as a classification - predicting the result, i.e., win, draw, or defeat - or as a regression - predicting the final score or predicting the total threat [59]. Although the majority of studies in the literature address this issue as a classification problem [60],[61],[62], some studies address result prediction in the context of the score [59],[62] and expected goal [28],[25] based regression problems. In the context of both regression and classification challenges, Bunker [63] proposed a roadmap to address result prediction problems. Peters [59] used Bunker's framework to address the issue of predicting the match result as a regression problem based on the total number of goals scored by the teams. He used clubs' lineups and playing styles to create different models for predicting home and away goals. The support Vector Machine and K-Nearest Neighbor models gave the best results in this study, where players and teams were represented in the model using data from the field [59]. In another study, Lindberg [62] addressed this problem from a classification and regression perspective. He used the LSTM(Long-Short Term Memory) model to predict a player's performance in a Fantasy Premier League game using player information and positions. Hubáček [60], who considered the problem within the context of classification, predicted the match result by using factors such as the match's location, ranking difference between the teams, form, and importance of the match. Stübinger [64] used player statistics from the FIFA video game combined with random forest, gradient boosting, support vector machine, and linear regression models to predict match scores regarding goals. He measured the economic returns of the outputs of his study using old odds. In his study, the random forest model was the most successful model, with an accuracy rate close to 81%, and the betting strategy provided an economic return of 1.58% per match.

Herbinet [25] uses the expected goal parameter as the target variable. While he used the ELO rating for team strength, he measured the teams' performance in the match based on the expected goals scored by the teams. He addressed the prediction of the match result from regression and classification perspectives, obtaining an F1 score of 0.382 with an accuracy of 0.51 for classification and a root mean square of 1.153 for regression.

2.1.6 Best Lineup Selection in Football

Lineup selection, one of the most significant prerequisites for success, is one of the biggest obstacles for football coaches [5]. There are numerous factors to consider when choosing the lineup, including team tactics, opposing teams, and environmental factors. The main task of coaches managing football teams is to evaluate all of these criteria and select the ideal players for the team. Best lineup selection often means more than just bringing together the best players in the team; it can be determined by examining the team's performance as a whole, not by the performance of its players [5],[6].

As it directly affects the result of the game, lineup selection has received the attention of researchers. There are many studies in the literature on ideal team selection, but only a few are related to football. Due to the scope of the study, only data-driven lineup studies are presented in this section. In his research, Cortez [5] combined training data collected with the help of wearable technologies with in-match data and determined the most important physical and technical parameters reflecting player performance for five positions using the recursive feature analysis method. The best player for each position was then chosen by ranking each player according to the technical and athletic standards necessary. He tried to validate the study's results by comparing them with the actual player preferences of coaches. The proposed model has the drawback of omitting the tactics used by the opposing team and environmental factors. In another lineup study, after determining the key performance indicators for each position using the analytical hierarchy process (AHP) method, the ideal team selection for the Fenerbahçe Club in the 4-4-2 formation was determined using an integer programming algorithm. The model used in this study also excludes the opposing team's tactics and environmental factors while also reducing tactical diversity to a single formation [7].

In another study, in-game statistics from 18 matches, determined by 50 experts, were used to evaluate player performance and select the ideal 11 for a 3-6-1 tactical formation. Contrary to the common belief in literature, all positions were assessed using the same statistics but ranked according to their own position [6]. On the other hand, [65], using a statistical method based on the frequency of repetition of moves in each position, identified the KPIs for each position, ranked the players by those rankings, and then chose the highest-ranked players for each position to make up the ideal team composition. Tavana [8] proposed a lineup selection method using a two-stage fuzzy method in another approach. He initially ranked the players based on position-specific key performance indicators. In the second stage, he evaluated team cohesion by analyzing previous matches. In a similar study, the authors calculated the impact of tactical formations on performance using players' ratings of tactics in earlier games for five well-known formations using a binary integer programming model. This study is valuable for examining the relationship between tactical formation and lineup selection [66].

None of the lineup selection models and methods proposed for football in the literature offers a comprehensive viewpoint by considering tactical formations, environmental factors, and opposing team characteristics. While studies differ in the models used and how to identify KPI (Key Performance Indicator)’s that determine player performance, efforts are often directed toward optimizing the resources of a single team. Only a few studies have considered opposing teams and tactical formations as part of the lineup selection problem. Therefore, literature lacks a holistic perspective, and this study aims to fill this gap.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The main research objective of this study is to examine whether the result of a football match and team performance can be predicted by a data-driven method and to determine the best lineup selection for football coaches based on the opposing team, tactical preferences, score, and environmental conditions in a data-driven manner. The results of this study can be used to help the technical staff of a football team make lineup decisions both before and during the game. The study also aimed to develop robust statistical measurements, explore data fusion possibilities, and assess the use cases and applicability of various machine learning methods for such a challenging dataset. Research questions of the study are presented in this section, then elaborate on the research methodologies in more depth.

3.2 Research Motivation and Questions

Football is still the most popular sport, but it is discussed less in the literature on sports analytics and data-driven analysis than in other sports. Its complexity, low-scoring nature, and high randomness factor [26] make football more difficult to analyze than other sports. In football, it can be exceedingly challenging to predict how one player or factor affects the results of a match. In sports such as basketball, volleyball, or baseball, match results, and individual performances have a substantial correlation; however, in football, the correlation is much weaker. These sports are more straightforward to evaluate than football because they involve fewer players, and each player's performance directly affects the game's results.

On average, two or three goals are scored in a football match at the first league level [67],[68],[69], therefore, analyzing the 90-minute individual and cumulative performance of 22 players over such limited outputs may not yield accurate results. On the other hand, while having more players than football, other sports, such as rugby and American football, are distinguished by their gradual development and lack of continuous play. The continuous nature of football makes the cause-and-effect relationships in the game less evident and the relationship between the score and the actions that affect it less visible. Football has always been a challenging topic for data-driven studies because of all these factors, but pioneering studies from the past ten years and the ability of machine learning models to learn complicated linkages and interactions have accelerated research in this subject [28],[9],[62],[63].

The studies reviewed in Chapter 2 show that most studies in the field of football focus on specific areas such as player performance, team performance, KPIs that determine player and team performance, result prediction, athletic performance comparisons, and lineup decision models. Moreover, numerous studies in the literature attempt to predict the result of a game [63],[64],[59],[62],[60],[25],[61],[9],[12], but none of them have combined the effects of the opposing team, team tactical preferences, game characteristics, player performance, and environmental factors to determine how they affect the results of the game. Similarly, many models and approaches are developed in the literature for the lineup decision problem of football teams [5],[6],[7],[8],[65],[66]. However, in these studies, the ideal team is selected independently of the parameters of the opposing team, and only in some studies team tactics and the formation on the field are considered as a parameter affecting the lineup selection [6],[7],[66].

The literature review shows that despite the increasing number of studies on the proliferation of computational power and data-driven solutions, many unexplored areas remain in football research. Due to the competitive nature of sports and the potential financial implications of their outcomes, much of the research in this field is unfortunately not included in the literature.

This study aimed to identify the best lineup by considering all the variables a coach might consider when choosing a lineup, including the opposing team, score, environmental conditions, and team tactics. Table 1 displays the data and derived parameters utilized to determine the ideal lineup selection.

Table 1: Data for Ideal Lineup Selection

Parameter	**Source of Parameter**
Player statistics (stats are saved separately for each match)	Include Dataset
Team statistics (stats are saved separately for each match)	Include Dataset
Expected goal changes by minutes for both home and away teams	Include Dataset
Home team lineup by minutes	Include Dataset

Table 1 continued:

Away team lineup by minutes	Include Dataset
Environmental conditions (wind, temperature, humidity)	Include Dataset
Referee	Include Dataset
Score of the match by minute	Include Dataset
Both teams' tactical formation by minutes	Include Dataset
Both teams rank	Include Dataset
Position specific key descriptive factors	Taking Expert Opinion via Survey
Team playing characteristic key descriptive Factors	Taking Expert Opinion via Survey
Team playing characteristics	Derived from team statistics
Player positions	Derived from player statistics
Player roles in position/player vectors	Derived from player statistics

To fill the research gap mentioned in Section 1.3, this study aims to answer the following research questions using a new holistic approach to the problem of lineup selection in football.

RQ-1. Which data and methods should be used to solve the lineup selection problem in football?

RQ-2. How can we examine the relationship between player performance and team performance in football using machine learning?

RQ-3. How can player roles and team characteristics be analyzed using data-based approaches?

RQ-4. What are the impacts of expert opinion and the contribution of expert opinion to data-driven sports analytics when determining the defining factors of players and teams?

RQ-5. Can data-based machine learning approaches replace the coach's duties in football?

RQ-6. What are the determining factors in team selection? Can these factors be parameterized?

3.3 Research Approach

The main objective of this study is to develop a framework that integrates player roles, team selection, team tactics, opposing team parameters, and team performance using data-driven methods. Additional study goals are to evaluate the use cases of various machine-learning algorithms and their suitability for such a complex data set, develop robust statistical metrics, and explore the possibilities of data fusion.

This study uses the expected goal parameter to measure team performance. This parameter represents the statistical conversion rate of a player's shots into goals and is calculated using machine learning methods based on historical data and factors, such as the location of the shot, its angle, and the position of opposing players close to the player who took the shot. The expected goal value of each shot takes a value between 0 and 1 [23]. For a football team, a high expected goal parameter is used to measure how close the team is to scoring a goal in a match. Although the expected and actual goal values are not parallel for a single match or short durations, they converge over numerous matches [70]. Conversely, a higher expected goal value for the opposing team indicates a more dominant performance for the opposing team and a higher number of positions the team sees in the goal. Therefore, this study uses the net expected goal parameter, which is the difference between the expected goals for the home team in a match and the expected goals for the opposing team, to measure team performance.

The ideal team selection is determined in two stages. In the first stage, a machine learning model with a target value of the net expected goal is trained using historical match data called the "predictive model." The predictive model used the parameters listed in Section 3.2 Table 1. In the second stage, while keeping the data on the opposing team constant, the targeted team's selection possibilities were tested using the predictive model. The ideal team selection is determined by producing the highest net expected goal data. The model used at this stage is called the "optimization model." The diagrams below show the details of the data flow in the proposed methodology and how the predictive and optimization models are built.

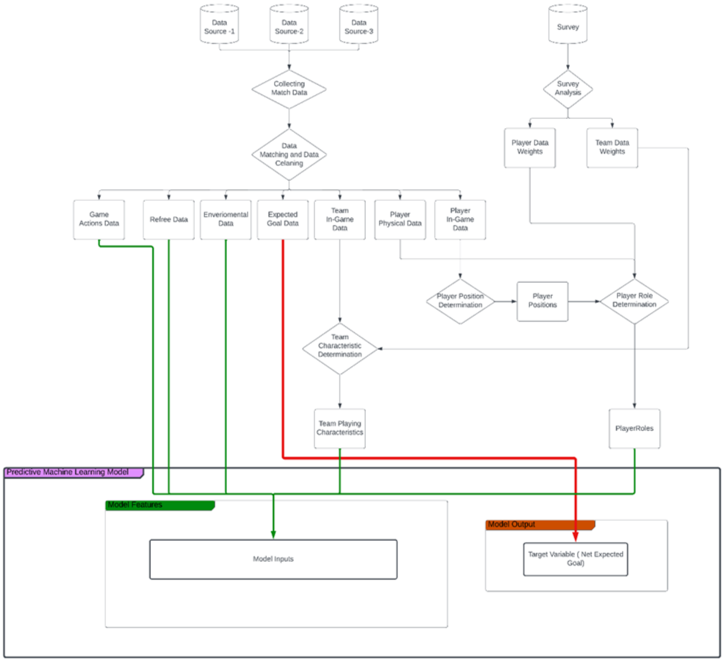


Figure 1: Data flow of predictive model

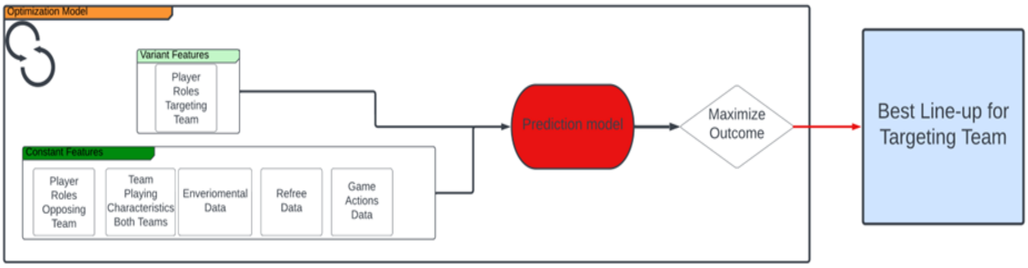


Figure 2: Data flow of optimization model

3.4 Descriptive Parameters Selection Through Survey

After the football community recognized the economic success of sports analytics, investment in this area increased exponentially. These efforts have changed and considerably improved the statistics that characterize the game. While only basic statistics such as goals, assists, and shots were recorded until the early 2000s, the amount of data and statistics in the football environment increased exponentially thanks to the rapid development of data collection tools in the 2000s. With the advent of automated and semi-automated data collection methods and sophisticated statistical parameters, many data providers have obtained over a hundred in-match statistics.

Additionally, team and player statistics can now be kept separately and subdivided based on the opposing team, specific players on the opposing team, or even environmental conditions. Due to this increase in data, a new field of study has emerged on which statistics and parameters best reflect player and team performance. Seeking expert opinion is one of the most popular techniques used in these studies to determine the metrics (KPIs) that best reflect the performance of a player or a team [71],[44],[35]. Therefore, expert opinion was used in this study to identify the leading performance indicators that impact player performance and the game characteristics of the teams. A survey was created to obtain expert opinions, the content of which is discussed in more detail in the following sections.

3.4.1 Survey Methodology

The survey consisted of 3 parts. The objective and intended use of the survey are described in the first part, along with the information and terminology needed to complete the survey. The second part contained three descriptive questions regarding the experts. In the third part, there were 57 questions about the descriptiveness of the in-match statistics for player positions and team characteristics. These questions were divided into three groups: offensive parameters, defensive, and team parameters. While the questions under offensive and defensive parameters were answered separately for the five on-field positions and team characteristics, the questions under team parameters were answered only on team characteristics. The survey consisted of 60 questions, with 285 answers when it was completed.

Table 2: Question Distribution in Each Sub-group of the Survey

Question Groups	**Number of Questions**	**Number of Answers**
Descriptive Questions	3	3 x 1 Answer for Each Question = 3
Offensive Parameters	33	33 x 6 Answers for Each Question = 198
Defensive Parameters	12	12 x 6 Answers for Each Question = 72
Team Characteristics	12	12 x 1 Answer for Each Question = 12
Total	**60**	**285**

Players other than goalkeepers were grouped according to the grouping approach based on spatiotemporal data proposed by Pappalardo et al. [36] and replicated by Behravan et al. [37] to identify player performance KPIs in the survey. Using this method, eight positions are determined during the first stage. Then, considering the symmetrical structure of the football pitch, players were grouped into five positions by combining the symmetrical positions. In the survey, experts were asked how well in-match statistics from Football Reference [72] and InStat [18] data sources describe the requirements of these five positions.

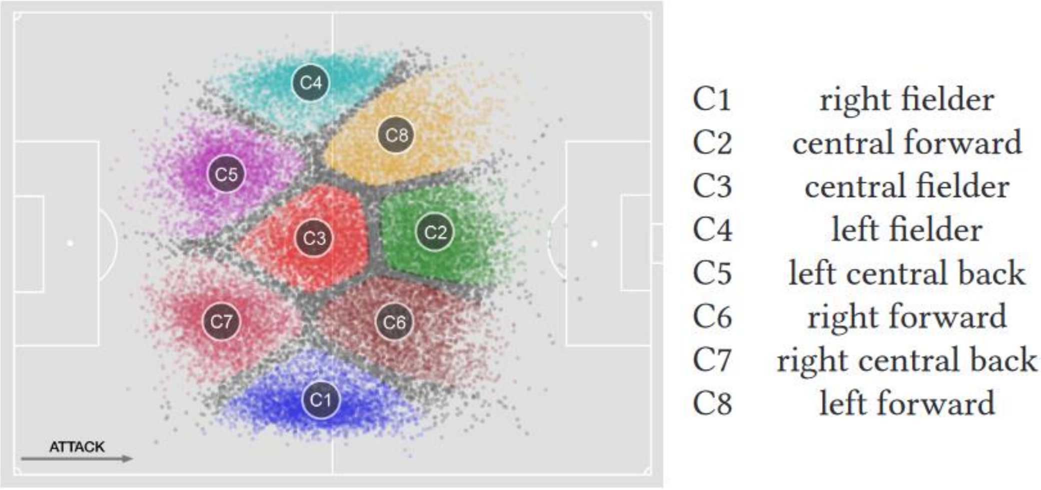


Figure 3: Initial positions [36]

Table 3: Regrouped Positions

Initial Positions [36]	Grouped Positions	Position Definition
C5	P1	Central Defenders
C7		
C1	P2	Back Players
C4		
C3	P3	Central Middlefielders
C6	P4	Side Forwards / Wingers
C8		
C2	P5	Forwards

The survey employed a 10-point Likert scale. A score of 1 on the Likert scale means that the statistic asked does not describe the requirements of the position, whereas a score of 10 means that the statistic asked ultimately represents the requirements of the position.

In the survey, 57 questions were created to determine the game characteristics of the teams by compiling in-match statistics from Football Reference [72] and InStat [18]. In the survey design phase, the study was conducted with a football coach working with UEFA A- license in the Turkish Football Federation and a sport scientist who is an Associate Professor at the Middle East Technical University Department of Physical Education and Sports. The survey was designed following the 5-step design steps suggested by Brace [73] in his book "How to Plan, Structure and Write Survey Material for Effective Market Research.” The steps followed in the questionnaire design phase were as follows.

1. *Identification of Research Question and Objectives:* The survey aimed to determine the game characteristics of the teams and the KPIs of the players according to their positions using in-match data.
2. *Participation Profile and Sampling Planning:* The survey participants were football coaches with a UEFA Pro License, the highest level of license provided by the UEFA. The number of participants was determined as five.
3. *Question Design*: In the direction of a UEFA A-licensed coach and sports scientist, subject matter experts wrote the questions most understandably. The survey questions were examined to ensure that they were clear and consistent with the football terminology used in Turkey, where the survey was conducted. The coaches’ observations determined the correlation between the questions and in-game statistics collected from the data provider.
4. *Scale Design:* A 10-point Likert scale was used in the survey. Definitions for 1 point and 10 points, which are the border points of the scale, were given in the first part of the survey.
5. *Structuring the Survey:* After its initial creation, the survey was delivered to two football coaches with UEFA B licenses at 2-week intervals, and comments regarding its comprehension, design, and completion time were collected. The survey was finalized after this feedback was obtained in two iterations, considering the feedback. The full version of the survey and the results are presented in the Appendix.

3.4.2 Survey Reliability

The reliability of the survey was checked in two stages. First, an internal consistency test was performed on the survey results. Test-retest reliability testing was carried out in the second stage.

To test internal consistency, Cronbach's Alpha reliability coefficient was used. Cronbach's Alpha measures internal consistency and reliability for a group of items or questions in a survey or test. It indicates how closely a tool's components measure the same construct or idea. A higher Cronbach's Alpha means that the items are more internally consistent [74]. This score assesses the scale's consistency within the study. For Cronbach's Alpha value, it is mainly used in studies in psychology and medicine. Bland et al. [75] accepted 0.7 and higher values for psychology as successful and 0.9 and higher values for medicine. Cronbach's Alpha reliability coefficient is calculated as follows:

$$\alpha = \frac{k}{k-1}(1 - \frac{\sum_{i-1}^{k} \sigma_i^2}{\sigma_x^2})$$

where:

- k is the number of questions in the survey
- σ_i^2 is the variance of the responses within each question
- σ_x^2 is the variance of responses from each participant across all questions

Cronbach's Alpha reliability coefficient ranges from 0 to 1, and it has been concluded that when the value gets closer to 1, the survey's internal consistency and reliability rise. The Cronbach's Alpha reliability coefficient was calculated as 0.885 using the survey results as input. This value shows that the internal consistency of the survey is high. According to a study by Takavol et al. [76], Cronbach's Alpha reliability coefficient should be between 0.7 and 0.95 to be acceptable for academic studies.

To measure the test-retest reliability of the survey, the survey was administered to two UEFA B licensed football coaches at two-week intervals. Then, the reliability of the survey was determined by the Pearson correlation coefficient and Intraclass Correlation Coefficient (ICC). Pearson's correlation coefficient (r) measures the linear relationship between two variables. It has a range between -1 and +1, where -1 indicates a perfect linear negative relationship, +1 indicates a perfect linear positive relationship, and 0 shows no linear relationship [77].

Pearson correlation coefficient is calculated as follows:

$$r = \frac{\sum_{i-1}^{n}\left(X_i - \overline{X}\right)(Y_i - \overline{Y})}{\sqrt{\sum_{i-1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i-1}^{n}(Y_i - \overline{Y})^2}}$$

where:

- X_i and Y_i are individual data points (survey question scores for each participant)
- $\overline{X}$ and $\overline{Y}$ are the means of X and Y, respectively.

The results of calculating the individual and combined Pearson correlation coefficient values for the responses of the two participants are shown in Table 4. The results indicate a strong linear correlation between the responses and confirm the understandability and reproducibility of the survey.

Table 4: Test-Retest Pearson Correlation Coefficient Values

Participants	**Pearson Correlation Coefficient**
Participant 1	0.86
Participant 2	0.73
Participants 1 and 2 (mean values)	0.88

The Intraclass Correlation Coefficient (ICC) was also used to measure test-retest reliability. This measure is often used in statistical analysis and research to assess the consistency or repeatability of measurements [78].

ICC is calculated as follows:

$$\text{ICC2} = (\text{MSB} - \text{MSW}) / (\text{MSB} + (k - 1) \times \text{MSW})$$

where:

- MSB is between participants' mean square of the variance of survey rates.
- MSW is between groups' mean square of the variance of survey rates.
- k is the number of participants

The ICC values for each participant's individual and combined test-retest answers are given in Table 5.

