MGI

Mestrado em Gestão de Informação
Master Program in Information Management

FOOTBALL ANALYTICS: A LITERATURE ANALYSIS FROM 2010 TO 2020

Leonardo Mendes Serra Fontanive

Dissertation presented as the partial requirement for obtaining a Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

FOOTBALL ANALYTICS: A LITERATURE ANALYSIS FROM 2010 TO 2020

by

Leonardo Mendes Serra Fontanive

Dissertation presented as the partial requirement for obtaining a Master's degree in Information Management, Specialization in in Knowledge Management and Business Intelligence

Advisor: Mauro Castelli

January 2021

ABSTRACT

The overall goal for the current study is to present a literature review of analytics, precisely machine learning (ML) reference authors in terms of methods and applicable scopes of study, in football where is a field that historically there are empirical decisions and the usage of analytics has been growing intensely. The research aims to list relevant academic contributions published between 2010 and 2020, performing a comparable picture per authors across the following subsets: player individual technical skills and team performance. Furthermore, the approach will provide a summary of studies for machine learning methods applied in football.

Such outcomes of this study would contribute to the discussion about football analytics. Regarding that these summaries can drive researchers to have a deep dive into the fields of interest straight to references preview studied in the thesis. Results indicate that football analytics has broadly vast opportunities in terms of research, regarding machine learning methods and a high potential to have a deep exploration of team and player perspective. This study can leverage and pavement new further in-depth and targeted investigation toward football analytics.

KEYWORDS

Machine Learning; Football; Sports; Analytics; Knowledge

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

Football has been improving the data-driven approach in terms of the decision. Following this trend, data researches have been emerging in the clubs and the interest in data is growing. Also, with the increase of the Internet of Everything Technologies in Football, new sources of data are leveraging the opportunities of analysis.

Combining data, technology, and football, (Américo, 2013) points out innovation introduced in football led to the creation of game analysis systems. Thus, these systems analyze players' moves collectively and individually to obtain statistics that support coaches' plans during the games. Fundamentally, the mentioned approach uses sophisticated techniques and algorithms to deal with important information about broad aspects of the game, such as studies of the aptitude and tactical capacity of the players and teams present in the game.

According to (De Silva, et al., 2018), it´s possible to apply data analytics from some specific football perspective. Also, it would be a key success to balance player indicators. Thus, for instance, tracking in football would enable monitoring and, consequently, would be possible to determine if an athlete is performing accordingly on matches and training sessions.

(Pires & Santos, 2018) presents a study which states that increasingly sports results often lie in the details that can be noticed with the use of technology or device that can make the difference. Also, nowadays in the big data era, sports are part of that, because, regarding there are large amounts of data collected that can be applied for analysis, and sports technology is in constant growth and development, as it has been seen the increase in science in the sports area.

Joining within the above findings, (Memmert, et al., 2011) concludes that many experts regard tactics as the factor which gets the least attention in the training process and the work is heavily empirical on those deficits not using a scientifically based analysis of football-specific group tactics.

Exploring the scenario presented in this study broadly, the study of (Mohammadi & Sorour, 2018) and (Rein & Memmert, 2016) illustrates also the effects on jobs and the workforce in general when the subject is Machine Learning, regarding that knowledge work automation addresses creative problem-solving. Furthermore, ML techniques are enablers for big data analytics and knowledge extraction.

(Novatchkov & Baca, 2013) summarizes, regarding gathering data in sports that it is possible to have an insight into significant characteristics. In one way these inputs can be applied for the development of intelligent methods adapted from conventional machine learning concepts, allowing an automatic assessment, and providing appropriate feedback. Thus, in practice, the implementation of such techniques could be crucial for the investigation of the quality of the execution, the assistance of athletes but also coaches.

Analyzing the researched scenario, in the literature, there is an expressive group of machine learning methods for football analytics. Also, among all the perspectives that involve the game, the context of

player individual technical skills and team performance seems to be findings that can have interesting results, regarding the lack of scientific decision support approach.

1.2. PROBLEM (RESEARCH QUESTION)/ GENERAL OBJECTIVE (MAIN GOAL)

Data analytics is like a storm that is gathering in football. It is one that will wash away all certainties and change the game we know. Football analytics has not progressed as far as it could. Every shared knowledge will support us to love more the game. Analytics in football is the newest frontier. It´s not just a matter of collecting data. You have to know what to do with them (Anderson & Sally, 2013).

Also, technology has been changing the data collection process in high-performance sport. So it increases the successful athletes and teams will be those who are supported by a strong analytical capability to create all the required analysis to create competitive advantage (Schulenkorf & Frawley, 2017).

Figuring out the background and problem identification, the research aims to present relevant academic contributions published between 2010 and 2020, performing a comparable picture per authors across the following subsets: player individual technical skills and team performance. Furthermore, the approach will provide a summary of studies for machine learning methods applied in football. It´s expected to reach this general objective, giving summaries and a basis for discussion toward football analytics in terms of methods and the referred subsets per authors.

1.3. SPECIFIC OBJECTIVES

We can divide the primary goal of this project into the following objectives, respecting the order of steps:

1. To investigate and study relevant academic contributions published between 2010 and 2020 toward machine learning methods applied for football analytics and the respective used fields: player individual technical skills and team performance;

2. To evaluate and discuss the most relevant ML models per authors to perform a summary;

3. To evaluate and discuss the most relevant authors per chosen fields to perform a summary;

4. To give future directions for the next steps using this study.

1.4. ASSUMPTIONS

The present research will underline references to build summaries about the subject of football analytics. The proposal is to target the relevant authors and drive a literature review, not a comparison between algorithms, regarding the accuracy and the weight of each subset approached in the football context.

1.5. STUDY RELEVANCE AND IMPORTANCE

Regarding relevance and importance, it can be highlighted that coaches can also take advantage of the data extracted from the players' performance. They are moving toward integrating analytics more dynamically. Decisions may use technology to support and justify them. Coaches can make tactical decisions, helping scouts decide what players their team should recruit as well. Coaches have access to live analytics of the players' physical condition and overall performance made accessible by the analytics department that is also working in real-time in the stands (Korte, 2014).

Joining the idea, analytics has influenced the tactics in professional baseball and basketball in recent years. Ultimately, it may have just as great an impact on football, which traditionally hasn't relied on statistics to figure out much of anything. In 2019, Liverpool, a recognized team in England that invest a lot in Analytics used data to big decisions and supported them to arrive at the last UEFA Champions League final (Schoenfeld, 2019).

Regarding the economic impact, the market for sporting events is worth $80 billion in 2014—with impressive growth projected for the foreseeable future. For a content industry wracked with uncertainty, sports analytics is a beacon of hope. On a sport-by-sport basis, growth occurred nearly across the board, but football remains the runaway leader. Football revenues increased from $25.1 billion in 2009 to $35.3 billion in 2013. (Collignon, 2019).

The work of (Morgulev, Azar, & Lidor, 2018) provides an introduction that underlines big data characteristics of sport as a uniquely authentic arena for exploring research ideas, an excellent source for analytics and specifically those concerning certain contexts of human behavior.

Additionally, the study of (Kumar, 2013), presents that if we search at the performance analytics literature related to football, we realize that most of the research is being done with few performance variables and is dependent on understanding the structure of the game. The possible interaction of multiple factors impacts the complexity of football match analysis. Thus, it makes it necessary to improve the contribution in terms of studies and cases.

1.6. METHODOLOGY

The design of the research will be descriptive. The first target of this research will have a consolidated literature review about football analytics. The second step will evaluate and discuss the most relevant ML models per authors to perform a summary, regarding the literature analysis. The third step will be a set of relevant authors per the following fields: player individual technical skills and team performance. At the end present the results and discussions with a conclusion and recommendation for further researches in the future.

2. LITERATURE REVIEW

2.1. FOOTBALL

2.1.1. Overall Concept

(Ali, 2011) approaches football as the premier audience sport in the world, due to its growing popularity, as well as the amount of financial interest in the game, it is one of the most extensively researched intermittent team sports. Highlighting, there are plenty of subject areas that have taken advantage of scientific knowledge gained from football including the natural and physical sciences, medicine, and social sciences.

Another important aspect to add on a timeline perspective until now, (Doidge, et al., 2019) shows that football has been facing an economic transformation over this time and there has been a significant change in how football clubs are managed, within an increasing focus on commercial and media growth. Also, it has been performing investments in stadiums and branding. Furthermore, sponsorship and contracts have been increasing exponentially to achieve global audience penetration. For instance, the Premier League and Champions League were formed in 1992 and these have acted as economic models that leverage the other leagues and turned it into a successful model and push the economic transformation.

Adding new components to conceptualize football, regarding (FIFA, 2019), the world governing body, since its birth, football has been part of our communities. It is more than a game, more than a sport, it is a way of life that we all embrace, regardless of nationality, creed, ethnicity, education, gender, or religion. And It is about supporting the growth and development of football by promoting the integrity and quality of the game for everyone, and not just for today, but for generations to come.

According to (Ali, 2011) and raising the bridge to arrive on the study subject, football is a complex sport, requiring the repetition of many disparate actions. For instance, there are several proofs of concepts that are currently being used such as assess the physical prowess of players, approaching simple running tests using monitor speed, agility proofs, and repeated sprint performance.

Joining what (Ali, 2011) concludes, (De Silva, et al., 2018) presents that performance management of top football players is a complex system involving enhancement of physical performance, skill-based training, tactical training, minimization of injury risk, and psychological support. Managing practice is vital to allowing players to perform at an optimal level throughout a play season's length.

(Constantinou & Fenton, 2017) agrees with (Ali, 2011) and (FIFA, 2019) that Football is the most popular sport in the world and it leverages the inspiration of several researchers to use football activities as a real-world application field to test various statistical, probabilistic, and machine learning techniques.

Continuing to explore the complexness, referring (Qing, et al., 2020), football is influenced by many factors such as technical, tactical, mental, and physiological, however as the matches have a high-level complexity and dynamic behavior, some other aspects needed to be addressed as situational variables as match location, team quality, quality of opposition and match outcome. Also, it´s important to underline other game nature dynamics as interactions between players and positions.

2.1.2. Player individual technical skills: the complexity, samples, and limitations

The football player's technical skills are directly connected with data and performance, (Caya & Bourdon, 2016) proposes that Business Intelligence & Analytics techniques can represent individual athletes well in their pursuit of positive achievements. Also, individual athletes are anxious to leverage their athletic performance in their respective sport and aspire to be good at what their sport demands in terms of physical and competitive accomplishments.

Continuing with the topic and setting practical samples, relating to football player's technical skills, data, and performance, the study of (Spearman & Basye, 2017) presents a model for ball control in football based on the concepts of how long it takes a player to reach the ball (time-to-intercept) and how long it takes a player to control the ball (time-to-control). Thus, players would keep the advantage in understanding this physics-Based modeling of pass probabilities translated and applied to the field, regarding that these metrics are constructed at the per-player level.