Table 5: Test-Retest ICC Values

Participants	Intraclass Correlation Coefficient
Participant 1	0.715
Participant 2	0.723
Participants 1 and 2 (mean values)	0.785

According to Kuo et al. [78], ICC values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability. When the results are analyzed on Kuo's scale, it can be said that the survey's reliability is between moderate and good.

3.4.3 Survey Evaluation

The survey results were calculated by taking the arithmetic average of the answers given for each question. According to the survey results, the importance of statistics related to each position and team characteristic was recorded separately. The following sections will explain how these values are used in the study.

3.5 Determination of the Player Roles

In the literature on football, player roles and positions are often used interchangeably [38],[39],[40],[41],[42]. However, as studies have progressed, it has become increasingly important to distinguish between and define the ideas of player roles and player positions. To achieve this differentiation and to identify player role-specific characteristics, early studies in the literature [44],[45],[47],[48],[28] used labeled data from video games [45],[44], machine learning models [47],[48], and vectorization techniques [41],[49]. This study used unsupervised machine learning models and vectorization methods to determine the player roles.

To better understand the methodology and models used, it is necessary to define the concepts of the player's position on the field and the player role. The concepts used in the study are as follows.

Player Position: The concept of player position is defined in this study as the space in which the player performs on the field and is calculated by grouping the spatiotemporal positions of the players. The labeling is based on the player's average position and is directly related to tactical formation. It determines the spread of the team on the field and the player's duties and responsibilities.

Player Role: In this study, the term player role refers to the duties and abilities of players within their position on the field. Two players playing in the same field area can have different responsibilities, determined by their skills, the position's requirements, and the coach's tactical demands. By definition, player roles can be expressed as a subset of player positions.

3.5.1 Determination of the player position

Player positions from the InStat[18] match the dataset used in this study. In this dataset, players are labeled into eight groups according to their spatiotemporal data within the field as proposed by Pappalorda et al. [36]. Corresponding labels and their meanings are presented in Table 6.

Table 6: InStat[18] Player Position Labels

InStat Position Labels	**Meaning**
CD	Central Defender
RD	Right Defender
LD	Left Defender
DM	Defensive Midfielder
CM	Central Midfielder
LM	Left Midfielder
RM	Right Midfielder
FW	Forward

These labels are grouped independently of the team's tactical formation. However, evaluating player positions independently of the teams' tactical formation can lead to misleading results. Tactical formations assign different tasks and characteristics to players' positions on the field [37]. Therefore, the player position labels provided by InStat [18] were regrouped according to the five positions given in Figure 3 (central defender, back players, central midfielder, side forwards, forwards), considering the tactical formations of the teams. In the InStat dataset, the tactical formations used by the teams for all, or part of the match, are divided into seven groups. These tactical formations were determined by the InStat with the help of automatic and semi-automatic systems according to the position of the players on the field, their distance from each other, and how they would parcel out the space on the field [18]. The specifications of the tactical formation were coupled with data gathered from InStat after consultation with a sports scientist and a professional football coach. These player positions, assigned according to the tactical formation, can be seen in Table 7.

Table 7: Player Position Assignment According to Tactical Formation

Tactical Formation / Assigned Position	Central Defender	Back Players	Central Midfielder	Side Forwards / Wingers	Forwards
4-3-3	CD	RD /LD	CM /DM	LM /RM	F
4-2-3-1	CD	RD /LD	CM /DM	LM /RM	F
3-5-2	CD	RD /LD/ LM /CM	CM /DM	-	F
3-4-3	CD	RD /LD	CM /DM	-	F
4-4-2	CD	RD /LD	CM /DM /	LM /RM	F
4-4-2 Diamond	CD	RD /LD	CM /DM / LM / RM	-	F
4-1-4-1	CD	RD /LD	CM /DM	LM /RM	F

3.5.2 Selection of Attributes that Represent the Player Performance

The players' in-game statistics from InStat and Football Reference were used to transition from player positions to roles. The data is available for every player in every match in the study, and the set of parameters used is the same for all positions. The significance coefficient of each statistic for each position was determined based on a survey administered to selected experts. In addition to in-match statistics, data on players' physical characteristics such as age, height, weight, dominant foot, and the ranking system used by the data provider were also used to determine player roles. The in-match statistics used in the study are presented in Table 8, and the physical data are shown in Table 9.

Table 8: In-Game Statistics for Player Performance

Parameter Type	**Feature Set**
Offensive Goal	Goals, Chances, Chances Conversion Rate (%), Shots, Shots on Target, Shots on Target (%), Expected Goal per Shot
Assist and Chances Create	Assists, Expected Goal Created, Passes, Accurate Passes Percentage (%), Key Passes, Crosses, Accurate Crosses (%), Key Passes Accurate (%), Expected assists (xA), Progressive Passes, Passes into Final Third, Passes into Penalty Area, Long Pass Attempted, Long Pass Completed, Total Distance Covered with Pass, Progressive Distance Covered with Pass, Dead Ball Passes
Dribbling (Ball Carrying)	Fouls Suffered, Attacking Challenges, Attacking Challenges Won Percentage (%), Dribbles, Successful Dribbles (%), Progressive Dribbling, Progressive Dribbling Distance
Negative (Turnover) Parameters	Fouls, Offsides, Lost Balls, Lost Balls in Own Half, Lost After Tackle
Defensive	Ball Recoveries, Ball Recoveries in Opponent's Half, Challenges, Challenges Won Percentage (%), Defensive Challenges, Defensive Challenges Won Percentage (%), Air Challenges, Air Challenges Won (%), Tackles, Tackles Won (%), Ball Interceptions, Free Ball Pick-ups, Blocked Shot, Clearance, Causing Error
Activity	Total Actions, Successful Actions (%), Touches, Touches in Penalty Area, Touches in Forward

Group in-match statistics in Table 8 according to terminologies commonly used in football literature. These terminologies are goal-scoring parameters, assists and chances created parameters, ball-carrying parameters, turnover parameters, defensive parameters, and activity parameters.

Table 9: Physical Attributes

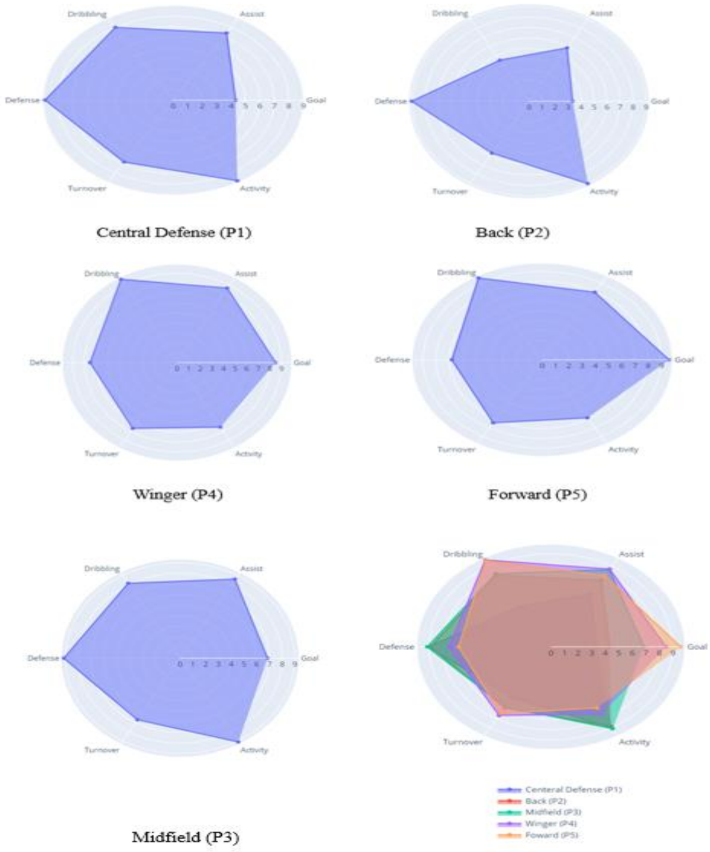
Physical Attributes			
Age	Weight	Height	Dominant Foot

In this study, 56 in-game statistics, four physical parameters, and one ranking value were used as inputs to characterize player types. Since the physical parameters and ranking values did not vary according to the position, the coefficient of significance was accepted as one for all positions. In contrast, the coefficients of the in-match statistics were determined from the survey conducted with the experts.

All in-match statistics were recorded in units per minute, considering the player's minutes on the field. To correctly use the players' in-match statistics, each player's statistics were recorded separately according to the matches they played and their tactical formation in the match. For example, if a player who played in the central midfield area in one match played in the striker area in another, this player's statistics were kept separately in the data collection of the relevant positions since the player was in different position groups in the two matches. In addition, if the player's position group changes due to tactical formation changes during the match, the statistics are recorded separately for each position group per minute format. This way, the player's performance in each position was interpreted with tactical formation information. The statistics' per-minute format enabled a performance measurement commensurate with the players' time on the field.

3.5.3 Survey Results

In the survey, experts were asked to evaluate to what extent the in-match statistics given in Table 8 for each of the five positions fulfill the requirements of the respective position. The questions in the survey, the in-game statistics indicated by the questions, and the survey results are presented in the Appendix. For example, the coefficients determined from the survey for the central defender position are shown in Figure 4.



Central Defense (P1)

Back (P2)

Winger (P4)

Forward (P5)

Midfield (P3)

Figure 4: In-game statistics from survey results

3.5.4 Player Role Determination

Player roles are defined as players' capacity to fulfill the position's duties on the field as described in Section 3.5. Players playing the same position can satisfy the requirements of that position at different levels and in different ways, depending on their skill sets. For example, a player who plays as a forward has many tasks, such as scoring goals, controlling long balls, making runs behind the opposition defense to find open positions, making key passes to players playing behind him, and putting them in scoring positions. However, the player can rarely consistently meet all these requirements with the same quality. Most of the time, the player can accomplish one or more of the position's requirements at a high level while having lower or average performance in other ones. The player's strengths and what they often prefer to do reflect their role and tasks within the position. Therefore, by analyzing in-match statistics, the player's role in the position can be determined by examining which of the requirements of the position the player does well and which he does poorly. However, due to the nature of the dataset used, in-match statistics cannot measure the quality and difficulty of on-field action and instead focus on the frequency of action and percentage of success. Three different approaches have been discussed in the literature to address this issue. The first is to determine player roles by interpreting players' roles in the position with domain knowledge and terminology [44],[45] used player roles used in the game series as input and grouped players according to their strengths and weaknesses based on the game descriptions. On the other hand, another study [28] used domain knowledge to define player roles according to their main requirements and labeled players with in-match statistics that best express the needs of the roles.

The second approach uses in-match statistics, clustering, and classifier algorithms to identify player roles. Kalenderoglu [48] used in-match statistics and hierarchical clustering to determine player roles. Using in-match statistics and a position classifier model, the authors of another study suggested a supervised identification model for player roles [47].

The final approach is to describe in vectorial representation what players can do in their position rather than directly grouping players into a role [41],[49]. Use the Non-Negative Matrix Factorization (NMF) method to examine the strengths and weaknesses of players in a position. Similar to the second approach, the generated vectors were grouped according to their similarity in the space dimension, and the players were matched to the definitions of player roles commonly used in domain terminology [41].

This study used the second and third approaches to determine player roles. However, unlike the existing studies in the literature, in both clustering and vectorization applications, the weight scores obtained from the expert survey were assigned to in-match statistics. Thus, each player's role was determined by considering the position's requirements and the order of importance of these requirements. As clustering methods, hierarchical and k-means clustering algorithms were employed.

Hierarchical clustering is a technique for grouping data points into clusters based on their similarities or differences [79]. Clusters at lower levels are combined to form larger clusters at higher levels, resulting in a hierarchical structure of clusters. Dendrograms [80] visually depict the data's hierarchical structure, facilitate the discovery of groups inherent in the data, and help recognize connections and similarities between pieces of data. Quantitative metrics like the Calinski-Harabasz score [81] or the Silhouette coefficient [82] are used to identify the ideal number of clusters. These metrics evaluate the separation and compactness of clusters, and they can be used to determine the ideal number of clusters to maximize overall clustering quality.

In this study, the NMF method used in studies in the literature [41],[49] was used for the vectorial representation of players. NMF [83] is a type of PCA (Principal Component Analysis) that produces only components with positive values. Factoring a data matrix into two low-dimensional matrices, the primary purpose of NMF is to approximate the matrix and demonstrate the underlying structure of the original data. The terms "base matrix" and "coefficient matrix" frequently describe these two matrices [84]. Once the players were represented as vectors, the Manhattan Distance [85] between the vectors was used to cluster the players.

3.6 Team Characteristics Determination

In football, the outcome of the game and the final score are significantly influenced by strategies and tactics [86]. A strategy is a detailed plan developed and implemented to achieve a goal or a specific objective and is usually formed through specific tactics [87]. The overall strategy used by a team to achieve its offensive and defensive objectives throughout the game is called team characteristics [88]. Both team characteristics and tactics are crucial for the performance of individual players in a match. A strong correlation exists between the teams' playing characteristics and the player's performances. The same player can perform differently depending on the team's style of play and the coach's preferred tactical formation. Therefore, the relationship between team characteristics and player performance is within the scope of this study.

The playing characteristics of teams are related to the players and tactics on the field, but also to the playing characteristics of the opposing team and the difference in quality between the teams [89]. Therefore, in this study, while the teams playing characteristics of the teams are analyzed and categorized, and their relationship with performance is examined, the performances of teams with different tactical formations and game characteristics are included in the model.

3.6.1 Team Performance Representative Features Selection

The approach described in Chapter 3.5 was also used to determine the game characteristics of the teams. In-match statistics were used to differentiate the game characteristics of the teams. This analysis is based on the teams' performance throughout the season, regardless of the match, opponent, or tactical formation. Team performance statistics are presented in the Appendix. In the survey conducted with football coaches, they were asked to what extent these statistics reflect the game characteristics of the teams. The survey results were used as importance weights when grouping teams according to their playing characteristics.

To determine the game characteristics of the teams, 65 in-game statistics were used. All of these statistics were taken from InStat [18]. Of these statistics, 27 are team statistics, and 38 are the sum of the individual statistics shared in Table X10. In addition, the average of InStat's rating values for the teams, updated every match, was also used. The significance coefficient of the rating value was taken as one, and the significance coefficient of the other statistics was determined according to the survey results with football coaches.

Table 10: In-Game Statistics for Team Performance

Parameter Type	**Feature Set**
Offensive Goal	Goals, Chances, Chances Conversion Rate (%), Shots, Shots on Target, Shots on Target (%)
Assist and Chances Create	Passes, Accurate Passes Percentage (%), Key Passes, Crosses, Accurate Crosses (%), Key Passes Accurate (%)
Dribbling (Ball Carrying)	Attacking Challenges, Attacking Challenges Won Percentage (%), Dribbles, Successful Dribbles (%)
Negative (Turnover) Parameters	Fouls, Offsides, Lost Balls, Lost Balls in Own Half
Defensive	Ball Recoveries, Ball Recoveries in Opponent’s Half, Challenges, Challenges Won Percentage (%), Defensive Challenges, Defensive Challenges Won Percentage (%), Air Challenges, Air Challenges Won (%), Tackles, Tackles Won (%), Ball Interceptions, Free Ball Pick-ups,
Activity	Entrances to the Opposition Half, Entrances to the Final Third, Entrance to the Penalty Box, Total Actions, Successful Actions
Dead Ball Activities	Corner, Set Pieces Attacks, Free-kick Attacks, Corner Attacks, Throw-in Attacks, Penalties
Offensive Expected Goal	xG (Expected Goals), Net xG (xG - Opponent's xG), xG Conversion, xG per Shot
Defensive Expected Goal	Opponent's xG, Opponent's xG per shot
Pressing	Team Pressing, Pressing Efficiency, %, Ball Possession, %, Team Pressing Successful, Opponent's Passes per Defensive Action, High Pressing, High Pressing, Low Pressing, Low Pressing, %
Possession	Building-ups, Building-ups without Pressing, Ball Possession (sec), Ball Possessions (quantity), Positional Attacks, Positional Attacks with Shots
Counterattack	Counterattacks, Counterattacks with a Shot

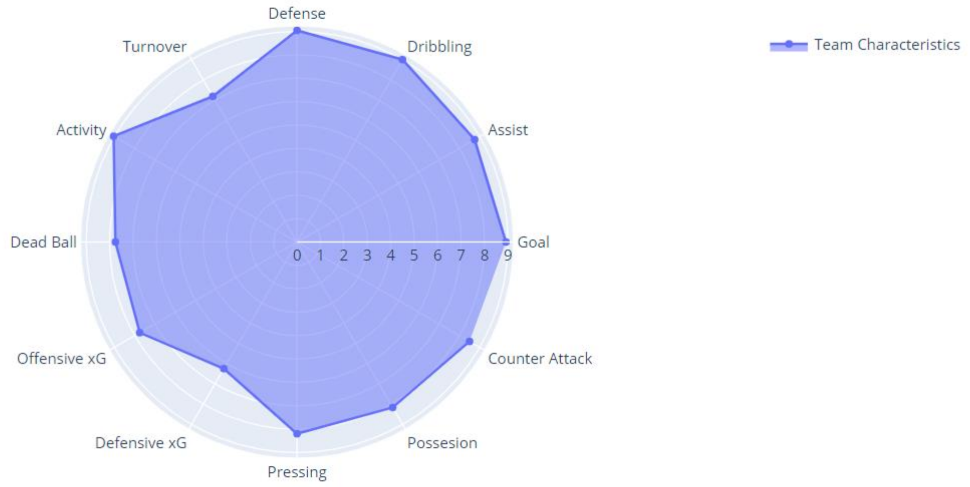


Figure 5: Survey ratings of team characteristics

3.6.2 Team Characteristics Clustering

In the literature, clustering [45], PCA [88],[89],[90], statistical analysis [91], and literature review [92] methods have been used to determine the game characteristics of teams. In this study, clustering and vectorization methods described in Section 3.5.4 were used by considering the survey results conducted with football coaches as importance weights.

3.7 Expected Goal

The expected goal (xG) is used to calculate the likelihood of scoring a goal from a particular goal attempt. It returns a number between 0 and 1, indicating the probability that a shot will result in a goal [23]. The expected goal parameter is shot-based. Thus, a position that does not result in a shot, regardless of its potential, does not change the expected goal parameter. Every shot has an expected goal value calculated independently of teams, leagues, matches, and other external factors. This value is determined by the shot's position, its angle to the goal, and the distance between the shot and the opposing defense [30].

Football literature has numerous expected goal calculation methods [26],[30],[31]. Furthermore, each data vendor contributes a different perspective to this metric by adopting its own calculating method. In this study, we use InStat [18] and Football Reference [72] data, which record the xG value of all shot movements during the match with time information. Team performances are measured as the difference in the xG values of the shots taken by the teams during the observed time interval. This measurement is kept separately for each goal variation, team tactics, and line-up characteristics of the teams, defined as inputs to the prediction model given in Figure 1.

Herbinet [25] used machine learning, deep learning, and statistical methods to test how effective the expected goal parameter is in measuring team performance using the expected goal, score, match result, and match ratio and found that models that use the expected goal parameter outperform other models. As a result of these considerations, the expected goal parameter is employed as the performance evaluation criterion for the teams in this study.

3.8 Performance Evaluation Methods

There are two fundamental models in the proposed methodology described in Section 3.3. The first one is a predictive model that uses in-match variable parameters such as score, team tactics, and team formations and parameters that remain constant throughout the match, such as environmental conditions, referee, match location, and team rankings as inputs, with the target value being the net expected goal value. The objective of this model is to accurately predict the teams' performance using in-match variables and constants. The details of the predictive model are explained in Section 3.8.2.

3.8.1 Parameters and Feature Engineering

Football literature has many studies on match results and score prediction. Bunker et al. [63] proposed a general framework for performance prediction models using machine learning techniques. The first steps in this framework are domain understanding, data understanding, and feature selection. In parallel, the feature sets that significantly influence team performance have been investigated in most studies attempting to predict or explain team performance. In some of the studies [25],[63],[93], in-match statistics were used, while in others [59],[60],[63],[64], the effects of out-of-match conditions were also tested.

This study collects the factors affecting the performance according to the feature sets used in the literature for match results and score predictions. In addition to the player roles and teams’ game characteristics features described in Chapters 3.5 and 3.6, tactical formation, score, opposing team parameters, referee, and environmental conditions are also included. Of these parameters, the score is used as a direct input to the model, while the tactical formation and referee parameters are encoded. Opponent parameters are integrated with the opponent team's tactical formation, lineup preferences, and rank features. The environmental conditions were transformed using a clustering algorithm into an input for modeling using temperature, wind speed, pressure, and humidity attributes. The factors affecting the performance are grouped under match-related and external features, as suggested in the framework [63]. The features used in the predictive model by this grouping are given in Table 11.

Table 11: Grouping of Features

Match-Related Features	**External Features**
Line-up players’ roles	Team ranks
Team tactical formation	Team playing characteristics
Score	Referee
	Environmental conditions

During the match, the attributes under match-related features may change. When these features change, a breakdown in the dataset is generated, and the match is examined based on these separated segments. External features are match-specific parameters that do not change during the match.

3.8.1.1 Effect of the Team Tactical Formation

Playing tactics and style are the distinctive behaviors of a team during a competition. The complexity of tactical decisions, such as preferred game formations or tactics, has grown over time, and the public continuously evaluates coaches' tactical expertise. For measurements of the variables that determine team tactical formation to be accurate, these measurements need to be repeated frequently under specific scenarios. Player and ball movement and player interaction are essential variables, usually composed of speed, time, and space components [92].

In today's elite football, tactics are crucial to success [94]. There are strong correlations between tactical formations and team performances, which can be further strengthened by combining them with player preferences and playing styles [66]. In this context, tactical formations were used in the predictive model. The data were collected from InStat [18] for each match, considering tactical changes within the match. The time information of the tactical changes made during the match was recorded, and the match was divided into segments according to the tactical formation preferences of the two teams. The dataset has seven tactical formations: 4-3-3, 4-2-3-1, 3-5-2, 3-4-3, 4-4-2, 4-4-2 Diamond and 4-1-4-1. These formations are the most frequently used tactical formations in soccer literature and terminology. All data about team tactical formation is encoded and directly fed into the predictive model.

3.8.1.2 Effect of the Score

Many technical, tactical, and environmental factors that influence team performance and match results have been identified in the literature by evaluating a football match as a whole. However, when a match is examined in parts rather than as a whole, it becomes evident that the performance does not progress linearly throughout the match, with each team performing better or worse in various parts. This is because football is a team sport.

When the factors that determine the variation of performance within a match are examined, the psychological effects of the score are found to have a substantial link with performance [95]. For example, scoring the first goal in a match has been correlated significantly with victory. Thus, there may be times when the correlation between the match score and team performance takes precedence over other factors. For instance, in a match where one of the teams enters the last 15 minutes leading by three or more goals, it may be challenging to explain the teams' performances in this part of the match regarding tactical formations, team preferences, or team quality. Motivation, psychological ability, and mental toughness are critical psychological determinants of football [96]. In a study examining the impact of score difference on team performance in professional football matches, it was found that teams that lost by a large margin (a score difference of three goals or more) had lower levels of performance determinants such as ball possession, successful passes, and shots on goal [97]. This indicates that teams' performance can be negatively impacted when they are significantly behind in terms of score.

In light of findings in the literature, the current match score was added to the data set used by the predictive model. Match scores are not normalized and are used as positive integers when in favor of the home team and negative integers when in favor of the away team. Each match in the dataset is segmented according to score, player, and tactical changes. As a result, the prediction model is intended to provide more realistic and accurate predictions by considering pre-match factors and score-related variables during the match.

3.8.1.3 Effect of the Opposing Parameters

The saying "Football is a game played with the opponent" is common in football terminology. This statement, which football commentators frequently repeat, can also be a challenge for sports scientists who study team performance. A sports scientist who attempts to examine a team's season based on its performance would be making an incomplete measurement if he just looked at the opposing team's performance. Only when the performance of each team is analyzed along with the performance of the opposing team a more precise measurement can be obtained [98]. Unfortunately, there aren't many studies that combine performance analysis with opponent parameters. Instead of directly comparing the performance of teams with that of opponents, such studies categorized the opponents’ performance using characteristics such as team ranking, standing, and form. For instance, teams can be classified into tiers or groups based on their placement in the respective league table (i.e., top-tier, middle-tier, bottom-tier teams, etc.). This classification enables a team’s performance to be assessed against a variety of levels of competition.

Studies examining the direct or indirect effects of various factors on the match performance of the opposing team have examined physical and technical performance criteria [99],[100] and psychological factors through the number of fans and club reputation [101].

In the predictive model, the opposing team parameters are the opposing team’s tactical formation, player roles determined concerning the formation, player characteristics, and rank. Incorporating such extensive information on the opposing team in the predictive model is the most significant difference between the proposed model and other models in the literature.

3.8.1.4 Effect of the Environmental Conditions and Referee

Environmental factors have been shown to substantially impact game performance in football. High temperatures have been linked to decreased high intensity running, increased fatigue, and dehydration toward the end of a match [102]. Like severe heat, extreme cold can also reduce muscle function and increase the likelihood of injury [103]. Humidity can also impact physical performance since it influences how heat is dissipated and how hard an effort is felt [102]. Altitude can lower aerobic capacity and increase fatigue in athletes who have not yet become accustomed to these conditions [102]. Wind affects the pass quality and the shot’s precision [104].

The proposed predictive model considers the effects of environmental conditions. Using the website wunderground.com [105], the environmental conditions at the time and location of each match in the dataset were retrieved from the nearest measurement station to the match location using web scraping. For each match, a weather dataset was created by collecting temperature, humidity, wind speed, and pressure values at the time of the match. Environmental condition data were categorized instead of used directly in the predictive model. First, the ideal number of clusters was determined using the Elbow Method and Silhouette Score [82]. In the Elbow method [106], the within-cluster sum of squares is plotted against the number of clusters, and the objective is to find the plot's "elbow" or point of turning point. The level that each data point fits into its assigned cluster is measured through Silhouette Analysis. The number of clusters with the highest Silhouette Score (108) is determined as the ideal cluster number. Graphs of the analysis can be seen in Figure 6. Then, grouping was performed using the K-Means Clustering Algorithm. K-means Clustering [106],[107] is an algorithm that efficiently separates data points into a certain number of clusters with similar characteristics by considering the attributes of the data.

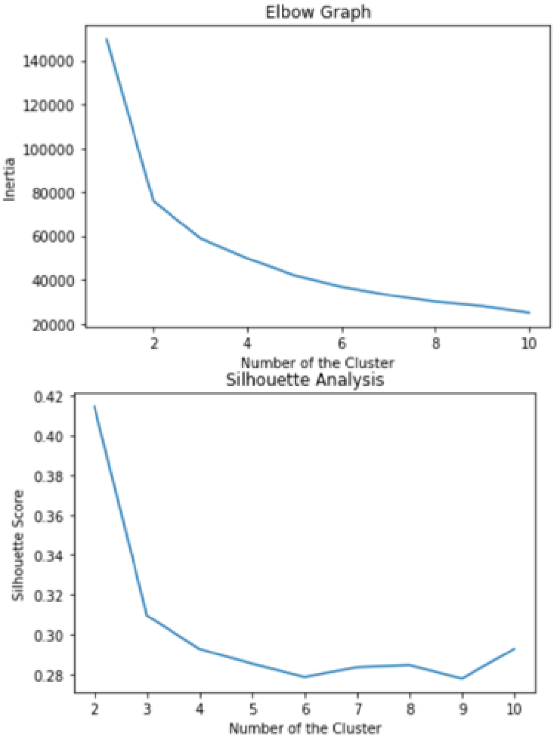


Figure 6: Elbow Graph and Silhouette Analysis Results

The Elbow Graph and the Silhouette Analysis both show that two clusters are the optimal number of clusters. Three clusters were further investigated in the predictive model to make it more sensitive to the effects of environmental conditions, assuming that environmental factors significantly impact match performance. The analysis of the grouping for two and three clusters is given below.

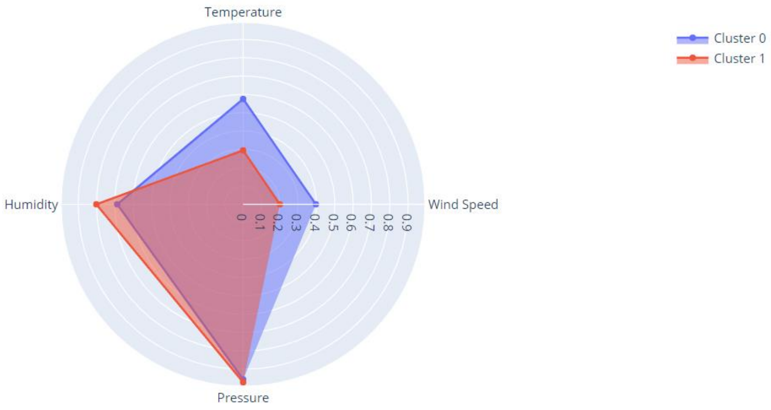


Figure 7: Two cluster numbers for environmental conditions

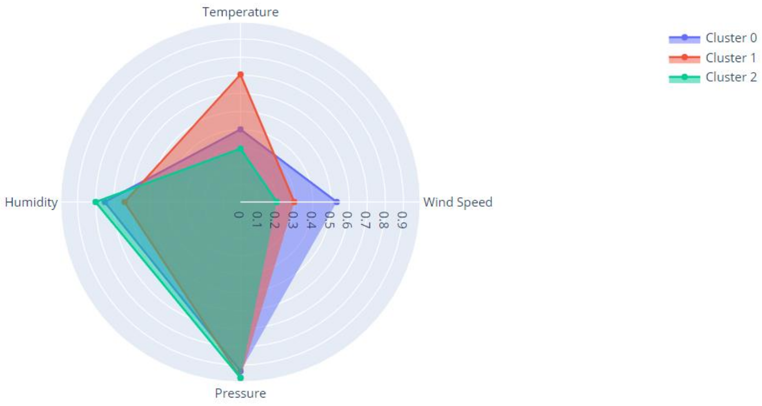


Figure 8: Three cluster numbers for environmental conditions

To distinguish between environmental conditions, the temperature and wind speed data shown in Figure 7 and Figure 8 provide more variability than the other environmental condition data. In the predictive model, a 3-cluster environment condition feature is used as it allows for a more precise grouping of environmental conditions.

The referee's influence on football is a commonly discussed topic on social media and in the conventional media. Football referees are responsible for ensuring the game is played fairly and from start to finish. However, referee errors greatly impact the outcome in low-scoring games like football. In the studies conducted on this subject, referees' mistakes and tendencies have been examined by considering variables related to social pressure [108], the crowdedness of the tribunes [109],[110], and referees' skills [111]. According to the findings of these studies, social and environmental pressures cause referees to make more errors in favor of the home team. This does not apply to all of them but is related to the referees' styles. As a result, the effect of referees must be analyzed in a performance prediction model. Therefore, the referees in the dataset are encoded and used directly in the predictive model.

3.8.2 Modeling

Despite the size of the football economy and the prevalence of football worldwide, research on performance measurement methodologies is relatively scarce [59]. Most research employs classification models since they are easier and more successful at forecasting accuracy and betting systems. Classification models are trained by labeling matches according to their results, such as wins, losses, or draws. Although these models are more accurate than regression models and are suitable for betting structures, they are insufficient for performance evaluation studies, as mentioned in Section 3.1 [59]. Therefore, classification models were not included in the scope of this study. When examining the models that classify match results and performance prediction as a regression problem, numerous approaches are identified in the literature.

Peters and Pacheco [59] used in-match statistics and player and team breakdowns as in the proposed approach for predicting match outcomes. They forecast the number of goals scored by the home and away teams independently using various machine learning models, and the match result was calculated as the difference between these two forecasts. They validated their proposed model using Kendall Rank Correlation and the teams' end-of-season standings. They converted the proposed methodology into a betting system and determined the best-performing model and dataset according to profitability. The in-match statistics of the line-up players determine the highest correlation to goals scored. At the same time, the best betting performance is obtained by applying the K-Nearest Neighbor algorithm and in-match statistics. Although their study and this study employ in-match statistics to gauge performance, their model does not consider player roles, game characteristics of the teams, the opposing team, score, environmental conditions, or the referee. As a result, while this study cannot be utilized as a benchmark model, it follows the same methodology as the predictive model in demonstrating the effect of in-match statistics on performance.

Another study predicts performance based on differences in goals in matches [64]. Physical attributes, ball handling skills, passing skills, shooting skills, defensive skills, and mental attribute scores of lineup players from the FIFA game series were tested using multiple machine learning models. The home and away teams’ lineup attributes are utilized to predict the goal difference in the matches. The goal difference reference values were determined, and the match result was forecasted if the goal difference was estimated above the reference value. Although the technique of this study is similar to that of the predictive model, it differs from it in terms of the use of artificial data, the uncertainty in calculating the reference goal difference values, and the direct use of the match score rather than the expected goal parameter. Even though this study cannot be used as a benchmark due to these variations, the mean square error value of 1.87 generated by the Random Forrest model, the most successful model in the study, in goal difference prediction is compared with the predictive model's output.

Since no other model in the literature is comparable to the predictive model and works with similar data sets, there is no benchmark model. The model proposed by Herbinet [25] was determined for the comparison of the results obtained in this study since it is most similar to the proposed model. The performance values of the teams are derived as ELO ratings in this study using the expected goal parameter. In addition to the expected goal value based on shots, the likelihood of scoring goals from positions that do not result in shots, named Match xG, is estimated using the spatiotemporal data. The model was evaluated using root mean square error (RMSE) and mean absolute error (MAE). With an MAE of 0.861 and an RMSE of 1.153, the neural network model produces the most accurate results, whereas the final model predicts the match score with an average error of 0.861 goals per team score.

The Dixon and Coles Model [112] has been utilized in most football-related studies to estimate match result probability and forecast the number of goals each team would score [113]. This model [112] employs the Poisson distribution and examines each team's goal-scoring and goal-conceding statistics. Herbinet [25] used expected goal data to train the Dixon and Coles model [112], which he used to compare the proposed approach in his study. As a result, although the Dixon and Coles model predicted the match score with an RMSE of 1.138, the MAE value was lower than that of Herbinet's model [25].

The proposed predictive model is a performance prediction model that tries to predict the match performance of teams in terms of net expected goals per minute based on player roles of the lineup, tactical formations of teams, game characteristics of teams, and environmental factors. It is used to select the team line-up that produces the highest net expected goal per minute for the proposed best lineup framework. Because of this scale and measurement difference there is no benchmark model determined to compare results of the predictive model.

3.8.2.1 Predictive Model Features

The features of the predictive model are broken down according to the tactical formation of both teams in the match, player preferences, and changes in the match score. Table 12 shows how these features are obtained.

Table 12: Predictive Model Features

Feature Name (In model name)	**Feature Explanation**	**Data Type**
Referee (**Referee**)	Chapter 3.8.1.4: Main referee of the game. They are used in categorical values as one hot-encoded version.	Categorical
Home Team Playing Characteristics (**h_type**)	Chapter 3.6: Teams are clustered according to team stats and expert weighting using hierarchical clustering and used in categorical value as one hot encoded version.	Categorical
Away Team Playing Characteristics (**a_type**)	Chapter 3.6: Teams are clustered according to team stats and expert weighting using hierarchical clustering and used in categorical value as one hot encoded version.	Categorical
Score (**Score**)	Chapter 3.8.1.2: Used as positive and negative integers showing score difference. Positive integers refer to the score advantageous in the home team, while negative integers refer to the score advantage in the away team.	Continuous (integer)
Environmental Conditions (**Env**)	Chapter 3.8.1.4: Environmental conditions at match time are clustered using k-means clustering. Cluster labels are used in categorical values as one hot-encoded version.	Categorical
Home Team Tactical Formation (**h_tac**)	Chapter 3.8.1.1: Seven tactical formations are labeled and used in categorical value as one hot encoded version.	Categorical

Table 12 continued:

Away Team Tactical Formation (**a_tac**)	Chapter 3.8.1.1: Seven tactical formations are labeled and used in categorical value as one hot encoded version.	Categorical
Net Rate (**Net rate**)	Chapter 3.8: Difference in team ratings. The net rate is calculated by subtracting the away team's rating from the home teams. A positive net rate indicates that the home team has a stronger rating and is stronger, while a negative net rate suggests that the away team is stronger. This attribute is normalized before use.	Continuous (float)
Home team player roles (**h0, h1,.., h9 or home embedding_0, home embedding_1…. or NMF_H)**	Chapter 3.5.4: The on-field roles of the players were determined using hierarchical clustering, embedding, or NMF. The determined roles are presented in the predictive model according to the methodology used. While the player roles determined by hierarchical clustering are added to the predictive model using one-hot encoding as categorical data, with NMF and embedding methods, player roles are vectorized and used as vector values in the predictive model.	Categorical / Vectorial
Away team player roles (**a0, a1,…, a9 or away embedding_0, away embedding_1…. or NMF_A)**	Chapter 3.5.4: The on-field roles of the players were determined using hierarchical clustering, embedding, or NMF. The determined roles are presented in the predictive model according to the methodology used. While the player roles determined by hierarchical clustering are added to the predictive model using one-hot encoding as categorical data, with NMF and embedding methods, player roles are vectorized and used as vector values in the predictive model.	Categorical / Vectorial

Table 12 continued:

Net Expected Goal Per Minute (**Net_xg_per-min**)	Chapter 3.7 Net expected goals per minute is the target feature for the predictive model. It is calculated by subtracting the expected goal value created by the home team from the expected goal value created by the away team in the selected section of the match and dividing it by the match duration into minutes in the selected match section. Positive values of net expected goals per minute indicate a better home team performance, while negative values indicate a better away team performance. When the value moves away from 0 to a positive or negative direction, it means that the performance difference between the two teams increases.	Continuous (float)

3.8.2.2 Predictive Model Preprocessing

The data was preprocessed before training the predictive model using the below-mentioned techniques.

1. *Determining player roles by filtering playing time:* As explained in Section 3.5.4, spatiotemporal data from the players on the field, tactical formation data from the teams, and position labels from the data provider are combined to determine players' roles in a football match. However, depending on the conditions of the match, players may be assigned to different positions and roles in their initial tactical assignments. This variation may be due to in-match variables such as score, changes in the tactical formation of the teams, red cards, or the personal preferences of the coach. For this reason, in the dataset used for the predictive model, instead of grouping the players under a single position, their positions were determined according to the grouping in Table 7 with the tactical formation they were on the field in each match they played.

   Players' in-match statistics were recorded separately for each position. However, the playing time of some players in some positions covers a very small time interval compared to all the data for that position. This can lead to misleading results in determining player roles using the hierarchical clustering method. To overcome this problem, the performance data of the players in the five positions were filtered by one of three different filtering methods before separating the players according to their roles in their position. The predictive model is tested separately for these three different filtering methods.

As a first filtering method, players who played only one match in the relevant position during the season were removed from the relevant position's data. The performances of these players were not included in the algorithms used to determine their roles in the respective position. These players are marked as 0 in the dataset used to train the predictive model. This marking means that the player is playing in a non-ideal position. Although this filtering method solves the problems of determining player roles, the data loss was high compared to other filtering methods. As the second filtering method, players who played more than 15 minutes per match in the relevant position were included in the role determination algorithms. Although this filtering method provides less data loss than the first method, it causes the players who constantly substitute in the final part of the match to be removed from the dataset. The last filtering method considered the total playing time in the position. Players who played at least 30 minutes at the relevant position during the season were included in the role-determination algorithms within the position. This method reduced the loss of data compared to the first method. Also, it allowed the model to have players who played regularly in relevant positions for short periods. The predictive model was tested for all three datasets. The filtering methods' effects on the dataset's size and the predictive model's success are explained in Section 4.3.

2. *Selecting player role representative method:* In the predictive model, player roles are represented by three different approaches. The first is using the categorical values the hierarchical clustering method determines. The second one is to vectorize the categorical values determined by the hierarchical clustering method using an embedding structure. The last one is to transform the player performances into a vector using the weights of the relevant position. The predictive model is tested separately for these three approaches. The impact of these approaches on model performance is shown in Section 4.3.
3. *Converting team ratings to a single rate*: Team rating values, calculated by considering the teams' current form, the players' total economic value, and their ranking in the league, were obtained separately for each match from the data provider. In the predictive model, a single rating value was obtained and used by subtracting the home team's rating value from the away teams.
4. *Encoding categorical features: Categorical* features in the predictive model were encoded using the One Hot Encoder method. It is a method used in machine learning and data preprocessing to represent binary vectors of categorical data and involves transforming categorical variables into binary vectors.

5. *Normalizing continuous features:* Normalization is a data preprocessing technique used to put data into a consistent format, usually on a scale ranging from 0 to 1. This is because machine learning algorithms can be sensitive to the scale of input features, and normalization is necessary to ensure that each feature contributes equally to the learning process. Therefore, all continuous features in the predictive model are normalized before model training.
6. *Calculating the target feature:* Net expected goals per minute are calculated by subtracting the expected goal value created by the home team from the expected goal value created by the away team in the selected interval of the match and dividing by the duration (in minutes) of the relevant part of the match.
7. *Removing the red card games:* A red card is a typical occurrence during a football match. A player who receives a red card is sent off, and his team has to play the remainder of the match with one man down. Because this alters the balance of power within the match and the primary goal of the predictive model is the line-up decision, the parts of the matches following the red card are excluded from the dataset that will be used to train the model.
8. *Removing outliers:* The data is segmented for each score, tactical formation, and substitution change in the match. Therefore, the data for the predictive model consists of parts that reflect specific sections of the match. Some of these parts represent very small time intervals compared to the total duration of the match. Expected goal changes occurring at these small time intervals can lead to outliers that mislead the data. For example, after a substitution in the 75th minute, an attack with an xG value of 0.8 took place and resulted in a goal. Immediately after the goal, a substitution in the 76th minute created a new control point in the data. This example shows that the expected net goals per minute will be calculated as 0.8. However, this value is significantly higher than the expected net goals per minute of the overall data as it analyzes only 1 minute of the match. To avoid this, only match segments longer than 15 minutes are included in the predictive model. The effects of this pre-processing on the data are analyzed in Section 4.4.

3.8.2.3 Models

The predictive model has been tested using a variety of machine learning and deep learning models, which are listed in Table 13. It is a regression problem because the predictive model's target is the expected net goals per minute. Because the dataset is predominantly categorical, models such as CatBoost Regressor and LGBM Regressor achieved better results. Five machine learning and deep learning models outperformed the other models. These are Multilayer Perceptron (MLP), Attention with Embedding, Gradient Boosting Regressor (GBR), LGBM Regressor, and CatBoost Regressor (CBR). In Section 4, the results obtained through these models are presented.

Table 13: Machine Learning Models

Multi-Layer Perceptron (MLP)	Attention with Embedding Layer
Gradient Boosting Regressor (GBR)	LGBM Regressor
Cat Boost Regressor	Scholastic Gradient Decent Regressor (SGD Regressor)
Stacking Regressor	Logistic Regressor
Elastic-Net	Support Vector Machines (SVM)
Gradient Tree Boosting	Decision Tree Regressor

Multi-Layer Perceptron (MLP): Multilayer perceptron (MLP) is a form of neural network that has input, hidden, and output layers. Figure 9 illustrates the structure of a multi-layer perceptron, which consists of interconnected nodes that represent the non-linear transformation between an input and an output [114]. The main purpose of nodes is to structure the learning process according to model input by assigning importance weights to the data to be learned. They are used in a variety of contexts, including classification, regression, and natural language processing.