Additionally, according to the figure below from (Pappalardo, Cintia, Pedreschi, Giannotti, & Barabási, 2017), there is another sample of events produced by a player during a match which would be considerable in terms of player individual technical skills. The study uses as a source a game followed by reporters from three sports newspapers, then they assign an individual player rating according to personal interpretation of each player´s performance.
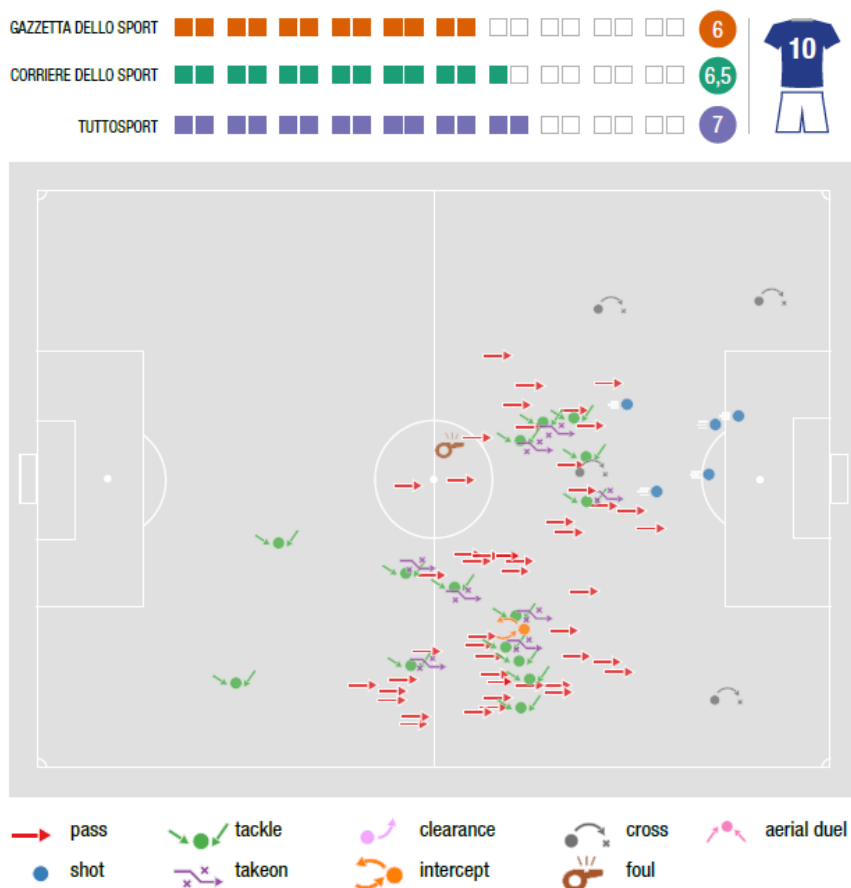
Figure 1 - Technical performance and soccer player ratings (Pappalardo, Cintia, Pedreschi, Giannotti, & Barabási, 2017)

Furthermore, the research of (Pappalardo, Cintia, Pedreschi, Giannotti, & Barabási, 2017) points out one specific attention needed before analysis which is that the human evaluation process has a limitation of features which attract their attention and then construct the evaluation. Thus, an important step is to understand how the human evaluation process can be leveraged with the support of data science and artificial intelligence. In the following figure, charts indicate the importance of every attribute, normalized in the range [0; 1], to the human rating process for football typical positions as Goalkeepers (a), Defenders (c), Midfielders (d), and Forwards (b). Hence, taking advantage of machine learning models, as the plots indicate, most of the features have a negligible influence on the human judge's evaluation process.

Figure 2 – Importance of technical and contextual features to human rating process (Pappalardo, Cintia, Pedreschi, Giannotti, & Barabási, 2017)

Continuing to approach the importance of context, the study of (Aquino, Alves, Fuini, & Garganta, 2017) identifies the lack of definition and classification of the skill-related variables. Also, it was detected two additional limitations: the contextualization of the sample omitted and the influence of match situational variables (e.g. location, quality of opponent, status); and the absence of representative task design to measure skill-related performance. The following figure shows definitions proposed by the research according to the literature.

Skill-related performance		Definitions
Offensives	Goal attempts	Directing the ball into the opposing goal, with any part of the body allowed by the game rules, in an attempt to make it cross the line between the goalposts. Successful: results in a goal. Unsuccessful: does not result in a goal
	Passes	Transferring the ball from one player to another of the same team, using the body parts allowed by the game rules (long passes: > 37 m; short passes: ≤ 37 m). Successful: the team continues with the ball possession. Unsuccessful: the team does not continue with the ball possession
	Carry/protect the ball	(i) Progressing in any direction when carrying the ball; (ii) when carrying the ball, making one or more contacts with it without progressing on the field, but in order to protect it from the opponent. Successful: the team continues with the ball possession. Unsuccessful: the team does not continue with the ball possession
	Received balls (oriented / not oriented)	Receiving and controlling the ball from a teammate, in a way: (i) oriented, allowing to continue sequentially the other actions with the ball possession; (ii) not oriented, not allowing to continue sequentially the other actions. Successful: the team continues with the ball possession. Unsuccessful: the team does not continue with the ball possession
	Dodge/feint	When carrying the ball, overtaking the direct opponent in an attempt to: (i) overtake and promote an offensive advantage for the team; (ii) create an imbalance for the defence opponent. Successful: the team maintains the ball possession. Unsuccessful: the team does not maintain the ball possession
Defensives	Crosses	Action of the player who is to take the ball from the player in possession of the ball impeding him to continue his action. Successful: the team wins the possession of the ball. Unsuccessful: the team does not win the possession of the ball
	Interceptions	Anticipation of the pass/goal attempt trajectory performed by the opponent, not allowing the ball to reach the destination when transmitted by the opponent. Successful: the team takes ball possession. Unsuccessful: the team does not take ball possession

Figure 3 - Proposed definitions for skill-related performance in soccer based on a literature search (Aquino, Alves, Fuini, & Garganta, 2017)

Moreover, (Aquino, Alves, Fuini, & Garganta, 2017) focuses that a fundamental task in sports science and performance analysis is to understand the relationship between skill acquisition and the development of players to achieve sports excellence. Hence, it is essential to develop theoretical principles to guide the concession of skill acquisition programs. The improvements for decision-making and regulation of action in dynamic environments, for instance in football, come out from the continuous performer-environment interactions.