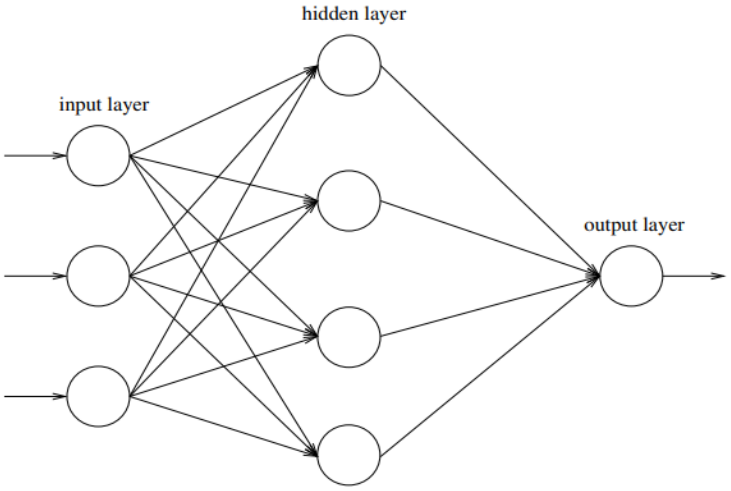


Figure 9: Multilayer Perceptron [115]

Attention: Attention is a form of neural network model. The attention model is a mechanism used in neural machine translation (NMT) and other tasks to selectively focus on relevant parts of the input during the translation or processing process. It has been shown to improve the performance of NMT systems by allowing the model to attend to different parts of the source sentence at different time steps [116]. It consists of a query, key-value pairs representing the input, and a scoring function. The scoring algorithm determines how similar the query and keys are, producing attention scores. The SoftMax operation is used to convert these scores into attention weights. A context vector is created by combining the values related to the keys depending on these weights. The most important information is combined with the initial query and subsequently processed in the model's later layers to create accurate and context-aware results.

For problems involving sequence-to-sequence mapping, such as machine translation, text summarization, and picture captioning, attention models are a critical achievement in the fields of machine learning and artificial intelligence. The idea of attention seeks to imitate how human perception functions, which involves concentrating just on specific inputs while information is being processed.

A model's ability to efficiently extract pertinent information from a given input which could be a list of words, an image, or any other type of structured data is fundamentally improved by an attention mechanism. The model's capacity to handle long-range relationships and various significance levels within the input is significantly enhanced by this attention-driven method.

Gradient Boosting Machines (GBM) Regressor: A potent machine learning method that is a member of the gradient boosting method family is the Gradient Boosting Regressor (GBR). It is frequently utilized and has achieved great success in a number of real-world applications [117]. In addition to being taught with regard to various loss functions, GBR is very adaptable and may be adjusted to the particular needs of the application [117]. By fusing the strengths of several weak learners to produce a strong and precise prediction model, it excels in predictive tasks, especially regression issues. The goal of GBM is to reduce prediction errors by continually enhancing the shortcomings of earlier models.

GBM frequently outperforms individual models in its ability to capture complicated relationships in data. Categorical and numerical features are two data kinds that GBM can handle. GBM offers perceptions into the significance of features, assisting in feature understanding and selection. If not adequately regulated, GBM can overfit. Careful tweaking is required for variables like tree depth and learning rate.

LGBM Regressor: A machine learning technique that is a member of the gradient boosting family of models is the LGBM Regressor, often referred to as the Light Gradient Boosting Machine Regressor. The LGBM Regressor is built on the gradient boosting framework, which entails training a group of weak prediction models—typically decision trees—to iteratively fix the mistakes caused by the group's earlier models [118]. Due to its effectiveness and superior performance across a range of applications, it has grown in popularity. Credit risk analysis, energy forecasting, water quality prediction, and sepsis-associated acute brain damage prediction are just a few of the areas where the LGBM Regressor algorithm has been applied.

Similar to other gradient boosting techniques, LightGBM minimizes the loss function using a gradient descent algorithm to optimize models. However, it differs in that it speeds up training by adopting a method known as "Gradient-Based One-Side Sampling" to choose the most useful data points for creating decision trees. Histograms are used by LightGBM to group feature values into discrete values. Finding the ideal split spots during tree construction is sped significantly as a result. LightGBM employs a leaf-wise strategy as opposed to conventional depth-wise tree growth. This indicates that the leaf nodes with the greatest loss reduction are expanded as the tree grows. This results in trees that are more complicated and may capture complex data patterns.

Catboost Regressor: A gradient boosting technique called the CatBoost Regressor was created expressly to handle category information in machine learning problems. It is an open-source library that has been created to perform better on well-known datasets than current gradient boosting implementations in terms of quality [119].

CatBoost proposes "ordered boosting," where the algorithm considers all potential splits for each feature and chooses the optimum split based on the feature values' ordered structure. Thus, fewer trees are required, and information is captured more effectively. CatBoost supports categorical characteristics natively. It transforms categorical variables into numerical values using a cutting-edge method known as "permutation-driven computation," enabling them to be employed right in the algorithm. CatBoost optimizes models by iteratively minimizing the loss function using gradient descent, much as other gradient boosting methods. Its treatment of categorical data makes a difference in how well it can learn from categorical features.

3.8.2.4 Hyperparameter Optimization

Hyperparameter optimization refers to finding the ideal set of hyperparameters for a machine learning model to achieve optimal performance. Hyperparameters are model parameters the user specifies before training rather than learning from data during training. They significantly impact the model's architecture, behavior, and generalization capabilities. Optimizing these parameters improves the model's performance and outcomes. The approach typically entails exploring various hyperparameter combinations and evaluating the model's performance on a validation set.

Grid search hyperparameter optimization is a popular method for determining a model's best set of hyperparameters [120]. A grid of hyperparameter values is constructed in this method, and all possible combinations are carefully investigated to discover which one delivers the best model performance. The grid search method was utilized to obtain the predictive model's parameter values.

3.8.3 Metrics and Evaluation

The predictive model's success was measured using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The dataset was partitioned into 80% of training and 20 % of test sets after the last ten games of seasons were separated for validation. The training set was used to train the predictive model. The model's performance was evaluated using the test set and the MAE and RMSE measures.

Mean Absolute Error is defined as the absolute difference between the ground truth value and the predicted value and is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i |$$

where y_i indicates ground truth values, and $\hat{y}_i$ indicates predicted values.

Root Mean Squared Error is defined as the square root of the average of the squared differences between the ground truth values and the predicted values and is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i indicates ground truth values, and $\hat{y}_i$ indicates predicted values.

3.9 Best Lineup Selection

The optimization model, a variation of the predictive model, determines the best lineup selection. It determines the best team by evaluating the payoff of all available alternatives using the predictive model while keeping the opponent's tactical formation and player preferences constant concerning the selected team's tactical formation. For varied team selection probabilities, the optimization model uses the predictive model to calculate the performance of the selected team against the opponent's tactical formation and player preferences. It presents the predictive model’s inputs that predict the highest performance return as the best line-up choice.

A review of the best lineup research in the literature reveals that some of the studies skip validation [121],[122]. The complexity of the problem, as well as the presence of difficult-to-measure features such as the opponent team and the current form of the team and the players, make comparing the performance of the choices and suggestions provided to other potential solutions challenging. In other studies in the literature [5],[65],[66],[35], the proposed best lineup was compared with the coaches' player preferences in actual cases, and the models were validated according to the similarity values obtained. In Cortez's study [5], in addition to the similarity between the coach's preference and the proposed best lineup, the validation step is presented in a more comprehensive framework by measuring team performance when similar choices are made to the proposed best lineup. However, Cortez [5] performed the validation step from a single-team perspective for only three positions and interpreted the validation for each position independently of other choices. Therefore, no model or metric in the literature can be directly compared with the optimization model.

3.9.1 Parameters

The optimization model has two parts, the selected team's player list, and the predictive model. In the optimization model, except for the predictive model's specifications of the selected team's player roles, the other specifications are fixed. In the first step, the optimization model determines the number of player positions required by the tactical formation of the selected team. Then, it predicts the net expected goals per minute parameter with the predictive model by trying all the players who are available for team line-up selection. As a result of this estimation, it determines the player set that provides the highest net expected goal per minute parameter as the best lineup.

3.9.2 Validation

An approach similar to Cortez's [5] approach was developed to measure and validate the performance of the optimization model. The last ten matches of the season were excluded from the predictive model and utilized as a validation set. Using the optimization model, the best lineup output was generated for both teams in these ten matches, for 20 teams. These results were compared to the actual match divisions, and their similarity to the coaches' preferences was examined. Aside from the similarity rate, how the performance return is increased is also investigated. For the optimization model to be validated, the similarity between the optimal lineup provided by the model and the coaches' preferences is intended to be directly associated with the performance output.

CHAPTER 4

EXPERIMENTAL SETUP

This section details the processes followed and the results obtained when testing the methodology proposed in this study using real-world data gathered from multiple data sources. The scope of the experimental setup is set to one season, and the English Premier League season 2021-2022 is employed.

4.1 Data Sources

In the study, team-based and player-based performance data for the English Premier League 2021-2022 season were obtained from two data providers, InStat [18] and Football Reference [72]. In addition, the environmental conditions data used in the study were obtained from the Weather Underground Website [105] from the historical data sets of the closest measurement station to the match location according to the date and time of the match. Because InStat shares data through paid membership, data was gathered via a premium membership account. Football Reference is a website that collects data from a public crowd. This data was compiled by web scraping with the Python Selenium package. Weather Underground is a public crowd-sourced website that records data from weather measurement stations in the United Kingdom together with date, location, and time information. The same procedure was used to acquire data from this source. Table 14 shows the different types of data obtained from data providers.

Table 14: Different Data Types

Data Providers	**Data Types**
InStat [18]	Premier League 2021-22 Season All Matches In Game Statistics (42 features) Premier League 2021-22 Season Team Statistics (66 features) Premier League 2021-22 Season All Players Physical Data (4 features)

Table 14 continued:

Football Reference [72]	Premier League 2021-22 Season All Matches In Game Statistics (17 features) Premier League 2021-22 Season All Matches Expected Goal Change in Minute
Weather Underground [105]	Premier League 2021-22 Season All Matches Environmental Conditions (4 Features)

4.2 Data Explorations

The English Premier League 2021-2022 season consists of 380 matches between 20 teams and 495 players. The home team won 163 matches during the season, the away team won 129, and 88 were drawn. A total of 35374 minutes of football was played, including extra time. Throughout the season, 22 different head referees were in charge. The average temperature was 52.6 degrees Fahrenheit, the average humidity was 72.5 percent, the average wind speed was 9.9 miles per hour, and the average outside air pressure was 29.6 inHg.

According to the data received from the InStat data source, seven tactical formations were employed throughout the season. Figure 10 depicts the usage rates of tactical formations for home and away teams during the season. The tactical changes made by the teams during the match were considered when creating these graphs. The 4-3-3 was the most popular formation for both home and away teams, totaling 20157 minutes. The 4-1-4-1 tactical formation, on the other side, was the least popular by both home and away teams, with only 1326 minutes played during the season. During the season, 1004 goals were scored.

Seven tactical formations were used throughout the season in the data obtained from InStat. The usage rates of these tactical formations during the season are given in Figure 10 from the perspective of home teams and away teams. While preparing these graphs, the tactical changes made by the teams during the match were considered. The 4-3-3 tactical formation has been the most preferred formation by both home and away teams. Teams stayed on the pitch with a 4-3-3 formation for 20157 minutes. On the other hand, the 4-1-4-1 tactical formation was the least preferred formation for both home and away teams and was preferred for only 1326 minutes throughout the season. A total of 1004 goals were scored during the season. An average of 2.65 goals were scored per match and a goal per 35.2 minute when extra time is included.

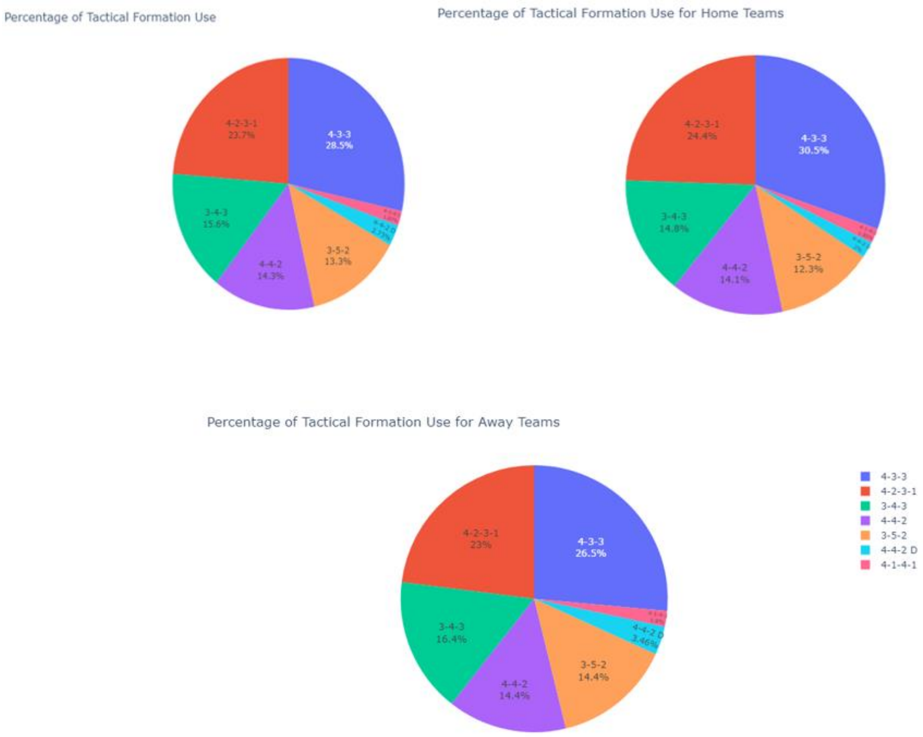


Figure 10: Tactical formation preference percentages

The dataset is segmented for each substitution, score, and tactical formation change within the match. In this context, the dataset consists of match segments ranging from 1 minute to 78 minutes, with an average segment length of 11.03 minutes. Figure 11 depicts graphs showing the distribution of match segments. There was a net expected goal difference of 0.297 per segment. During the season, 1.6 net expected goals per match were measured. This means an average difference of 1.6 expected goals per match between one team and the other.

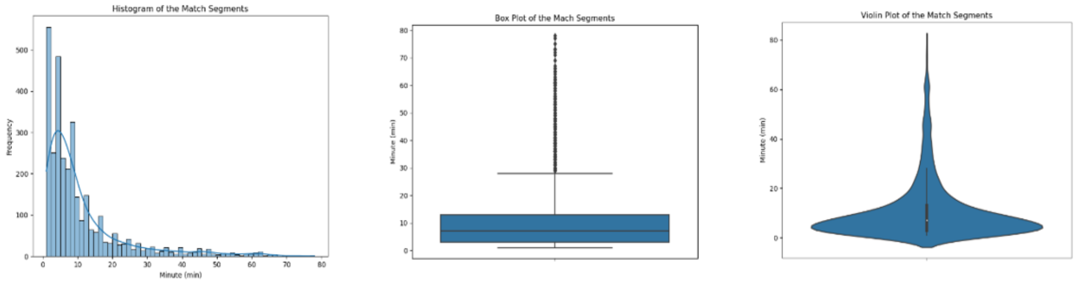


Figure 11: Distribution of the match segments in minute

Net expected goals per minute is the target feature for the predictive model. This feature is obtained by dividing the net expected goal value generated per segment by the segment duration. Graphs showing the distribution of this feature in the dataset are given in Figure 12. On a segment basis, an average of 0.053 net expected goal difference per minute is measured, while this value is 0.0177 per match. On a segment basis, the variance of the net expected goals per minute parameter is 0.014, while the standard deviation is 0.118.

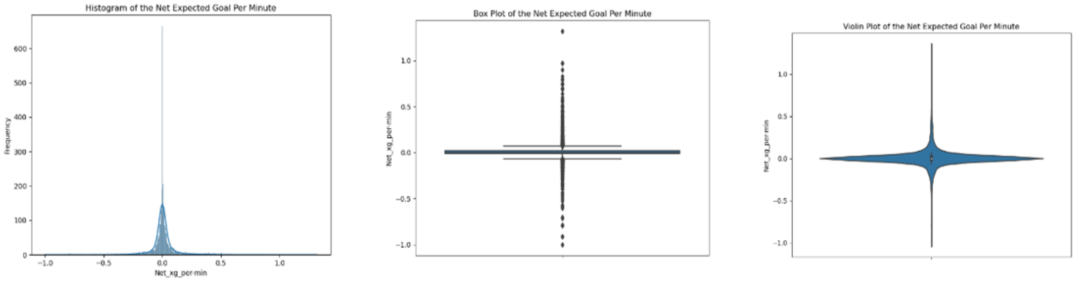


Figure 12: Distribution of the net expected goal per minute

Figure 13 shows the relationship between the expected net goal value per minute and segment duration. As seen in the figure, there is an inverse correlation between the increase in the duration of the game in the segment and the goal value. As the game duration represented by the segment decreases, the goal value can take very high values. This may cause outliers in the data. The outlier removal process described in the "Preprocessing" section details the filtering based on segment duration to remove these outliers from the data.

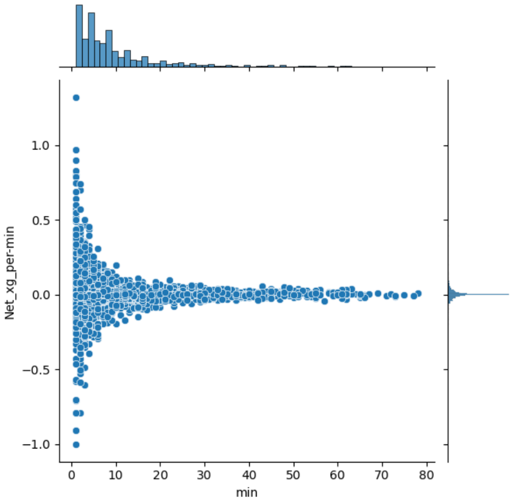


Figure 13: Net expected goal per minute and segment game duration

Correlation measures the link between two or more variables. A high correlation between variables shows that they have a strong relationship. When data points are highly correlated, the model can become excessively dependent on specific correlations in the training data, making generalization to new, untested data difficult. As a result, a model may outperform training data but perform poorly on real-world data [123]. Overfitting refers to these kinds of situations.

Multicollinearity is a potential problem that arises when using correlated data and can occur when predictor variables in a regression model are strongly correlated. As a result, estimations of model coefficients that are unstable and inaccurate can be challenging to interpret [123]. In this context, the relationships between the features in the predictive model were examined, and Figure 14 depicts the correlation matrix. This matrix does not include columns representing player roles because multiple representation options (vectorization, embedding, categorization) were investigated.

When the correlations of the predictive model's features with the target value are examined, it is discovered that Feature Score and Net Rating have the highest correlation values, with values of 0.23 and 0.22, respectively. These features may have a negative impact on the Optimization Model's outcomes. However, a feature analysis of the predictive model and the weights of the features in the model are required for this conclusion. After performing that, if Score and Net Rating values are found to be excessively dominating compared to player roles and other attributes, they will be eliminated from the predictive model to be used in the optimization model. Section 4.5 presents the feature analysis.

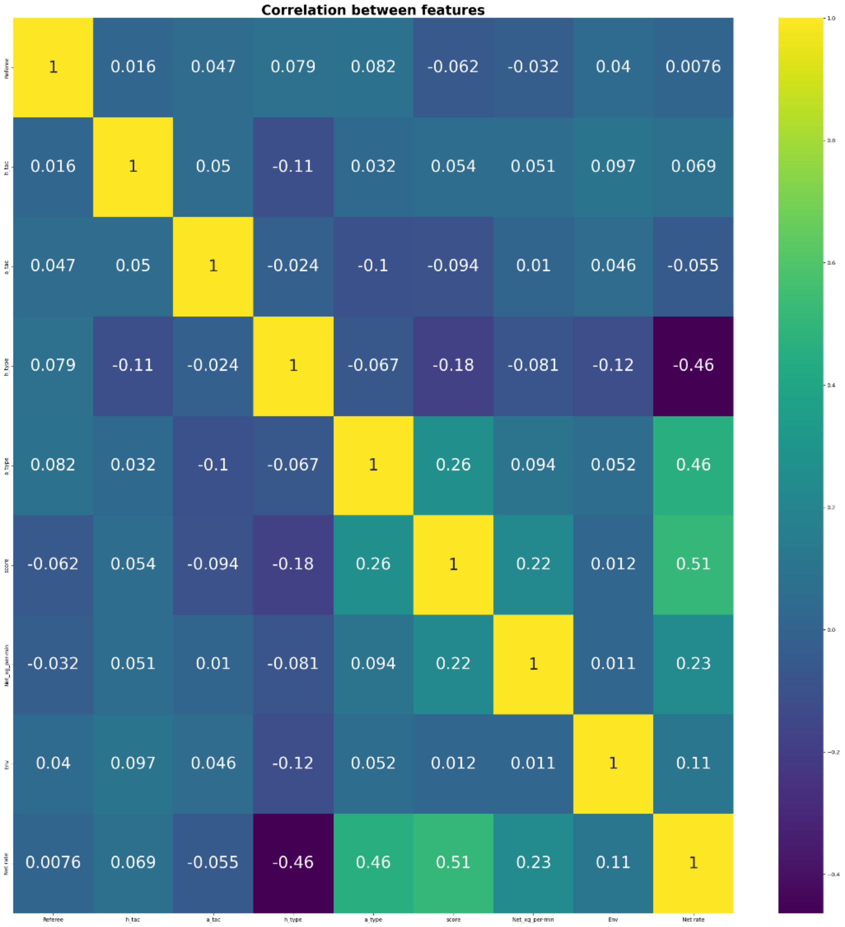


Figure 14: Correlation map of predictive model features

4.3 Player Role Determination

The study examined three distinct player role representations for the predictive model. Categorical representation, vectorization of the categorical representation with an embedding layer, and vectorial representation using NMF are the three methods. The Hierarchical Clustering algorithm mentioned in section 3.5.4 was employed for the categorical representation. The second representation involves categorizing the players based on their roles and then vectorizing the team's player selections with an embedding layer. NMF depicts each player as a vector based on the statistics of his position and the survey weights. The filtering procedures provided in Section 3.8.2.2 were tried for all three methods. Table 15 shows the effects of the filtering strategy on data size.