2.1.3. Team Performance and the associated challenge

Regarding the game, the last topic approached the player technical skills in terms of complexity, samples, and limitations, there is another perspective that will be addressed in this study: the team as a separate entity for performance analysis.

Collaborating with the statement above, the study of (Pratas, Volossovitch, & Carita, 2018) refers that target-scoring trends have been analyzed from two different perspectives, according to research studies performing football match analysis: the static and the dynamic. Also, the inherent randomness of football makes the analysis even more impactful, and what makes a difference is perhaps not the data itself, but the ability to use this data to formulate a theory that explains how a team increases their chances of winning. Furthermore, some relevant performance metrics correlated with goal-scoring are underlined and may be appropriate in each context of the game but would likely be insignificant in another, depending on several factors related to the quality of the teams and the style of the game. Thus, Football analytics determines that performance indicators are relevant, and the problem is how the importance of different performance metrics varies depending on the context.

Continuing to explore the challenge of analysis in terms of team performance, according to (Lepschy, Woll, & Wäsche, 2018), despite the popularity of football and while reviews on performance indicators in football are available, none focuses solely on the identification of success factors. Additionally, it appeared that the most significant variables are efficiency (the number of goals divided by the number of shots), shots on goal, ball possession, pass accuracy/successful passes as well as the quality of opponent and match location.

Following again the research of (Pratas, Volossovitch, & Carita, 2018), the characteristics that can be approached and analyzed on goal scoring as key performance indicators:

Pass accuracy	Number of passes
Temporal Analysis	Duration of possession
Scoring Efficiency	Number os passes
Types of passes	Game situation
Zones in which possessions started	First and next goals
Playing style	Space-time coordination
Scoring efficiency	First goal
Areas from which goals were scored	

Table 1 - key performance indicators (Pratas, Volossovitch, & Carita, 2018)

The research of (Sarmento, Campanico, & Marcelino, 2014) shows an overview in terms of team analysis, it indicates relationships between patterns of physical and efficacy of game actions (involvements with the ball, successful, passes, dribbling, shots, and shots on target). Also, high-intensity activity patterns can be a key success factor for team performance and players of more

successful teams covered greater total distances with the ball, and at very high-intensity running, had a high average of goals for total shots on target, performed more actions with the ball, higher number of passes, tackles, dribbling, and shots on target when compared with less successful teams.

2.1.4. Other Fields

Beyond the scope of this study, football has a broad scope under different fields, regarding (Morgulev, Azar, & Lidor, 2018), sports betting market is one of largest sports business sector where consumers and suppliers try to predict the results of future events correctly. Scientists are therefore continuing to develop a variety of models that are constructed using different methodologies for forecasting. Such models concentrate on predicting the outcomes of individual matches or the outcomes of the tournament. It has become an arena for making progress in computing and machine learning, with cutting-edge predictive analytics, due to the extremely competitive nature of the gambling industry.

The research of (Klyuchka, Cherednichenko, Vasylenko, & Yakovleva, 2015) aimed to find the most important factors that are not confidential information and can be easily determined before the start of the football match. It presents that forecasting rules are used to increase the accuracy of predicting the results of football matches by identifying the winning team based on data retrieved from results of previous games championship, adding substantial factors, to understand the influence of results.

The football codes are recurrent team sports with high-intensity action bursts that are intermix with low-intensity and rest events. A variety of pressures imposed on intermittent sporting team members contribute to temporary, acute, or chronic fatigue. Fatigue is dynamic and multifactorial and depends on various contextual factors such as physical ability, technical abilities, the role of play, training load, the importance of the game, and seasonal period. The number of competitive matches per season is often very high; thus, between training sessions and competition, athletes only have a short period to recover. There is evidence that too many matches can result in a lack of motivation and mental burn-out, a decline in physical and match results, and an increase in injuries. There is evidence that too many matches can result in a lack of motivation and mental burn-out, a decline in physical and match quality, and an increase in injuries. Recovery approaches are therefore required to relieve fatigue, recover efficiency, and reduce injury risk. (Clarke & Noon, 2019)

Such studies as (Morgulev, Azar, & Lidor, 2018), (Klyuchka, Cherednichenko, Vasylenko, & Yakovleva, 2015), and (Clarke & Noon, 2019) are outside the scope of this research, but It´s important to notice that, conceptually, there are other fields which can be combined to cope deeply team and individual performance, especially injury prevention according to (Clarke & Noon, 2019).

2.2.1. Concept

Moving forward with the football performance context in terms of team and individual player skills, it's important to connect the concepts of football analytics, (Babbar, 2019) shows that sports have been facing progress from just being a sport to the involvement of science in it. The referred study approaches Sports analytics as a combination of data collection, forecasting the game, and using tools and techniques to interpret the game strategy to improve a player's performance individually and for the team. Hence, it is expected that sports analysis will foster many new applications for end-users, sports coaches, and sports managers. Also, analytical goals in these applications include a comparison of results, prediction, and behavioral correlation of attributes between players and teams. Moreover, information can be either quantitative or qualitative and is usually collected from the athletes' biographical data, performance, medical reports, and scouting reports.

Following the rationale of (Babbar, 2019), football analytics can support reliable and systematic data enabling athletes and coaches to leverage their decisions. To put in place this kind of initiative, real-time systems can be used for finding key analysis points, capturing the position of ball and movement of players throughout the game, and combining it with advanced statistical algorithms and software would enable coaches, managers to alter their tactics to gain an upper hand on the competitor.