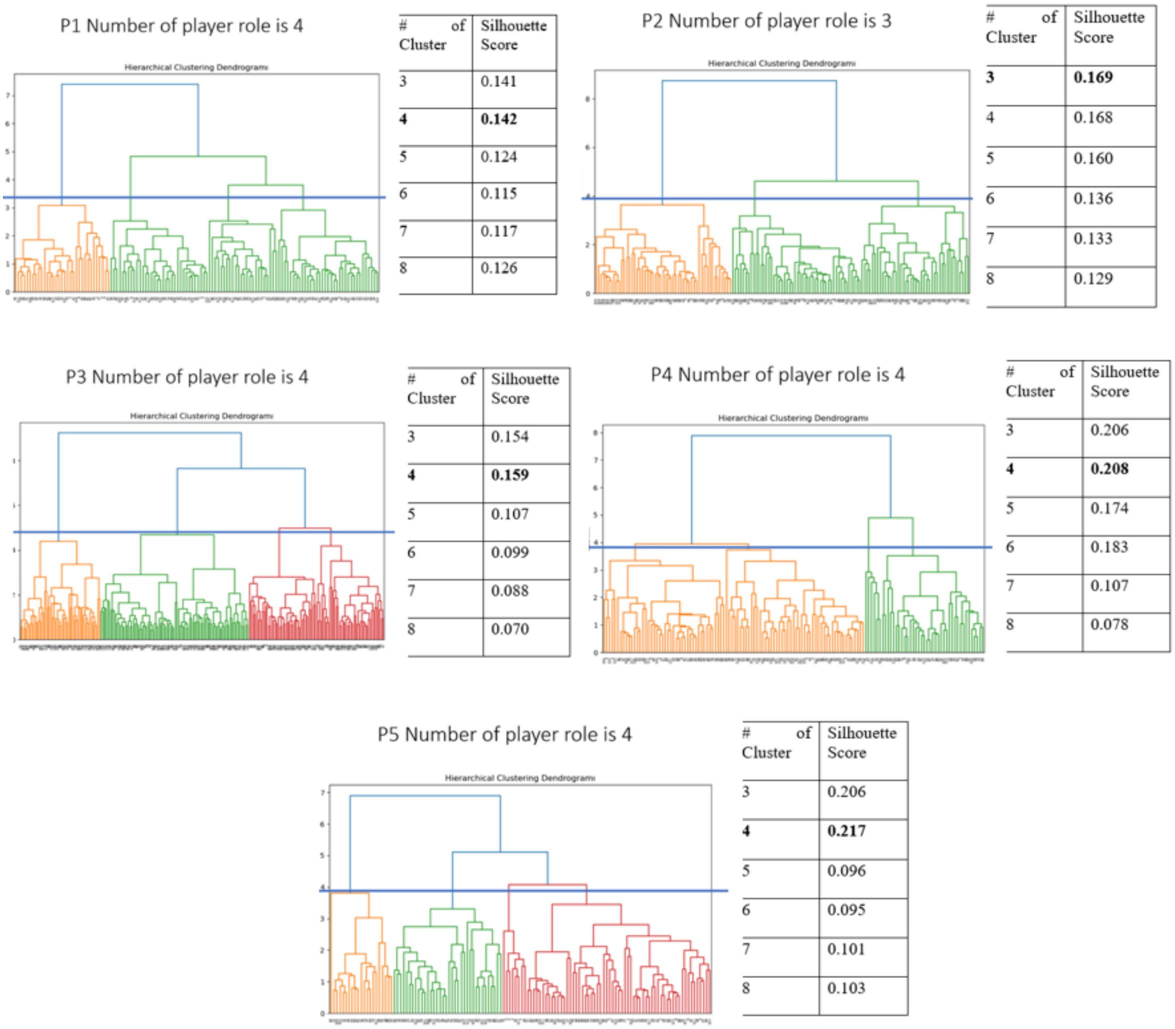
Table 15: Data Size and Filtering Method

Position/ Filtering Method	**No Filter**	**Only One Game Filter**	**15-Minute Average Filter**	**Total 30-Minute Filter**
P1: Central Defenders	131(# of players)	114 (# of players)	126(# of players)	126(# of players)
P2: Back Players	169(# of players)	122(# of players)	158(# of players)	145(# of players)
P3:Centeral Middlefielders	263(# of players)	209(# of players)	236(# of players)	226(# of players)
P4:Side Forwards/ Wingers	186(# of players)	125(# of players)	150(# of players)	137(# of players)
P5: Forwards	157(# of players)	110(# of players)	133(# of players)	126(# of players)

Considering the number of lost data and the effects of lacking information on the model, a total of 30 minutes filter was chosen as the most optimal method. Although the average 15-minute filter saves more data overall, the data cleaned with this filtering method contains players who take more time overall but frequently substitute in certain positions and perform in short periods.

4.3.1 Hierarchical Clustering

Hierarchical clustering, an unsupervised machine learning technique, groups related data points depending on their similarities or differences. The primary purpose of hierarchical clustering is to generate dendrograms, tree-like hierarchical representations of data points. The roles assigned to the players inside the grouped positions were defined using dendrograms. The dendrograms were created by weighing the players' in-match data based on their weight scores from the survey results discussed in Section 3.5. Figure 15 shows the dendrograms for each position and the ideal number of roles. When determining the optimal number of roles, tree-structured data points were selected to provide the broadest definition and most inclusive representation within the data group.



P1 Number of player role is 4

# of Cluster	Silhouette Score
3	0.141
4	**0.142**
5	0.124
6	0.115
7	0.117
8	0.126

P2 Number of player role is 3

# of Cluster	Silhouette Score
3	**0.169**
4	0.168
5	0.160
6	0.136
7	0.133
8	0.129

P3 Number of player role is 4

# of Cluster	Silhouette Score
3	0.154
4	**0.159**
5	0.107
6	0.099
7	0.088
8	0.070

P4 Number of player role is 4

# of Cluster	Silhouette Score
3	0.206
4	**0.208**
5	0.174
6	0.183
7	0.107
8	0.078

P5 Number of player role is 4

# of Cluster	Silhouette Score
3	0.206
4	**0.217**
5	0.096
6	0.095
7	0.101
8	0.103

Figure 15: Determination of ideal number of clusters using dendrograms

The categorization results for player roles can be compared to similar studies by Elvan [28], Li [41], and Kalenderoğlu [48]. By emphasizing the individual characteristics of their performance, player roles can also be matched with player role descriptions in video game series such as Football Manager and FIFA. The conceptual definitions of player roles are excluded from the scope of the study. However, to compare the player roles and groupings in the literature, the player role groupings in the three studies mentioned above are shown in Table 16. For comparison with the literature, the player characteristics of the players grouped in the study according to the skill sets classified in Table 10 are shown for each position in Figure 16.

Table 16: Player Roles Definitions from Literature

Player Positions	**Elvan (2021) [28]**	**Li (2022) [41]**	**Kalenderoğlu (2019) [48]**	**Our Study**
Central Defenders	2 Groups: -Aggressive CB -Stable CB	2 Groups: -Central Defender -R or L Ball Playing Defender	1 Group: -Stoppers	4 Groups
Back Players	2 Groups: -Wingback -Fullback	2 Groups: -R or L Back -R or L Wing Back	2 Groups: -Wing Backs -Defensive Backs	3 Groups
Central Midfielders	5 Groups: -Defensive Midfielder -Regista -Holding Midfielder -Box-to-box Midfielder -Central Attacking Midfielder	3 Groups: -R or L Defensive Midfielder -Playmaker -Wide Midfielder	3 Groups: -Defensive Midfielders -Central Midfielders -Advanced Playmakers	4 Groups
Wingers	2 Groups: -Classic Winger -Attacking Winger	2 Groups: -R or L Winger -Inside Forwards	1 Group: -Wingers / Set Piecers	4 Groups
Forwards	5 Groups: -Link-up Striker -Tank -False 9 -Speedy Striker -Second Striker	4 Groups: -Second Striker -Mobile Striker -Poacher -Target Men	5 Groups: -Winger Forwards -Inside Forwards -Target Men -Advanced Forwards -Shooters / Set Players	4 Groups

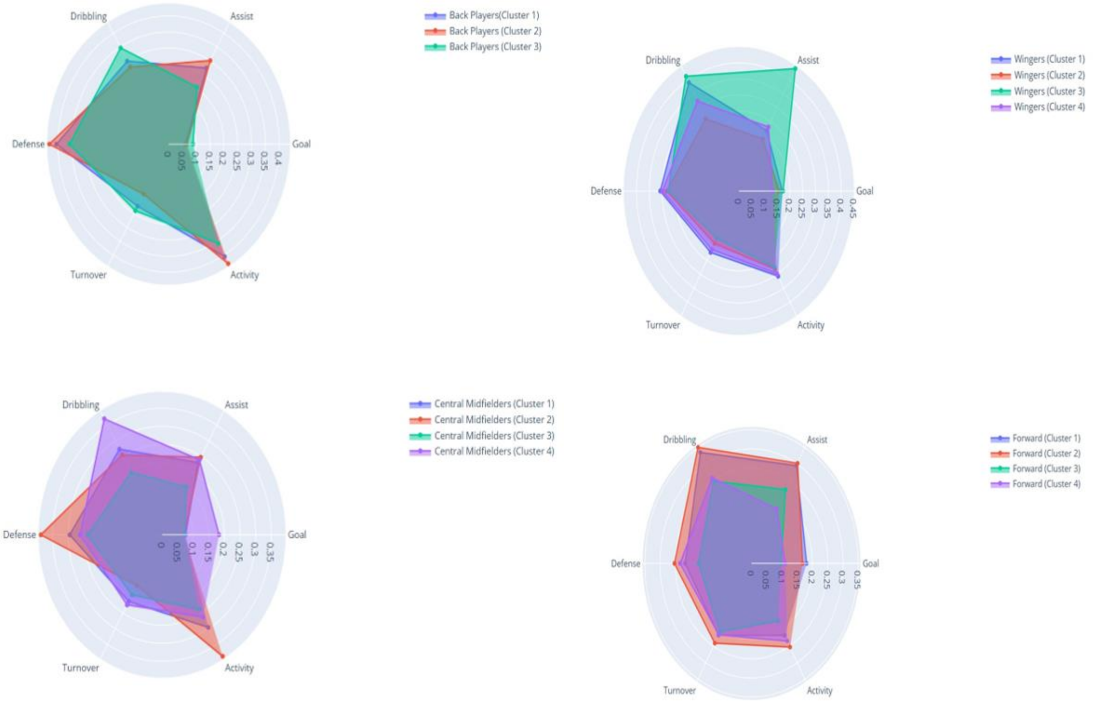


Figure 16: Player roles representation

4.3.2 Embedding

The method of transforming categorical variables into dense vectors in a lower dimensional space is called embedding. Using word embedding models is a popular way to create word vectors [124]. These vectors capture the natural patterns and correlations between many categories. Categorical variables are embedded according to a similar principle. Embedding categorical variables by projecting them onto continuous vector spaces where similar categories are closer together allows models to capture patterns and similarities between various categories better.

The embedding layer is employed in conjunction with hierarchical clustering in the study. This strategy demonstrates the coherence of players clustered according to their roles with other players' roles. The predictive model dataset includes home and opposing team lineups and categorical data on ten different player roles (excluding the goalkeeper). This data is then converted into a 5-dimensional vector with an embedding layer. Each vector dimension corresponds to one of the five on-field positions depicted in Figure 17. In this method, the study incorporates the players' roles in their positions and their compatibility with other team members. When a defensive central midfielder plays with an offensive central midfielder, the resulting vector will differ from when two defensive players play together. Figure 17 depicts an example of Arsenal FC's home and away lineup preferences utilizing 5-dimensional vectors created by the embedding layer and displayed on the 3D plane using PCA. When the home matches are examined, the formation vectors in matches where Arsenal Club loses points differ from those in matches where it wins.

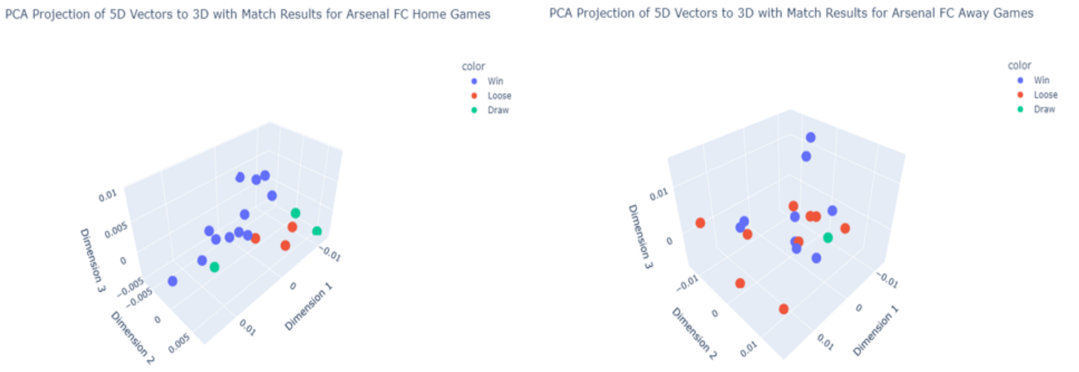


Figure 17: Arsenal FC home and away games lineup vectors obtained using embedding layer projected 3D space using PCA

4.3.3 NMF

The traditional NMF function attempts to depict the data's inherent structure. As a result, it cannot be used with importance weights. The findings of the expert survey produced position-specific importance weights for the statistics in each position. The Feature Weighted Non-Negative Matrix Factorization (FW-NMF) function, a variant of NMF, was used to apply these weights. Feature weights can be included since the basis vectors of the factor matrix F in FW-NMF must be convex combinations of data points. These weights show the importance or relevance of each feature in the factorization process [125].

The players for each position were turned into an N-dimensional vector using the position's importance weights and the FW-NMF method. In the literature, the number of representative vectors was calculated independently for each skill set of the players, such as shooting, passing, and defense [41],[49]. These vectors are added together to determine the total number of dimensions (N) of the vector representing all of the player's features. Within the predictive model, tests were run for variable N numbers. Figure 18 depicts the player vectors for the central defender position for the N is equal to three.

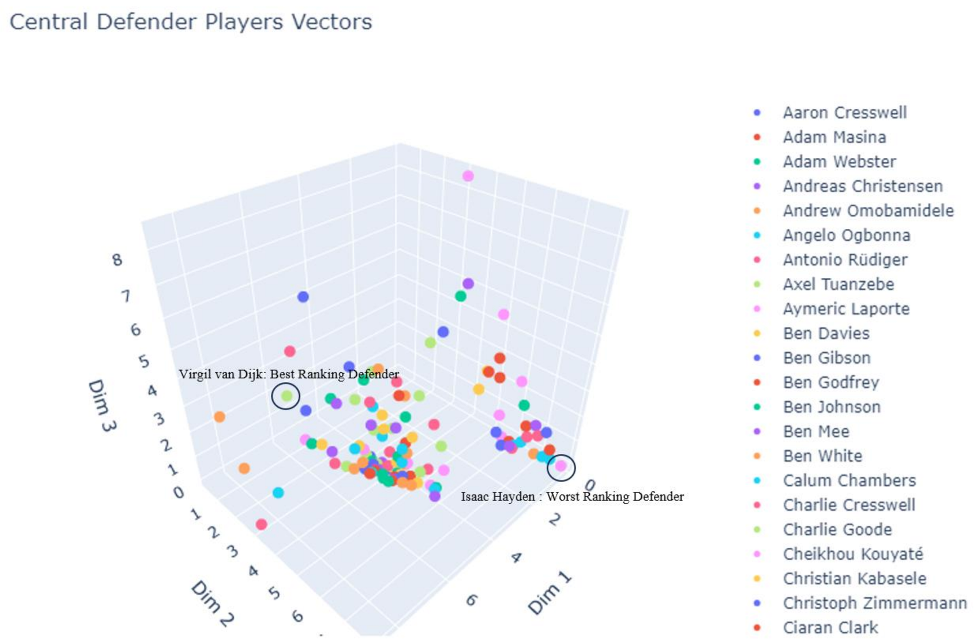
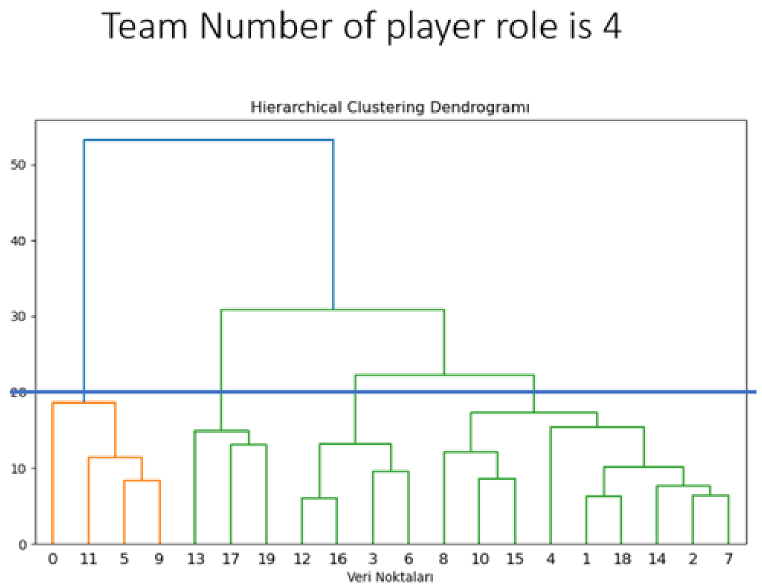


Figure 18: PL21-22 season central defender players player vectors

4.4 Team Characteristic Determination

The game characteristics of the teams were determined using hierarchical clustering and FW-NMF algorithms, similar to the player role determination step. The ideal number of clusters in the hierarchical clustering method was calculated using a dendrogram. For three and four clusters, the estimated Silhouette Scores are fairly close. The ideal number of clusters was identified as four based on a more detailed representation in the data and the region covered by the branches as seen in the dendrograms. Team values were acquired by web scraping method from Transfermrkt(13) website on September 1, 2021, the end date of the summer transfer season, in order to compare the clustering implementation and examine the association between playing styles of the teams and their values. The teams were clustered using the K-Means clustering algorithm based on the team values generated by Transfermrkt(13), and the cluster results are shown in Table 17.



# of Cluster	Sillhoute Score
3	**0.290**
4	**0.289**
5	0.207
6	0.194
7	0.182
8	0.172

Figure 19: Ideal number of cluster determination using dendrograms for team playing characteristics

Table 17: Team Clusters

Team Names / Clustering Method	**Hierarchical Clustering (# of cluster is 4)**	**K-Means Clustering using Market Value (# of cluster is 4)**	**Team PL21-22 Season Finishing Position and Cluster Every 5 Position**
Arsenal FC	1	2	1-5 Standings: Cluster 1 (# Seeding is 5)
Aston Villa	3	2	11-15 Standings: Cluster 3 (# Seeding is 14)
Brentford	3	3	11-15 Standings: Cluster 3 (# Seeding is 13)
Brighton & Hove Albion	2	3	6-10 Standings: Cluster 2 (# Seeding is 9)
Burnley	3	4	16-20 Standings: Cluster 4 (# Seeding is 18)
Chelsea	1	1	1-5 Standings: Cluster 1 (# Seeding is 3)

Table 17 continued:

Crystal Palace	2	3	11-15 Standings: Cluster 3 (# Seeding is 12)
Everton FC	3	2	16-20 Standings: Cluster 4 (# Seeding is 16)
Leeds United	3	3	16-20 Standings: Cluster 4 (# Seeding is 17)
Leicester City	3	2	6-10 Standings: Cluster 2 (# Seeding is 8)
Liverpool FC	1	1	1-5 Standings: Cluster 1 (# Seeding is 2)
Manchester City	1	1	1-5 Standings: Cluster 1 (# Seeding is 1)
Manchester United	2	1	6-10 Standings: Cluster 2 (# Seeding is 6)
Newcastle United	3	3	11-15 Standings: Cluster 3 (# Seeding is 11)
Norwich City	4	4	16-20 Standings: Cluster 5 (# Seeding is 20)
Southampton	3	3	11-15 Standings: Cluster 3 (# Seeding is 15)
Tottenham Hotspur	2	2	1-5 Standings: Cluster 1 (# Seeding is 4)
Watford	4	4	16-20 Standings: Cluster 4 (# Seeding is 19)
West Ham United	3	3	6-10 Standings: Cluster 2 (# Seeding is 7)
Wolverhampton Wanderers	4	3	6-10 Standings: Cluster 2 (# Seeding is 10)

The results of the hierarchical clustering to determine the game characteristics of the teams are compared with the results of the clustering based on the market value of the teams and the Premier League 2021-22 end-of-season point ranking, as shown in Table 17. Each grouping and ranking is broken down into four groups. The first group consists of the league's top five teams, while the last group consists of the league's bottom five teams. When the results is examined, it is clear that clustering based on teams' playing styles, as well as clustering based on teams' economic worth, have a substantial link with league standings. While the playing styles of the teams are more successful in identifying the teams that finish at the top of the league, the economic-based clustering model explains the teams that finish at the bottom of the league more successfully. It is also observed that the teams that finish at the top of the league have more possession of the ball and play with higher passing percentages and organize their attacks as counterattacks to a lesser extent. The teams finishing last in the league, on the other hand, show the reverse pattern. The Wolverhampton Wanderers (Wolves) are an exception to this rule. Although the Wolves are among the top fifteen teams in the upper center of the league in terms of the economic worth that defines the league, their counter-attacking style of play places them in the same cluster as the teams that finish at the bottom of the league.

The Kendall Tau Distance measure was used to determine which clustering approach best describes the final season ranking of teams. The Kendall Tau Distance computes the rank differences between two ordered sequences or series. This statistic assesses the link between two sequences by measuring rank consistency [126]. Kendall Tau Distance returns 1 if the two sequences are exactly the same and -1 if they are completely different. To compare the clustering results stated in Table 18, the Kendall Tau distance was utilized.

Table 18: Clustering Comparison

Clustering Pairs	**Kendall Tau Distance**
Hierarchical Clustering - Standings	0.6835
K-means using Market Values- Standing	0.6536

When the two clustering methods are compared, clustering based on team playing styles performs slightly better than clustering based on economic factors in explaining the end-of-season point ranking. However, both clustering methods produced similar and successful outcomes when it came to grouping teams based on their performance.

Figure 20 depicts the vectorial representation of the teams created with FW-NMF. As shown in the graph, the teams that finish in the top three spots in the league stand out from the others. Burnley is another team that draws attention in the graph. Despite finishing 18th in the league and having a limited lineup quality, Burnley plays a more attacking and possession-oriented game compared to teams of similar strength, which distinguishes Burnley from the other teams.

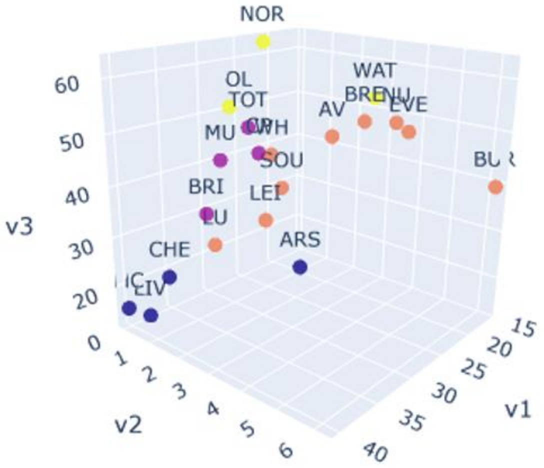


Figure 20: Team vectors using FW-NMF

4.5 Performance of the Predictive Model

In this section, the performance of the machine learning models used in the predictive model is compared using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics.

4.5.1 Modeling

Multi-Layer Perceptron (MLP), Attention, Gradient Boosting Regressor, LGBM Regressor, and Cat Boost Regressor models were tested for the predictive model. MLP and Attention models are classified as deep learning models in the machine learning sub-section. Because the Attention model has an embedding structure, the embedding step was skipped in the dataset created for testing this model. The other four models, with the exception of the Attention model, were tested on three distinct datasets. For player roles and team game characteristics, the datasets were parameterized using hierarchical clustering, hierarchical clustering with embedding, and FW-NMF methods. Because the embedding layer is included in the attention model, it is only tested with the hierarchical clustering dataset.

The Weight and Biases application [127] was used to compare models, record the hyperparameter sets used and visualize the data. Weights & Biases (WandB) is a platform for tracking and visualizing machine learning experiments. It offers a number of features to assist with various stages of the machine-learning workflow.

The model preprocessing steps are given in Section 3.8.2.2. The preprocessing for outlier removal outlined in Section 8 substantially impacts model performance. The predictive model's MAE drops as the length of the match segments rises. However, removing short-duration parts from the data resulted in significant data reductions. This reduction in data size prohibits the models from making accurate forecasts due to the fundamental principles of machine learning and deep learning models. As a result, the data was filtered to choose match segments longer than 15 minutes in order to select the most successful prediction model. Table 19 shows the influence of match segment length on model performance and data amount for the Attention model.

Table 19: Attention Model Metrics and Match Segment Filtering

Metrics/ Match Segments	**No Filter**	**≥ 5 Minutes**	**≥ 10 Minutes**	**≥ 15 Minutes**	**≥ 20 Minutes**	**≥ 25 Minutes**	**≥ 30 Minutes**
MAE	0.05732	0.02997	0.01998	0.01658	0.01551	0.01237	0.01088
RMSE	0.1185	0.04716	0.02848	0.02313	0.01959	0.01689	0.01329
Data Counts	3212	2142	1647	1105	884	549	368

The MAE and RMSE values of the five models tested according to the player role representation and team game characteristic representation are shown in Table 20.

Table 20: Results of the Different Models for 15 Minute or Longer Match Segments

Models / Data Preparation Technique	**Hierarchical Clustering**	**Hierarchical Clustering with Embedding**	**FW-NMF**
Multi-Layer Perceptron (MLP)	0.0207 (MAE) / 0.0282 (RMSE)	0.01771 (MAE) / 0.0239 (RMSE)	0.0158 (MAE) / 0.0214 (RMSE)
Gradient Boosting Regressor	0.0165 (MAE) / 0.0227 (RMSE)	0.0157 (MAE) / 0.0203 (RMSE)	0.0161 (MAE) / 0.0209 (RMSE)
LGBM Regressor	0.0184 (MAE) / 0.0250 (RMSE)	0.0175 (MAE) / 0.0240 (RMSE)	0.0172 (MAE) / 0.0226 (RMSE)
Cat Boost Regressor	0.1726 (MAE) / 0.02312 (RMSE)	**0.015 (MAE) / 0.0198 (RMSE)**	**0.0153 (MAE) / 0.020 (RMSE)**

Table 20 continued:

Attention	0.0166 (MAE) / 0.0231 (RMSE)	-	-

As shown in Table 20, the model with the lowest MAE among the tested models is the embedded version of Cat Boost Regressor. Considering these findings, the Cat Boost Regressor model and hierarchical clustering with embedding data structure were selected for the predictive model. The selected predictive model performed feature analysis to determine which features interact with the expected net goals per minute parameter and in which way.

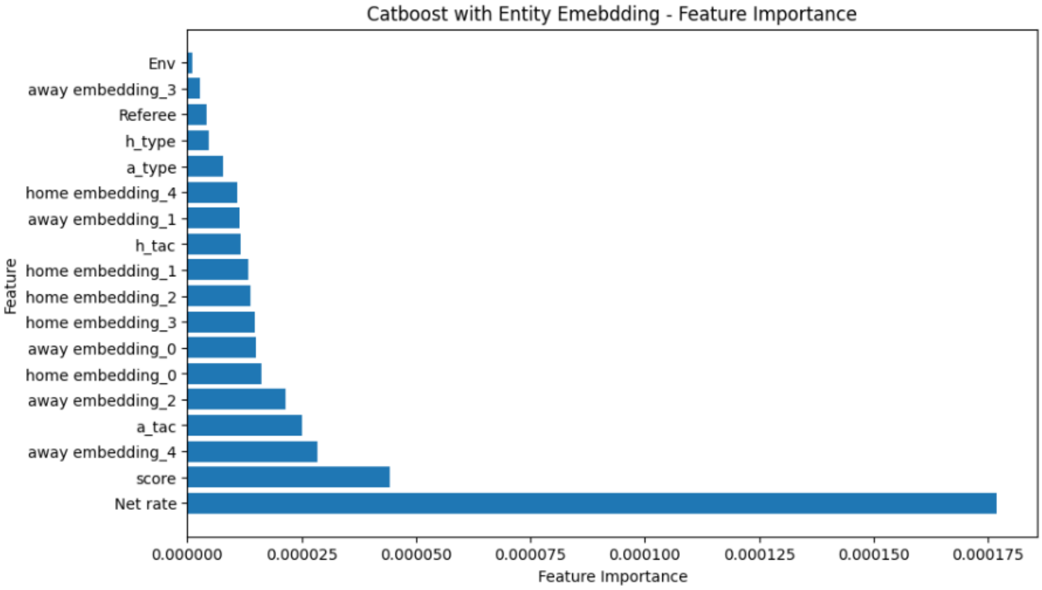


Figure 21: Feature importance of the selected predictive model

4.5.2 Parameters

The Grid Search method was used to optimize the hyperparameters of the Cat Boost Regressor algorithm. Table 21 shows the hyperparameters examined with Grid Search and the values selected as a result of the optimization.

Table 21: CatBoost Regressor Selected Hyperparameters

Parameter Name	**Optimized Value**
Learning Rate	0.01
Depth	3
L2 Leaf Regularization	0.1
Bagging Temperature	0.8
Number of Iterations	500
Sub Sample	0.9
Column Sample by Level	0.9
Minimum Data in the Leaf	1
Random Strength	0.1
Border Count	32
Number of Embedding	5

4.5.3 Comparison with the Literature

As stated in Section 3.8.2, there is no benchmark model against which the proposed predictive model's outcomes can be directly compared. As a result, the internal consistency of the predictive model must be evaluated. First, the relationship between the expected net goals per minute parameter, which is employed as a performance indicator in the predictive model, and the actual goal is examined. In order to better understand the results of the predictive model, the relationship between expected net goals and net score should be examined first.

The number of goals scored in the English Premier League and the total expected goal statistics generated by the teams for the 6-season period from the 2017-2018 season in which the Football Reference (14) website started to provide the expected goal parameter to the 2022-2023 season, were analyzed. According to this data, 6092 goals were scored in 2280 matches and the total expected goal value was 6022. The difference between goals scored per match and expected goals was calculated as 0.0307. This value represents the deviation between the expected goals per match parameter and the number of goals scored for each team. The distribution of goals scored and expected goals by season is presented in Table 22.

Table 22: Goal – Expected Goal Distribution by Seasons

Season	**Number of Goal**	**Total Expected Goal**	**Goal - Expected Goal Difference Per Match**
2017-2018	988	945.8	0.111
2018-2019	1040	1022.1	0.047
2019-2020	1002	973.7	0.074
2020-2021	986	982.4	0.009
2021-2022	1037	1017,9	0.050
2022-2023	1039	1080.1	-0.108
Total 6 Seasons	6092	6022	0.031

To assess the model's actual performance, a 30-minute filter was used, which produces the most efficient output within the match segment lengths indicated in Table 19. MAE and RMSE are scale-dependent model performance measuring metrics. Because the model predicts the net expected goals per minute measure, utilizing these metrics to comment on the score of a 90-minute match is challenging. As a result, a scale-independent indicator, the Mean Absolute Percentage Error (MAPE), is used to assess the most effective model's performance throughout a 90-minute match. However, the dataset's structure prevents the use of proportional error metrics. This is due to the fact that the target value of expected net goals per minute can be zero. When the target value is 0, the MAPE metric is calculated as infinite because the denominator value in the MAPE calculation is zero. To address this issue, rows in the dataset with zero target values were eliminated.

The predictive model's error metrics were 0.0113 MAE and 0.0138 RMSE when trained using the Cat Boost Regressor algorithm with five embeddings and a 30-minute filter. The error metrics after removing the 0 target values from the dataset to derive the scale-independent metric were 0.0135 MAE, 0.0177 RMSE, and 24.9 MAPE.

The derived MAPE value of 14.9 indicates that the model predicts the expected net goal value with an error of 24.9% in a 90-minute match. To compare this value to other studies, 2280 Premier League matches played between the 2017-18 season and the

2022-23 season were evaluated, and an average goal difference of 2.88 was calculated. When this value is predicted with a 24.9 percent error, the predictive model may measure the match result with a 0.717 goal difference error. This result outperforms Herbinet's [25] MAE of 0.861 when measuring the score of a single team. This calculation, however, is insufficient for a complete comparison; it is only a rough calculation to compare the predictive model with other studies in the literature and has the following drawbacks.

- The best version of the predictive model is not used.
- While Herbinet's model predicts per team score, the predictive model focuses on net expected goal value.
- Variation between net expected goal and net score was not included in calculations.
- Average goal difference data only contains 6 Premier league seasons. It is not large enough to generalize the results.
- In calculations, differences between the net expected goal value per minute and the net expected goal value per 90 minutes are ignored. Proportional correlation was used between the two metric scales through the MAPE value.

4.6 Best Lineup Selection Model

In this chapter, the results and performance of the optimization model for ideal lineup selection using predictive model's outputs are evaluated.

4.6.1 Modeling

The most successful version of the predictive model, the Cat Boost model trained with 30-minute filtering and five embedding, was employed in the optimization model. However, when Figure 21 is analyzed, it is observed that the "Net Rate" feature has a greater influence on the model result than the embedding characteristics. The "Net Rate" feature was not employed in the predictive model used in the optimization model to prevent this problem. The MAE and RMSE values of the model without this feature were 0.0102 and 0.01285, respectively. The improvement in model performance over the generic model was surprising but a positive improvement. Figure 22 depicts the feature importance values of the predictive model utilized in the optimization model.

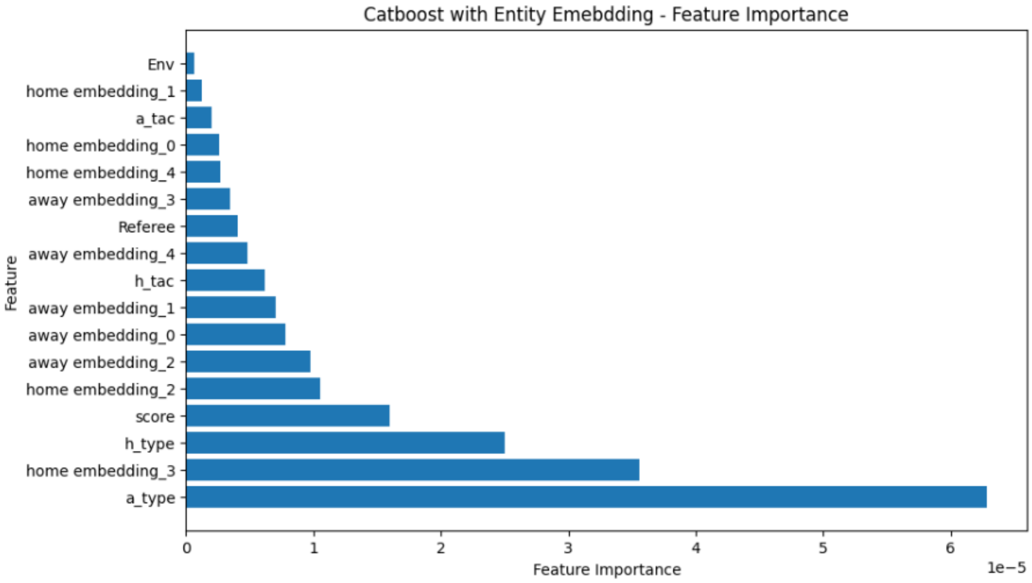


Figure 22: Feature importance of the predictive model used in optimization model

The optimization model tests the predictive model for all lineup selection alternatives, and the pick with the highest performance output is selected as the optimal lineup selection. For this purpose, it measures the performance output of the predictive model under the conditions in which all players in the match lineup, opponent team player preferences, opponent tactical formation, and preferred tactical formation are held constant. Only positions and player roles in which players have played at least thirty minutes during the season are given as input to the prediction model. Players are not tested in positions they have not played before in the optimization model.

The predictive model produces the same results when more than one player is tagged with the same player role, which is a limitation of the approach used. In such circumstances, the Cat Boost Regressor FW-NMF model is utilized to compare the predictive model's outputs of these two players using different representations. The player with the highest performance output is chosen in the FW-NMF model.

4.6.2 Validation

Ten matches were chosen and evaluated from both the home and away team viewpoints for model validation. In this manner, the optimization model outcomes for the match start for 20 teams are compared to the actual selections. Furthermore, in the validation dataset, the optimization model's suggested player substitutions and the actual substitutions are compared based on the match result, tactical change, and opponent player substitution.

Match squads in the Premier League are made up of 18 players. Of these 18 players, two are goalkeepers, one is an ace, and one is a substitute. The optimization model evaluates the 16 players based on their positions and roles to determine the best starting lineup of ten players. In the validation dataset, 200 players for starting lineup were chosen from a pool of 320 players for a total of 20 matches. An average of 6554.35 trials were carried out for each team, while keeping the opponent's tactical formation, player preferences, and the team's preferred tactical formation constant. The FW-NMF approach was used to develop a predictive model for players with the same player role while holding other variables constant. With a success score of 79.5%, the predictive model correctly predicted 159 of 200 players selected in 20 matches. 14 of the 41 incorrectly predicted players had the substitute player role label but were not accurately predicted due to the player's position and the team's position on the field, and his strong foot. The optimization model predicted 27 players incorrectly for the starting 11 preferences and 2 players for the starting line-up of 10 players for each match.

In the match between Leicester United and Southampton, the model incorrectly predicted the team selection of four players in Southampton's starting lineup. Similarly, in the match between Brighton and West Ham United, the optimization model incorrectly predicted 4 players in West Ham's starting lineup. Both West Ham and Southampton were defeated in these matches. The Aston Villa's starting 11 was the model's most successful forecast in the match between Manchester City and Aston Villa. Despite the model correctly forecasting the coach's player selection for Aston Villa with a ten to ten accuracy, Aston Villa lost this match. Even the best lineup selection in the optimization model failed to predict positively Aston Villa's expected net goals per minute against Manchester City. Another correct prediction was the Arsenal lineup in the match against Everton in which Arsenal won.

Despite having less similarities in their starting lineups, three teams left the match with an unfavorable result, according to the optimization model forecasts. These are Aston Villa against Manchester City, Wolverhampton against Liverpool and Brentford against Leeds United. Only Leeds United had a better league record than their opponents. Figure 23 depicts the number of correct predictions for the starting lineup in the matches.

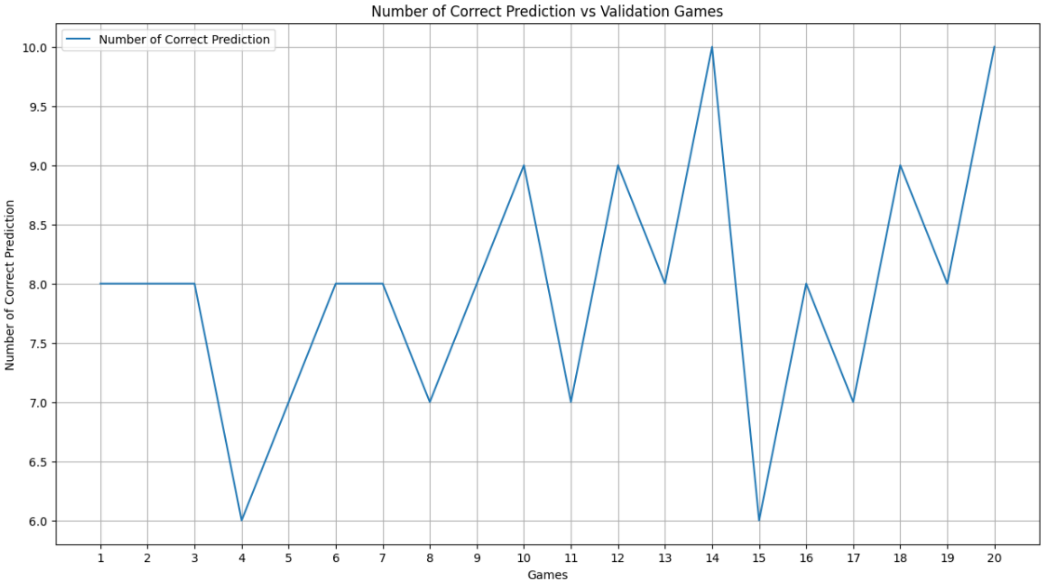


Figure 23: Number of correct predictions vs. validation games

There were 111 in-match changes in the 10 matches in the validation dataset, including 58 substitutions, 38 scoring changes, and 13 tactical formation changes, based on the match segments played without red cards. For 51 scoring and tactical changes the optimization model identified 9 optimal line-up changes. Only three of these changes occurred in the section that followed the change indicated by the optimization model. Only 7 of the 58 changes in the matches resulted in higher predictive model output than before the change. This is due to the optimization model not considering external factors such as player fatigue, in-match dynamics, and current league position. There is no benchmark model in the literature to which the optimization model can be compared.

CHAPTER 5

CONCLUSION

In this thesis, a best line-up approach is proposed that maximizes performance as measured by the expected goal per minute metric using player roles, team game characteristics, environmental factors, tactical formation, match score and opponent team parameters. This chapter focuses on the contributions of the proposed approach, its limitations and how it can be improved.

5.1 Contributions

This study has the following contributions to the literature:

- As a new approach to measuring the position-based effects of in-match statistics, the effects at each position were analyzed by taking expert opinion.
- A novel approach is proposed that incorporates the parameters of the opposing team when evaluating the performance of the teams.
- A novel approach to determine the playing characteristics of teams using in-match statistics and expert opinion.
- A novel approach to determine the playing roles of players based on playing positions using in-match statistics and expert opinion.
- A novel approach to measure the impact of environmental conditions, referee, match score and team strengths on match performance is presented.
- A novel best lineup selection approach is presented that incorporates the effects of opposing team, player roles, match score and environmental conditions.
- A novel approach on how to combine expert opinion and data driven solutions in the football domain is presented.
- A novel approach using expected goal parameter for performance measurement in football is presented.
- A novel approach is presented for a match time independent performance measurement metric using the net expected goal per minute parameter as a target value for the predictive model.
- A novel approach to parametrize the effect of teams' preferred tactical formation on match performance.

5.2 Limitations

One of the most significant duties of a football coach is lineup selection, which can vary depending on a variety of factors. Coaches assess numerous aspects for lineup selection, ranging from the players' personal relationship with the coach to their condition, training performance, and harmony with their teammates, as well as the differences they might contribute to their pairings with opposing players. In many seasons, coaches rotate and field less-than-ideal lineup selections to protect the player’s health and plan for future fixtures. In such a field, it is hard to make a definitive and absolute conclusion on player selection by analyzing solely the technical aspects of the game. The fact that football is a team, and a people-oriented sport is the study's fundamental constraint, making the best lineup prediction impossible. Therefore, through a data-driven approach, the objective of this study is to provide coaches with the most optimal player selection options based on the opponent team's play and player type. Every on-field and off-field events that is not included in the scope of the study is an obstacle to improving the study's outcomes.