According to (Morgulev, Azar, & Lidor, 2018), sports analytics can be supported as a framework, historical data can be either quantitative or qualitative and these data are typically collected from multiple sport-relevant resources and the collected data are standardized, centralized, integrated, and analyzed using different metrics. Thus, it is assumed that a reliable and systematic analysis of the data will enable different stakeholders to strengthen their decision-making processes. A sports analytics framework is described in the following figure.
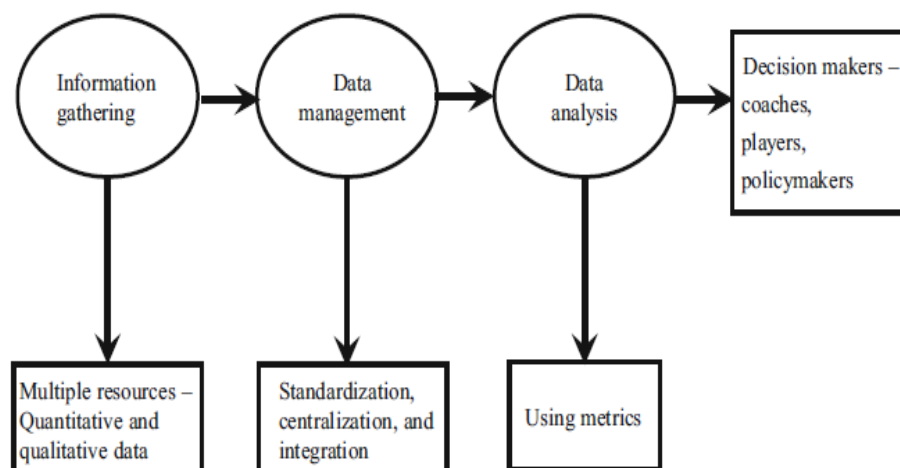
Figure 4 – A sports analytics framework (Morgulev, Azar, & Lidor, 2018)

Under this context and regard the study of (Morgulev, Azar, & Lidor, 2018), in football, teams in the English Premier League (EPL) became advanced in terms of performance analytics. Although, when

compared to basketball, for instance, assessing players' skills to score in football is slowed down by the low frequency of scoring events. Tactical factors, such as the number and length of possessions, passing sequences, and spatial analysis of the territory played are aggregated to optimize performance. Additionally, a specific example of how players and coaches may benefit from the assessment of large samples of events in football is the information combined with probability in the directions of penalty shots, based on the shooters' previous statistics provided by the analysts to the goalkeepers before critical matches.

As football is a competitive sport, an integrative framework was developed by (Caya & Bourdon, 2016) in which the potential value from Business Intelligence and Analytics project is tight to the actual focus of value where these investments are expected to happen. While this framework provides a very high-level representation of how and where this kind of project generates value in competitive sports, it conducts more accurate examples of value creation at each level of analysis (institutional, organizational, and individual levels). Within each "sub-model", specific detail about the nature of Business Intelligence and Analytics initiatives, along with particular conversion contingencies and measures of value created from those same initiatives.
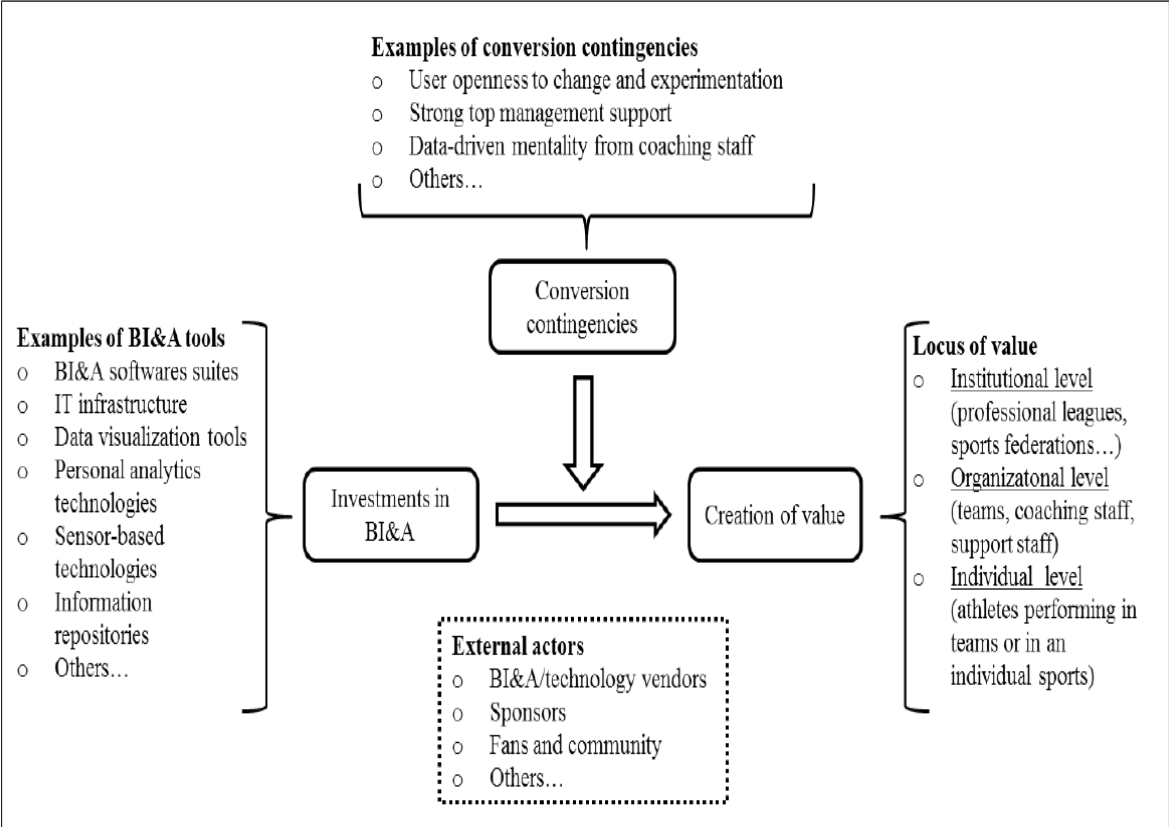
Figure 5 – Model of value creation from business intelligence & Analytics in competitive sports (Caya & Bourdon, 2016)