This study proposes a model that predicts in-match performance by detailing the characteristics of players and teams using past match statistics and expert opinions and presents a novel approach for the best lineup selection using this model. However, the success of the best lineup approach is closely tied to the performance of the predictive model. As mentioned in Section 4.5, there is a linear link between increased playing time in a match and predictive model success. However, the quantity of needed long pieces of data is rather limited when data is collected for a single season. As a result, the fundamental constraint of this study is the lack of data supporting the proposed approach. When the proposed approach is applied to more seasons and leagues, it is intended that both its comprehensiveness and performance would rise. At the same time, the existing dataset does not have enough data for deep learning models like Attention, which is evaluated for the predictive model, to perform efficiently.

Another limitation of the dataset utilized in the study is that it does not include difficulty measures for players and teams. The frequency of an action, not its difficulty, is measured by statistical data. Due to the dataset containing no metric distinguishing between actions performed under pressure and those performed without pressure, the method utilized to distinguish players and teams is only based on the frequency of actions and the percentage of success. Furthermore, because physical and training data are not included in the dataset, players are modeled as one-dimensional based solely on on-match statistics.

In this study, the best lineup is determined by iteratively testing the performance output of the predictive model for all players in the match lineup. Based on the five on-field positions depicted in Figure 3, the predictive model determines the tactical formation and player roles. However, three of these five positions (P1, P2, and P4) are generated by grouping player positions on the left and right sides of the field using the transversal symmetry of the football field. Therefore, the player positions used in the predictive model do not include information on which wing the player plays on. For example, a player who only plays on the left wing would be selected in the best lineup output of the predictive model, regardless of which wing of the field he plays on. This may result in multiple players who can play in the same area of the field being presented together in the ideal lineup selection in some cases. This limitation is due to the limited experimental data of the study mentioned in the second paragraph. In studies with a larger data set, players can be evaluated in positions independent of the symmetry of the field.

The approach is used to determine player roles, not only according to players playing styles but also according to their skills and quality. This means that when clustering with low size data, skill and quality parameters may dominate over game characteristics and player role parameters. This limitation can be solved by increasing the number of data, or by pre-grouping players into skill groups and then separating them according to their roles.

A general limitation is the lack of established applications and research in the field. Sports analytics is an emerging topic, with few studies providing evaluation measures and outcomes. As a result, there is less research to compare the proposed framework's outcomes. Even studies that provide evaluation metrics are incomparable since existing approaches are limited to a small fraction of the available data.

5.3 Future Works

The success of the predictive model is directly tied to the performance of the best lineup model. As a result, enhancements to the predictive model will improve the success of the optimization model. The most significant improvement to the study would be to the usage of a larger data set that includes multiple leagues and seasons.

The expected goal parameter was utilized in the study to assess the performance of the teams. On the other hand, the expected goal parameter is a shot-oriented metric that is insufficient for evaluating attacks and positions that do not result in shots. Using a more up to date statistic, such as Expected Thread, in future studies could eliminate this problem.

Another improvement would be to separately assess the effects of related and external factors rather than combining them in a single model. The predictive model's performance can be increased by investigating the effects of external factors such as referees and environmental conditions on player and team performance in a separate model.

On the other hand, the statistics in the data set used in the study are normalized. However, linear normalizing techniques are insufficient for describing these statistics in depth. While there is no significant difference between making 20 and 25 passes in a match, there can be notable differences between making 2 key passes and making 5 key passes. Similarly, the difficulty of getting particular statistics in a match grows as the frequency of them increases. A forward with three accurate shots on goal, for example, may not feel the pressure of the opposing defense as much on his fourth shot as a forward with his first shot. To overcome this issue, future studies could employ a normalization technique in which the logarithmic or exponential significance of the statistic is proportional to its frequency.

The game characteristics of the teams are determined in the predictive model by gathering their performances throughout the season. However, the game characteristics of the teams may differ depending on the opponent team's performance and tactical formation. A team may not show the same playing characteristics throughout the season, and the quality of the opposition, current standings, and other external factors may significantly impact the teams' playing characteristics. As a result, evaluating the game characteristics of the teams along with the tactical formation they utilize, the strength of the opponent team they play, their league rating, and the stage of the season would be a more comprehensive method.

There is no preliminary clustering of player roles based on their talents and qualities in the proposed method. Instead, player roles and skill sets are represented by a single clustering. Identifying and displaying these two parameters separately in the predictive model can give players a more precise approach.

The players in this study were vectorized using NMF with the entire set of statistics. Instead of this technique, vectorizing players into skill groupings such as passing, defense, and off-ball play allows the player's defensive and offensive skills to be parameterized independently. A similar method can be used to identify team game characteristics.

REFERENCES

[1] “Annual Review of Football Finance,” Deloitte United Kingdom, https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance-europe.html (accessed May. 19, 2023).

[2] M. Lewis, Moneyball: The Art of Winning an Unfair Game. New York: W.W. Norton, 2003.

[3] J. Castellano, D. Alvarez-Pastor, and P. S. Bradley, “Evaluation of Research Using Computerised Tracking Systems (Amisco® and Prozone®) to Analyse Physical Performance in Elite Soccer: A Systematic Review,” Sports Med, vol. 44, no. 5, pp. 701–712, May 2014, doi: 10.1007/s40279-014-0144-3.

[4] Z. G. Li, “Sports Policy and Training Decision Support Method Based on Wireless Sensor Network,” Wireless Communications and Mobile Computing, vol. 2021, 2021, doi: 10.1155/2021/1608340.

[5] A. Cortez, A. Trigo, and N. Loureiro, “Football Match Line-Up Prediction Based on Physiological Variables: A Machine Learning Approach,” Computers, vol. 11, no. 3, Art. no. 3, Mar. 2022, doi: 10.3390/computers11030040.

[6] S. A. Salles, H. R. da Hora, M. Erthal Júnior, A. S. Velasco, and P. R. Croce, “Multiple choice method with genetic algorithm for the formation of soccer teams,” Pesquisa Operacional, vol. 42, 2022. doi:10.1590/0101-7438.2022.042.00243537.

[7] E. O. Ozceylan, “A mathematical model using AHP priorities for soccer player selection: A case study,” South African Journal of Industrial Engineering, vol. 27, no. 2, 2016. doi:10.7166/27-2-1265

[8] M. Tavana, F. Azizi, F. Azizi, and M. Behzadian, “A fuzzy inference system with application to player selection and team formation in multi-player sports,” Sport Management Review, vol. 16, no. 1, pp. 97–110, Feb. 2013, doi: 10.1016/j.smr.2012.06.002.

[9] A. Ariyaratne and R. M. Silva, “Meta-heuristics meet sports: a systematic review from the viewpoint of nature inspired algorithms,” International Journal of Computer Science in Sport, vol. 21, pp. 49–92, Mar. 2022, doi: 10.2478/ijcss-2022-0003.

[10] J. Brooks, M. Kerr, and J. Guttag, “Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 49–55, Aug. 2016, doi: 10.1145/2939672.2939695.

[11] M. Frey, E. Murina, J. Rohrbach, M. Walser, P. Haas, and M. Dettling, “Machine Learning for Position Detection in Football,” in 2019 6th Swiss Conference on Data Science (SDS), Jun. 2019, pp. 111–112. doi: 10.1109/SDS.2019.00009.

[12] A. Joseph, N. E. Fenton, ve M. Neil, "Predicting football results using Bayesian nets and other machine learning techniques," Knowledge-Based Systems, cilt 19, sayı 8, ss. 544-553, 2006.

[13] L. O. Gavião, A. P. Sant’Anna, G. B. Alves Lima, and P. A. de Almada Garcia, “Evaluation of soccer players under the Moneyball concept,” Journal of Sports Sciences, vol. 38, no. 11–12, pp. 1221–1247, Jun. 2020, doi: 10.1080/02640414.2019.1702280.

[14] J.-C. Pomerol and S. Barba-Romero, “Multicriterion decision in practice,” International Series in Operations Research &amp; Management Science, pp. 299–326, 2000. doi:10.1007/978-1-4615-4459-3_11

[15] P. Rajesh, Bharadwaj, M. Alam, and M. Tahernezhadi, “A Data Science Approach to Football Team Player Selection,” in 2020 IEEE International Conference on Electro Information Technology (EIT), Jul. 2020, pp. 175–183. doi: 10.1109/EIT48999.2020.9208331

[16] “Football professional videos and Data Platform,” Wyscout, https://wyscout.com/ (accessed Dec. 12, 2022).

[17] “OPTA data from stats perform,” Stats Perform, https://www.statsperform.com/opta/ (accessed Dec. 12, 2022).

[18] “Instat football platform discontinuation next steps • hudl,” Hudl, https://football.instatscout.com/(accessed Dec. 12, 2022).

[19] Football transfers, rumours, market values and news,” Transfermarkt, https://www.transfermarkt.com/ (accessed Dec. 12, 2022).

[20] UEFA.com, “UEFA coefficients,” UEFA.com, https://www.uefa.com/memberassociations/uefarankings/ (accessed Dec. 12, 2022).

[21] L. Pappalardo and P. Cintia, “Quantifying the relation between performance and success in soccer,” Advs. Complex Syst., vol. 21, no. 03n04, p. 1750014, May 2018, doi: 10.1142/S021952591750014X.

[22] J. H. Hewitt and O. Karakuş, “A Machine Learning Approach for Player and Position Adjusted Expected Goals in Football (Soccer).” arXiv, Jan. 19, 2023. doi: 10.48550/arXiv.2301.13052.

[23] “Assessing the performance of Premier League goalscorers,” Stats Perform, https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/ (accessed Feb. 17, 2023).

[24] V. Barnett and S. Hilditch, “The effect of an artificial pitch surface on home team performance in football (soccer),” Journal of the Royal Statistical Society. Series A (Statistics in Society), vol. 156, no. 1, p. 39, 1993. doi:10.2307/2982859.

[25] C. Herbinet, “Predicting Football Results Using Machine Learning Techniques”, Jun. 2018.

[26] M. Brechot and R. Flepp, “Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals,” Journal of Sports Economics, vol. 21, no. 4, pp. 335–362, May 2020, doi: 10.1177/1527002519897962.

[27] “Assessing the performance of Premier League goalscorers,” Stats Perform, https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/ (accessed Jan. 13, 2023).

[28] A. E. Aydemir, “A DATA DRIVEN PERFORMANCE EVALUATION FRAMEWORK FOR SPORTS ANALYTICS,” SPOR ANALİTİĞİ İÇİN VERİ GÜDÜMLÜ PERFORMANS ANALİZ ÇERÇEVESİ, Sep. 2021, Accessed: Aug. 17, 2023. [Online]. Available: https://open.metu.edu.tr/handle/11511/93119.

[29] J. Tippett, “The Expected Goals Philosophy: A Game-Changing Way of Analysing Football”, 2019.

[30] M. Herold, F. Goes, S. Nopp, P. Bauer, C. Thompson, and T. Meyer, “Machine learning in men’s professional football: Current applications and future directions for improving attacking play,” International Journal of Sports Science & Coaching, vol. 14, no. 6, pp. 798–817, Dec. 2019, doi: 10.1177/1747954119879350.

[31] M. Cavus and P. Biecek, “Explainable expected goal models for performance analysis in football analytics,” arXiv.org, Jun. 14, 2022. https://arxiv.org/abs/2206.07212v2.

[32] W. Spearman, Beyond Expected Goals. 2018.

[33] T. Laakso, K. Davids, P. Luhtanen, J. Liukkonen, and B. Travassos, “How football team composition constrains emergent individual and collective tactical behaviours: Effects of player roles in creating different landscapes for shared affordances in small-sided and conditioned games,” International Journal of Sports Science & Coaching, vol. 17, no. 2, pp. 346–354, Apr. 2022, doi: 10.1177/17479541211030076.

[34] A. J. Barake, H. Mitchell, C. Stavros, M. F. Stewart, and P. Srivastava, “Classifying player positions in second-tier Australian football competitions using technical skill indicators,” International Journal of Sports Science &amp; Coaching, vol. 17, no. 1, pp. 73–82, 2021. doi:10.1177/17479541211010281

[35] D. Abidin, “A case study on player selection and team formation in football with machine learning,” Turk J Elec Eng & Comp Sci, pp. 1672–1691, May 2021, doi: 10.3906/elk-2005-27.

[36] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, and F. Giannotti, “PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach,” ACM Transactions on Intelligent Systems and Technology, vol. 10, pp. 1–27, Sep. 2019, doi: 10.1145/3343172.

[37] I. Behravan, S. H. Zahiri, S. M. Razavi, and R. Trasarti, “Finding roles of players in football using automatic particle swarm optimization-clustering algorithm,” Big Data, vol. 7, no. 1, pp. 35–56, 2019. doi:10.1089/big.2018.0069.

[38] G. Ermidis, M. B. Randers, P. Krustrup, and M. Mohr, “Technical demands across playing positions of the Asian Cup in male football,” International Journal of Performance Analysis in Sport, vol. 19, no. 4, pp. 530–542, 2019. doi:10.1080/24748668.2019.1632571.

[39] Q. Yi, H. Jia, H. Liu, and M. Á. Gómez, “Technical demands of different playing positions in the UEFA Champions League,” International Journal of Performance Analysis in Sport, vol. 18, no. 6, pp. 926–937, Nov. 2018, doi: 10.1080/24748668.2018.1528524.

[40] A. Kubayi, “Technical demands of the various playing positions in the qualifying matches for the European football championship,” International Journal of Performance Analysis in Sport, vol. 21, pp. 1–10, Mar. 2021, doi: 10.1080/24748668.2021.1901436.

[41] Y. Li, S. Zong, Y. Shen, Z. Pu, M.-Á. Gómez, and Y. Cui, “Characterizing player’s playing styles based on player vectors for each playing position in the Chinese Football Super League,” Journal of Sports Sciences, vol. 40, no. 14, pp. 1629–1640, 2022, doi: 10.1080/02640414.2022.2096771.

[42] M. Konefał, P. Chmura, T. Zając, J. Chmura, E. Kowalczuk, and M. Andrzejewski, “Evolution of technical activity in various playing positions, in relation to match outcomes in professional soccer,” Biol Sport, vol. 36, no. 2, pp. 181–189, Jun. 2019, doi: 10.5114/biolsport.2019.83958.

[43] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, “Actions speak louder than goals,” Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining, 2019. doi:10.1145/3292500.3330758.

[44] B. Aalbers and J. Van Haaren, “Distinguishing Between Roles of Football Players in Play-by-play Match Event Data.” arXiv, Sep. 13, 2018. doi: 10.48550/arXiv.1809.05173.

[45] S. Ghar, S. Patil, and V. Arunachalam, “Data Driven football scouting assistance with simulated player performance extrapolation,” in 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Dec. 2021, pp. 1160–1167. doi: 10.1109/ICMLA52953.2021.00189.

[46] “Elite Player Performance Plan (EPPP),” The PFSA, https://thepfsa.co.uk/eppp/ (accessed Jan. 11, 2023).

[47] A. García-Aliaga, M. Marquina, J. Coterón, A. Rodríguez-González, and S. Luengo-Sánchez, “In-game behaviour analysis of football players using machine learning techniques based on player statistics,” International Journal of Sports Science & Coaching, vol. 16, no. 1, pp. 148–157, Feb. 2021, doi: 10.1177/1747954120959762.

[48] U. Kalenderoğlu, “Football player profiling using opta match event data: hierarchical clustering,” Opta maç verisi kullanarak futbolcu profillme: hiyerarşik kümeleme, 2019, Accessed: Jan. 07, 2023. [Online]. Available: https://openaccess.mef.edu.tr/xmlui/handle/20.500.11779/1214.

[49] T. Decroos and J. Davis, “Player Vectors: Characterizing Soccer Players’ Playing Style from Match Event Streams,” 2020, pp. 569–584. doi: 10.1007/978-3-030-46133-1_34.

[50] C. Carling, T. Reilly, and A. M. Williams, Performance Assessment for Field Sports. London: Routledge, 2010.

[51] H. Sarmento et al., “Influence of tactical and situational variables on offensive sequences during elite football matches,” Journal of Strength and Conditioning Research, vol. 32, no. 8, pp. 2331–2339, 2018. doi:10.1519/jsc.0000000000002147.

[52] D. L. Alves et al., “What variables can differentiate winning and losing teams in the group and final stages of the 2018 FIFA World Cup?,” International Journal of Performance Analysis in Sport, vol. 19, no. 2, pp. 248–257, 2019. doi:10.1080/24748668.2019.1593096.

[53] J. Fernandez, and L. Bornn, “Wide Open Spaces: A statistical technique for measuring space creation in professional soccer,” 2018.

[54] Network-based measures for predicting the outcomes of football games, https://www.semanticscholar.org/paper/Network-based-Measures-for-Predicting-the-Outcomes-Cintia-Rinzivillo/df400b609347705c2f15ad1bea415ab1ef82c18e (accessed Aug. 17, 2023).

[55] G. H. P. Toemen, “Player Performance Prediction in Football Using Machine Learning Techniques”.

[56] T. Laakso, K. Davids, P. Luhtanen, J. Liukkonen, and B. Travassos, “How football team composition constrains emergent individual and collective tactical behaviours: Effects of player roles in creating different landscapes for shared affordances in small-sided and conditioned games,” International Journal of Sports Science & Coaching, vol. 17, no. 2, pp. 346–354, Apr. 2022, doi: 10.1177/17479541211030076.

[57] J.-P. F. Rojas, “Premier League extends £5.1bn TV broadcast rights deal to 2025,” Sky News, https://news.sky.com/story/premier-league-extends-tv-broadcast-rights-deal-to-2025-12305022 (accessed May. 10, 2023).

[58] “Annual Review of Football Finance,” Deloitte United Kingdom, https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance-europe.html (accessed Feb. 7, 2023).

[59] G. Peters and D. Pacheco, “Betting the system: Using lineups to predict football scores.” arXiv, Jan. 17, 2023. doi: 10.48550/arXiv.2210.06327.

[60] O. Hubáček, G. Šourek, and F. Železný, “Learning to predict soccer results from relational data with gradient boosted trees,” Mach Learn, vol. 108, no. 1, pp. 29–47, Jan. 2019, doi: 10.1007/s10994-018-5704-6.

[61] J. Hucaljuk and A. Rakipović, "Predicting football scores using machine learning techniques," *2011 Proceedings of the 34th International Convention MIPRO*, Opatija, Croatia, 2011, pp. 1623-1627.

[62] A. Lindberg and D. Söderberg, “Comparison of Machine Learning Approaches Applied to Predicting Football Players Performance,” 2020, Accessed: Apr. 01, 2023. [Online]. Available: https://hdl.handle.net/20.500.12380/301745.

[63] R. P. Bunker and F. Thabtah, “A machine learning framework for sport result prediction,” Applied Computing and Informatics, vol. 15, no. 1, pp. 27–33, Jan. 2019, doi: 10.1016/j.aci.2017.09.005.

[64] J. Stübinger, B. Mangold, and J. Knoll, “Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics,” Applied Sciences, vol. 10, no. 1, Art. no. 1, Jan. 2020, doi: 10.3390/app10010046.

[65]B. Kavi, “OPTIMIZING THE LINE-UP IN SOCCER WITH REAL-LIFE DATA”.

[66] Mahrudinda, S. Supian, Subiyanto, and D. Chaerani, “Optimization of the best line-up in football using binary integer programming model,” International Journal of Global Operations Research, vol. 1, no. 3, pp. 114–122, 2020. doi:10.47194/ijgor.v1i3.45.

[67] J. Lago-Ballesteros and C. Lago-Peñas, “Performance in team sports: Identifying the keys to success in soccer,” Journal of Human Kinetics, vol. 25, no. 2010, pp. 85–91, 2010. doi:10.2478/v10078-010-0035-0.

[68] B. Milanovic, “Globalization and goals: does soccer show the way?,” Review of International Political Economy, vol. 12, no. 5, pp. 829–850, Dec. 2005, doi: 10.1080/09692290500339818.

[69] A. Heuer and O. Rubner, “How Does the Past of a Soccer Match Influence Its Future? Concepts and Statistical Analysis,” PLOS ONE, vol. 7, no. 11, p. e47678, Nov. 2012, doi: 10.1371/journal.pone.0047678.

[70] D. W. Johnson and S. Johnson, “The effects of attitude similarity, expectation of goal facilitation, and actual goal facilitation on interpersonal attraction,” Journal of Experimental Social Psychology, vol. 8, no. 3, pp. 197–206, May 1972, doi: 10.1016/S0022-1031(72)80001-5.

[71] H. Sarmento et al., “Patterns of play in the counterattack of Elite Football Teams - a mixed method approach,” International Journal of Performance Analysis in Sport, vol. 14, no. 2, pp. 411–427, 2014. doi:10.1080/24748668.2014.11868731.

[72] “Football statistics and history,” FBref.com, https://fbref.com/en/ (accessed Aug. 17, 2023). (visited on 2020-11-12).

[73] I. Brace, Questionnaire Design: How To Plan, Structure And Write Survey Material for Effective Market Research. 2004.

[74] J. A. Gliem and R. R. Gliem, “Calculating, Interpreting, And Reporting Cronbach’s Alpha Reliability Coefficient For Likert-Type Scales,” 2003, Accessed: Aug. 18, 2023. [Online]. Available: https://hdl.handle.net/1805/344.

[75] J. M. Bland and D. G. Altman, “Statistics notes: Cronbach’s alpha,” BMJ, vol. 314, no. 7080, pp. 572–572, 1997. doi:10.1136/bmj.314.7080.572.

[76] M. Tavakol and R. Dennick, “Making sense of Cronbach’s alpha,” Int J Med Educ, vol. 2, pp. 53–55, Jun. 2011, doi: 10.5116/ijme.4dfb.8dfd.

[77] A. Ly, M. Marsman, and E.-J. Wagenmakers, “Analytic posteriors for Pearson’s correlation coefficient,” Statistica Neerlandica, vol. 72, no. 1, pp. 4–13, 2018, doi: 10.1111/stan.12111.

[78] J. P. WEIR, “Quantifying test-retest reliability using the intraclass correlation coefficient and the sem,” Journal of Strength and Conditioning Research, vol. 19, no. 1, pp. 231–240, 2005. doi:10.1519/00124278-200502000-00038.