2.2.2. Use Cases

To make the concepts of football analytics tangible, this sub-section presents a compiled of samples that show the applicability of the main purpose of the section. There are seven studies, each one with a different perspective in the football analytics field, such as tactical efficacy, offensive

behaviors, defensive playing method, analysis of match situations, statistical indicators for improving player performance, data-driven framework to evaluate player performance, situations that finished in goal and distribution of the positions of the events.

The first sample is the study from (Borges, Garganta, Guilherme, & Jaime, 2019), it refers to tactical efficacy and offensive game processes adopted by Italian and Brazilian youth soccer players, approached the performance of the team with the following scope: 218 offensive actions selected from 28 matches, including 18 matches of the Italian team U-15 in dispute for the Italian championship, season 2015/2016, and 10 matches of the Brazilian team – 5 matches U-15 and 5 matches U-17 – in dispute for the national and state championship, season 2016. Additionally, Matches were randomly selected along the season.

The research of (Borges, Garganta, Guilherme, & Jaime, 2019) takes advantage of data through observational analysis using Match Vision Studio Premium®, software that enables the researcher to create a categorical matrix according to the variables to be analyzed.
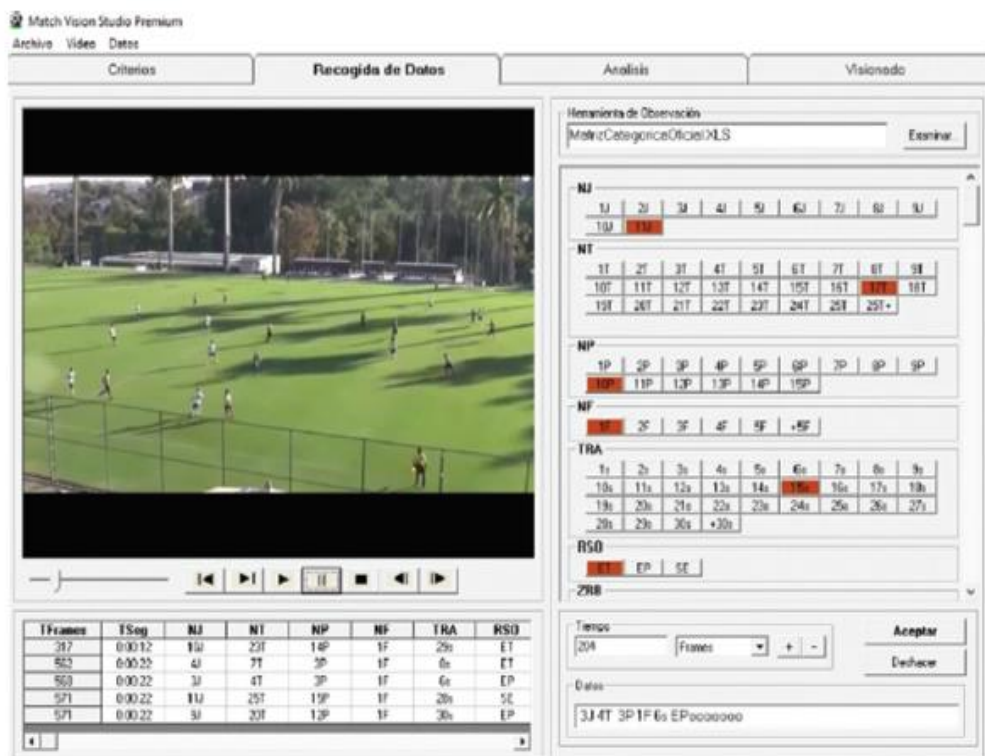
Figure 6 - Screen capture from Match Vision Studio Premium® software with the categorical matrix used in the study (Borges, Garganta, Guilherme, & Jaime, 2019)

(Borges, Garganta, Guilherme, & Jaime, 2019) concludes that all offensive sequences ended in shots according to the following variables: number of players involved, ball touches, passing, duration, corridor change. Also, defines offensive actions as three: a counter-attack, quick attack, and positional attack. Thus, with this context, the research suggests that all offensive methods adopted can be used to achieve success during a game of U-15 and U-17 soccer players.

The research of (Castelão, Garganta, Afonso, José, & Costa, 2015) targets to analyze offensive behaviors performed by six national football teams that were involved in the finals of the 2006 world cup and 2004 and 2008 Euro Cup.

The mentioned study of (Castelão, Garganta, Afonso, José, & Costa, 2015), supported by the software SIDS (Sequential Data Interchange Standard & GSEQ (Generalized Sequential Querier), uses sequential analysis by the lag method to verify the different offensive game patterns and analyzes 647 offensive game sequences. Also, with the bias and sample explored, it shows that top-performing football teams drive different patterns and methods of offensive play and yet be victorious. Furthermore, no patterns were found to be more effective than others, regarding any specific offensive behavior.

Continuing approach football analysis on a group perspective, (Frias, 2012) presents that changes in the defensive playing method influence the collective behavior of football teams, using Team AMS software application and Team Sense as the support tool of the study. Moreover, it aimed to understand the influence of specific performance constraints on the actions of teams, with the main goal was to analyze the influence of the defensive method (zone vs. man-to-man) in the collective performance of football teams. Then, the research analyzes two small-sided games played by two teams of 6 players (5 outfield players plus a goalkeeper) both using zone defense in the first experimental condition and man-to-man defense in the second one. Thus, the collective performance of teams was captured by 4 collective variables found: surface area, stretch index, length per with ratio, and teams' centers' distance.
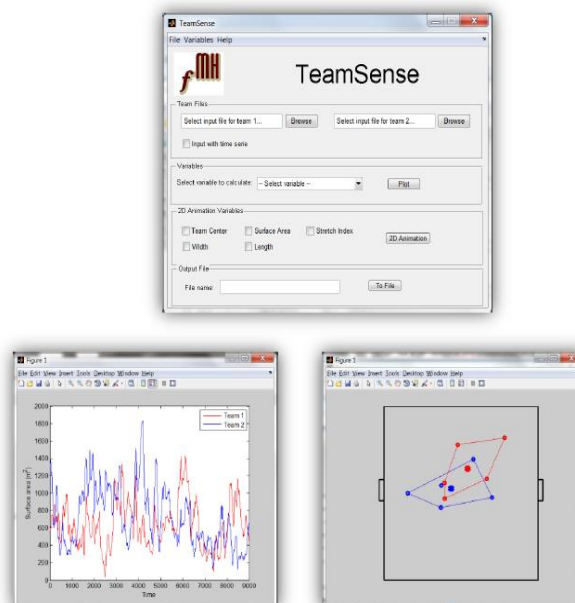
Figure 7 - TeamSense software application developed in Matlab with entry point view; an example of a time-plot variable output (e.g. surface area); an exemplar of photogram from a variable 2D video animation (Frias, 2012)

Regarding the study of (Frias, 2012), summarizes that the changes imposed in the collective behavior of a team by the adoption of a different defensive method can be less strong than the own differences between the two teams.

The last section approached one study that may drive a balance in terms of prioritization between defense tactical and the opponent analysis, also, in the perspective of improving approach, the paper from (Vanoye, Penna, Parra, & Díaz, 2017) proposes to use metrics or statistical indicators for leveraging the football performance. Besides the rating of the individual errors with negative points: Goals Shots Off Target, Not goals from direct free-kicks and indirect free-kicks, unsuccessful dribbles, caught opposition offside, unsuccessful shots free-kicks or indirect free-kicks, head Shots Off the target, shots off target, unsuccessful long /short passes, pass directions incorrectly, pass lengths incorrect, pass locations incorrect, duels lost on the offensive/defensive, aerial duels lost on the offensive/defensive, own goals, penalties conceded, defensive mistakes, fouls Committed, corner crosses / direct or indirect free-kicks conceded, Throw-ins conceded, yellow or red cards, substituted off, and others, which significantly affects the soccer performance of the team, to the metric called Motivation Index or lack of motivation.

To going further with this Index, the study of (Vanoye, Penna, Parra, & Díaz, 2017) takes advantage of a European football match to obtain the index of motivation and thereby determine the relationship of the index with the outcome of the match. In the meantime, the software NacSports supports the performance analysis, and indicators were distributed as individuals, according to figure 8, and as a team, regarding figure 9.
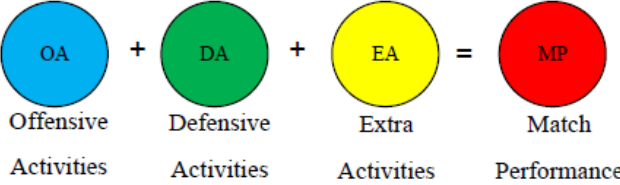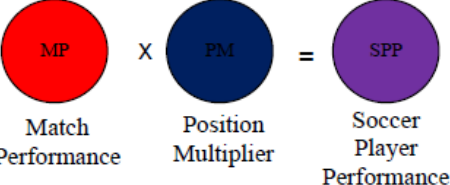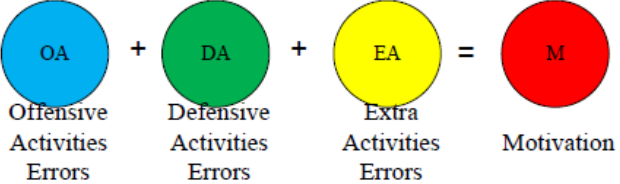
Figure 8 - statistical indicators for improving the soccer performance from an individual's perspective (Vanoye, Penna, Parra, & Díaz, 2017)
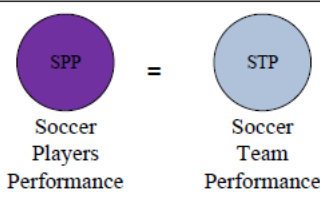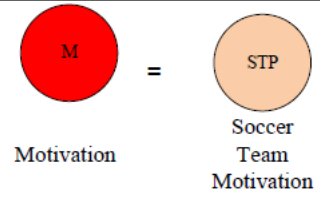
INDICATORS (TEAM)			PURPOSE
SUM SPP (Soccer Players Performance)	=	STP (Soccer Team Performance)	The soccer team performance is the sum of the soccer player's performance.
SUM M (Motivation)	=	STP (Soccer Team Motivation)	The soccer team motivations are the sum of soccer player's motivation.

Figure 9 - statistical indicators for improving the soccer performance from the team's perspective (Vanoye, Penna, Parra, & Díaz, 2017)

After the experimentation stage and results, (Vanoye, Penna, Parra, & Díaz, 2017) concludes that small individual errors affect the motivation state of the players, which affects the result of the match. Hence, indicators of motivation can be a key tool to be applied to make corrections during the match and avoid losing the match by negative values indicators of motivation.