[79] W.-B. Xie, Y.-L. Lee, C. Wang, D.-B. Chen, and T. Zhou, “Hierarchical clustering supported by reciprocal nearest neighbors,” Information Sciences, vol. 527, pp. 279–292, 2020. doi:10.1016/j.ins.2020.04.016.

[80] A. Modibo Sidibé, X. Lin, and S. Koné, “Assessing groundwater mineralization process, quality, and isotopic recharge origin in the sahel region in Africa,” Water, vol. 11, no. 4, p. 789, 2019. doi:10.3390/w11040789.

[81] H. Song, B. Park, H. Park, and W. M. Shim, “Cognitive and Neural State Dynamics of narrative comprehension,” The Journal of Neuroscience, vol. 41, no. 43, pp. 8972–8990, 2021. doi:10.1523/jneurosci.0037-21.2021.

[82] Y. Gu et al., “Abnormal dynamic functional connectivity in alzheimer’s disease,” CNS Neuroscience &amp; Therapeutics, vol. 26, no. 9, pp. 962–971, 2020. doi:10.1111/cns.13387.

[83] D. Cai, X. He, X. Wu, and J. Han, “Non-negative matrix factorization on manifold,” 2008 Eighth IEEE International Conference on Data Mining, 2008. doi:10.1109/icdm.2008.57.

[84] D. Lee and H. S. Seung, “Algorithms for Non-negative Matrix Factorization,” in Advances in Neural Information Processing Systems, MIT Press, 2000. Accessed: Aug. 18, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html

[85] B. R. de Oliveira et al., “Selection of soybean genotypes under drought and saline stress conditions using Manhattan distance and Topsis,” Plants, vol. 11, no. 21, p. 2827, 2022. doi:10.3390/plants11212827.

[86] A. Yiannakos and V. Armatas, “Evaluation of the goal scoring patterns in European Championship in Portugal 2004.,” International Journal of Performance Analysis in Sport, vol. 6, no. 1, pp. 178–188, Jun. 2006, doi: 10.1080/24748668.2006.11868366.

[87] C. Carling, A. M. Williams, and T. Reilly, Handbook of Soccer Match Analysis: A Systematic Approach to Improving Performance. London: Routledge, 2008.

[88] J. Fernandez-Navarro, L. Fradua, A. Zubillaga, P. R. Ford, and A. P. McRobert, “Attacking and defensive styles of play in soccer: analysis of Spanish and English elite teams,” J Sports Sci, vol. 34, no. 24, pp. 2195–2204, Dec. 2016, doi: 10.1080/02640414.2016.1169309.

[89] M.-Á. Gómez, M. Mitrotasios, V. Armatas, and C. Lago-Peñas, “Analysis of playing styles according to team quality and match location in Greek professional soccer,” International Journal of Performance Analysis in Sport, vol. 18, no. 6, pp. 986–997, Nov. 2018, doi: 10.1080/24748668.2018.1539382.

[90] L. Kong, T. Zhang, C. Zhou, M.-A. Gomez, Y. Hu, and S. Zhang, “The evaluation of playing styles integrating with contextual variables in professional soccer,” Front Psychol, vol. 13, p. 1002566, 2022, doi: 10.3389/fpsyg.2022.1002566.

[91] J. Castellano and M. Pic, “Identification and Preference of Game Styles in LaLiga Associated with Match Outcomes,” International Journal of Environmental Research and Public Health, vol. 16, no. 24, Art. no. 24, Jan. 2019, doi: 10.3390/ijerph16245090.

[92] A. Hewitt, G. Greenham, and K. Norton, “Game style in soccer: what is it and can we quantify it?,” International Journal of Performance Analysis in Sport, vol. 16, no. 1, pp. 355–372, Apr. 2016, doi: 10.1080/24748668.2016.11868892.

[93] M. Konefał, P. Chmura, T. Zając, J. Chmura, E. Kowalczuk, and M. Andrzejewski, “Evolution of technical activity in various playing positions, in relation to match outcomes in professional soccer,” Biol Sport, vol. 36, no. 2, pp. 181–189, Jun. 2019, doi: 10.5114/biolsport.2019.83958.

[94] R. Rein and D. Memmert, “Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science,” Springerplus, vol. 5, no. 1, p. 1410, 2016, doi: 10.1186/s40064-016-3108-2.

[95] R. Gauriot and L. Page, “Psychological momentum in contests: The case of scoring before half-time in football,” Journal of Economic Behavior & Organization, vol. 149, pp. 137–168, May 2018, doi: 10.1016/j.jebo.2018.02.015.

[96] H. Kristjánsdóttir, K. R. Jóhannsdóttir, M. Pic, and J. M. Saavedra, “Psychological characteristics in women football players: Skills, mental toughness, and anxiety,” Scandinavian Journal of Psychology, vol. 60, no. 6, pp. 609–615, 2019. doi:10.1111/sjop.12571.

[97] C. Peñas and D. A, “Ball Possession Strategies in Elite Soccer According to the Evolution of the Match-Score: the Influence of Situational Variables,” Journal of Human Kinetics, vol. 25, pp. 93–100, Jan. 2010, doi: 10.2478/v10078-010-0036-z.

[98] P. Nalepka et al., “Human Social Motor Solutions for human–machine interaction in dynamical task contexts,” Proceedings of the National Academy of Sciences, vol. 116, no. 4, pp. 1437–1446, 2019. doi:10.1073/pnas.1813164116.

[99] H. Wagner, J. Pfusterschmied, S. P. Von Duvillard, and E. Müller, “Skill-dependent proximal-to-distal sequence in team-handball throwing,” Journal of Sports Sciences, vol. 30, no. 1, pp. 21–29, Jan. 2012, doi: 10.1080/02640414.2011.617773.

[100] J. Miñano-Espin, L. Casáis, C. Lago-Peñas, and M. Á. Gómez-Ruano, “High speed running and sprinting profiles of elite soccer players,” Journal of Human Kinetics, vol. 58, no. 1, pp. 169–176, 2017. doi:10.1515/hukin-2017-0086.

[101] B. Popp, C. C. Germelmann, and B. Jung, “We love to hate them! Social media-based anti-brand communities in professional football,” International Journal of Sports Marketing and Sponsorship, vol. 17, no. 4, pp. 349–367, Jan. 2016, doi: 10.1108/IJSMS-11-2016-018.

[102] M. Mohr, L. Nybo, J. Grantham, and S. Racinais, “Physiological Responses and Physical Performance during Football in the Heat,” PLOS ONE, vol. 7, no. 6, p. e39202, Jun. 2012, doi: 10.1371/journal.pone.0039202.

[103] V. Mougios, “Reference intervals for serum creatine kinase in athletes,” British Journal of Sports Medicine, vol. 41, no. 10, pp. 674–678, Oct. 2007, doi: 10.1136/bjsm.2006.034041.

[104] P. Chmura et al., “Is there meaningful influence from situational and environmental factors on the physical and technical activity of elite football players? Evidence from the data of 5 consecutive seasons of the German Bundesliga,” PLOS ONE, vol. 16, no. 3, p. e0247771, Mar. 2021, doi: 10.1371/journal.pone.0247771.

[105] “Local weather forecast, news and Conditions,” Weather Underground, https://www.wunderground.com/ (accessed Feb. 18, 2023).

[106] K. P. Sinaga and M.-S. Yang, “Unsupervised K-means clustering algorithm,” IEEE Access, vol. 8, pp. 80716–80727, 2020. doi:10.1109/access.2020.2988796

[107] M. Ahmed, R. Seraj, and S. M. Islam, “The K-Means Algorithm: A comprehensive survey and performance evaluation,” Electronics, vol. 9, no. 8, p. 1295, 2020. doi:10.3390/electronics9081295.

[108] L. Garicano, I. Palacios-Huerta, and C. Prendergast, “Favoritism under Social Pressure,” Review of Economics and Statistics, vol. 87, no. 2, pp. 208–216, 2005. doi:10.1162/0034653053970267.

[109]B. BURAIMO, R. SIMMONS, and M. MACIASZCZYK, “Favoritism and referee bias in European soccer: Evidence from the Spanish League and the UEFA Champions League,” Contemporary Economic Policy, vol. 30, no. 3, pp. 329–343, 2011. doi:10.1111/j.1465-7287.2011.00295.x

[110] P. Pettersson-Lidbom and M. Priks, “Behavior under social pressure: Empty Italian stadiums and referee bias,” Economics Letters, vol. 108, no. 2, pp. 212–214, 2010. doi:10.1016/j.econlet.2010.04.023

[111] J. Mallo, P. G. Frutos, D. Juárez, and E. Navarro, “Effect of positioning on the accuracy of decision making of association football top-class referees and assistant referees during competitive matches,” Journal of Sports Sciences, vol. 30, no. 13, pp. 1437–1445, 2012. doi:10.1080/02640414.2012.711485.

[112] M. J. Dixon and S. G. Coles, “Modelling association football scores and inefficiencies in the football betting market,” Journal of the Royal Statistical Society Series C: Applied Statistics, vol. 46, no. 2, pp. 265–280, 1997. doi:10.1111/1467-9876.00065

[113] J. Goddard and I. Asimakopoulos, “Forecasting football results and the efficiency of fixed-odds betting,” Journal of Forecasting, vol. 23, no. 1, pp. 51–66, 2004. doi:10.1002/for.877

[114] Y. Zhao and J. Xiao, “An adiabatic method to train binarized artificial neural networks,” Sci Rep, vol. 11, no. 1, Art. no. 1, Oct. 2021, doi: 10.1038/s41598-021-99191-2.

[115] M. W. Gardner and S. R. Dorling, “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences,” Atmospheric Environment, vol. 32, no. 14, pp. 2627–2636, Aug. 1998, doi: 10.1016/S1352-2310(97)00447-0.

[116] T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015. doi:10.18653/v1/d15-1166

[117] A. Natekin and A. Knoll, “Gradient Boosting Machines, a tutorial,” Frontiers in Neurorobotics, vol. 7, 2013. doi:10.3389/fnbot.2013.00021

[118] F. Alzamzami, M. Hoda, and A. El Saddik, “Light gradient boosting machine for general sentiment classification on Short texts: A comparative evaluation,” IEEE Access, vol. 8, pp. 101840–101858, 2020. doi:10.1109/access.2020.2997330

[119] A. V. Dorogush, V. Ershov, and A. Gulin, “CatBoost: gradient boosting with categorical features support.” arXiv, Oct. 24, 2018. doi: 10.48550/arXiv.1810.11363.

[120] R. Ghawi and J. Pfeffer, “Efficient hyperparameter tuning with grid search for text categorization using KNN approach with BM25 similarity,” Open Computer Science, vol. 9, no. 1, pp. 160–180, 2019. doi:10.1515/comp-2019-0011

[121] G. Budak, İ. Kara, Yusuf Tansel, and R. Kasımbeyli, “New mathematical models for team formation of sports clubs before the match,” Central European Journal of Operations Research, vol. 27, no. 1, pp. 93–109, 2017. doi:10.1007/s10100-017-0491-x

[122] B. H. Boon and G. Sierksma, “Team formation: Matching quality supply and quality demand,” European Journal of Operational Research, vol. 148, no. 2, pp. 277–292, 2003. doi:10.1016/s0377-2217(02)00684-7.

[123] T. Hastie, R. Tibshirani, and J. Friedman, “The elements of Statistical Learning,” Springer Series in Statistics, 2009. doi:10.1007/978-0-387-84858-7

[124] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. doi:10.3115/v1/d14-1162

[125] M. Chen, M. Gong, and X. Li, “Feature weighted non-negative matrix factorization,” IEEE Transactions on Cybernetics, vol. 53, no. 2, pp. 1093–1105, 2023. doi:10.1109/tcyb.2021.3100067

[126] V. Cicirello, “Kendall Tau sequence distance: Extending Kendall Tau from ranks to sequences,” EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, vol. 7, no. 23, p. 163925, 2020. doi:10.4108/eai.13-7-2018.163925

[127] “Weights & Biases – developer tools for ML,” Weights & Biases – Developer tools for ML, https://wandb.ai/site (accessed Aug. 18, 2023).

APPENDICES

APPENDIX A

SURVEY QUESTIONS

Descriptive Questions

Age
Educational status
Level of experience

Parameters For Player Positions And Team Playing Characteristics

Number of total actions
Number of goals
Number of shoot and percentage of on target
Xg per shoot
Number of goal chances and conversion rate
Number of passes and success rate
Number of forward passes and success rate
Number of passes to the third zone and success rate
Number of passes into the opponent's penalty area and success percentage
Number of passes from set-pieces and success rate
Number of long passes and success rate
Number of key passes and success rate
Number of assists
Number of shot passes and expected (potential) assists
Number of crosses and success rate
Total of passing distances
Total number of meetings with pas
Number of forward pass meetings
Number of touches in the opponent's penalty area
Number of offsides
Number and success rate of dribbling forwards
Total dribbling number and success rate
Number of offensive dual tackles and percentage of success
Number of foul exposures
Total number of turnovers
Number of turnovers in the opponent half
Number of turnovers in own half

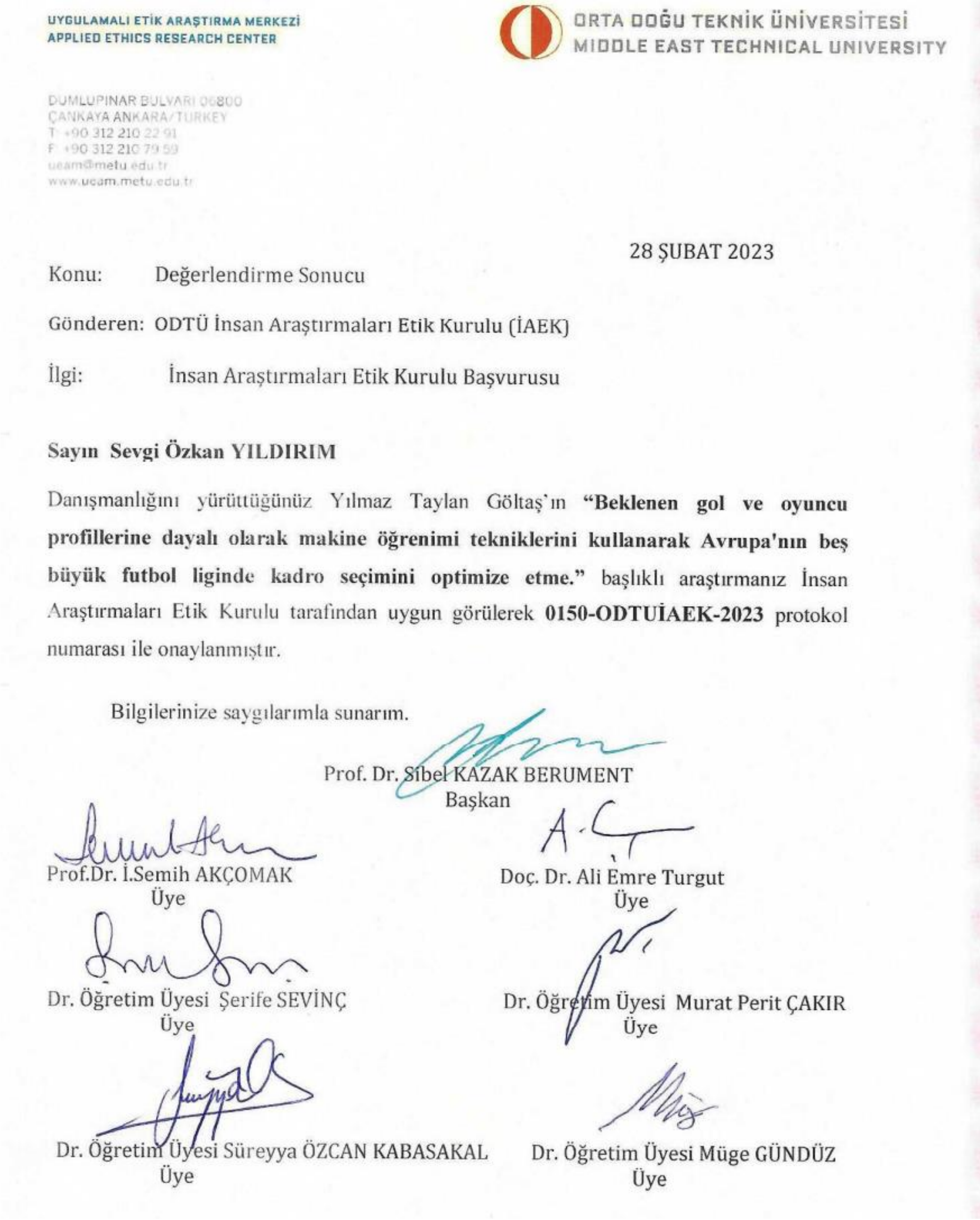
Number of turnovers after tackles
Total number of ball winnings
Winning the ball in the opponent's half
Number of turnovers in own half
Number of challenges and percentage of success
Number of air ball challenges and success percentage
Number of challenges for loose ball and percentage of success
Number of dribbling interventions and success percentage
Number of fouls committed
Number of forcing the opponent to mıstake
Number of shot blocking
Number of interceptions (pass interception)
Number of clearances

Parameters For Team Playing Characteristics

Ball possession (percentage, duration, number, ball possessıon per attack)
Expected goal (xg)
Net expected goal difference
Average expected goal of opponent shots
Total number of team presses and success percentage
Number of team presses in the opponent half and success percentage
Number and success percentage of team presses in own field
Average number of passes made by the opponent until the ball is won
Number of plays under press and without press
Number of set attacks and shot conversions percentage
Number of fast break attacks and percentage of conversions to shots
Number of set-pieces and success percentage

APPENDIX B

Ethics Committee Approval Letter



UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
ueam@metu.edu.tr
www.ueam.metu.edu.tr

28 ŞUBAT 2023

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Sevgi Özkan YILDIRIM

Danışmanlığını yürüttüğünüz Yılmaz Taylan Göltaş'ın **"Beklenen gol ve oyuncu profillerine dayalı olarak makine öğrenimi tekniklerini kullanarak Avrupa'nın beş büyük futbol liginde kadro seçimini optimize etme."** başlıklı araştırmanız İnsan Araştırmaları Etik Kurulu tarafından uygun görülerek **0150-ODTUİAEK-2023** protokol numarası ile onaylanmıştır.

Bilgilerinize saygılarımla sunarım.

Prof. Dr. Sibel KAZAK BERUMENT
Başkan

Prof.Dr. İ.Semih AKÇOMAK
Üye

Doç. Dr. Ali Emre Turgut
Üye

Dr. Öğretim Üyesi Şerife SEVİNÇ
Üye

Dr. Öğretim Üyesi Murat Perit ÇAKIR
Üye

Dr. Öğretim Üyesi Süreyya ÖZCAN KABASAKAL
Üye

Dr. Öğretim Üyesi Müge GÜNDÜZ
Üye

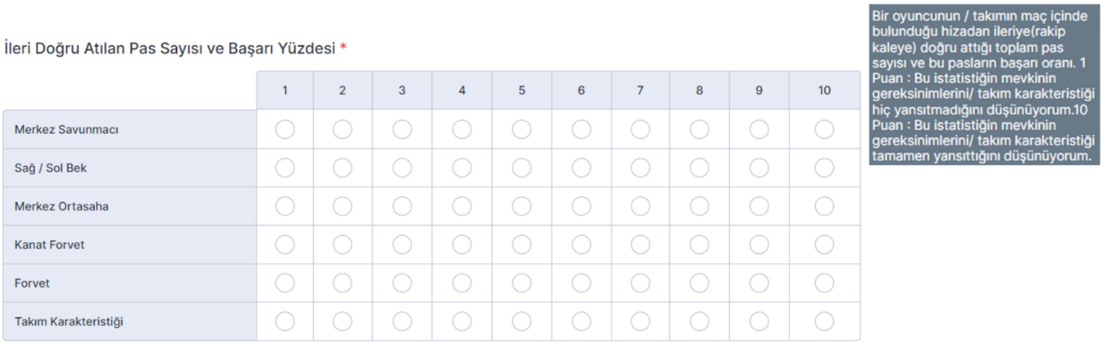
APPENDIX C

Examples of Survey Questions

Yaşınız

○ 30 ve altı ○ 31-40

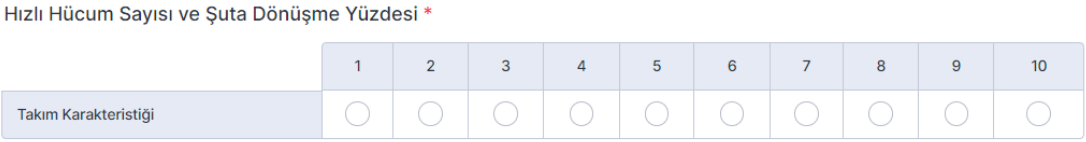
○ 41-50 ○ 51 ve üstü



İleri Doğru Atılan Pas Sayısı ve Başarı Yüzdesi *

	1	2	3	4	5	6	7	8	9	10
Merkez Savunmacı	○	○	○	○	○	○	○	○	○	○
Sağ / Sol Bek	○	○	○	○	○	○	○	○	○	○
Merkez Ortasaha	○	○	○	○	○	○	○	○	○	○
Kanat Forvet	○	○	○	○	○	○	○	○	○	○
Forvet	○	○	○	○	○	○	○	○	○	○
Takım Karakteristiği	○	○	○	○	○	○	○	○	○	○

Bir oyuncunun / takımın maç içinde bulunduğu hizadan ileriye(rakip kaleye) doğru attığı toplam pas sayısı ve bu pasların başarı oranı. 1 Puan : Bu istatistiğin mevkinin gereksinimlerini/ takım karakteristiği hiç yansıtmadığını düşünüyorum.10 Puan : Bu istatistiğin mevkinin gereksinimlerini/ takım karakteristiği tamamen yansıttığını düşünüyorum.



Hızlı Hücum Sayısı ve Şuta Dönüşme Yüzdesi *

	1	2	3	4	5	6	7	8	9	10
Takım Karakteristiği	○	○	○	○	○	○	○	○	○	○