Among the different supporting structures oriented to analysis in football, the work of (Pappalardo, Ferragina, Cintia, & Pedreschi, 2019) defines PlayeRank, a data-driven framework that offers a principled multi-dimensional and role-aware evaluation of the performance of football players. Moreover, it deployed a massive dataset of soccer-log, millions of match events on four seasons of 18 prominent soccer competitions.

The framework of (Pappalardo, Ferragina, Cintia, & Pedreschi, 2019) consists of 3 phases starting from a database of soccer-logs: rating phase, in charge of computation of the performance rating; ranking phase, PlayeRank assigns a player to a position according to a set of rules and based on players rating; learning phase, generates information used in the rating and the ranking phases performing two steps, weighting and role detector training.
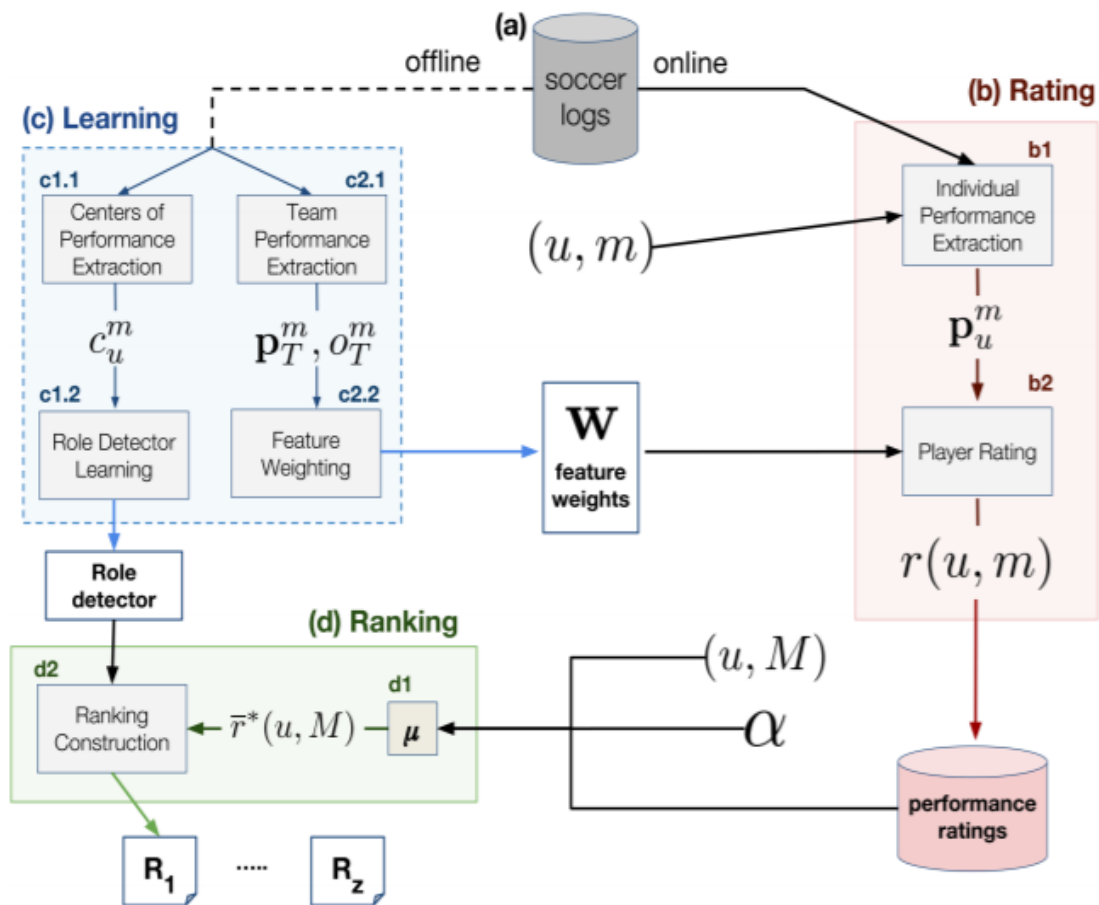
Figure 10 - Schema of the PlayeRank framework. Starting from a database of soccer-logs (a), it consists of three main phases. The learning phase (c) is an "offline" procedure: It must be executed at least once before the other phases since it generates information used in the other two phases, but then it can be updated separately. The rating (b) and the ranking phases (d) are online procedures, i.e., they are executed every time a new match is available in the database of soccer-logs (Pappalardo, Ferragina, Cintia, & Pedreschi, 2019)

The dataset of (Pappalardo, Ferragina, Cintia, & Pedreschi, 2019) covers a total of 64 soccer seasons, more than 31 million events and was provided by the company Wyscout, a leading company in the football industry that connects soccer professionals worldwide, supporting more than 50 soccer associations, and more than 1,000 professional clubs around the world. Following the table below, all the details will be provided.

competition	seasons	matches	events	players
Spanish first division	4	1,520	2,541,873	1,264
English first division	4	1,520	2,595,808	1,231
Italian first division	4	1,520	2,610,908	1499
German first division	4	1,124	2,075,483	1,042
French first division	4	1,520	2,592,708	1,288
Portuguese first division	4	1,124	1,720,393	1,227
Turkish first division	4	1,124	1,927,416	1,182
Greek first division	4	1,060	1,596,695	1,151
Austrian first division	4	720	1,162,696	593
Swiss first division	4	720	1,124,630	647
Russian first division	4	960	1,593,703	1,046
Dutch first division	4	1,248	2,021,164	1,177
Argentinian first division	4	1,538	2,450,170	1,870
Brazilian first division	4	1,437	2,326,690	1,790
European Champions League	3	653	995,363	3,577
Europa League	3	1,416	1,980,733	9,100
European Cup 2016	1	51	78,140	552
World Cup 2018	1	64	101,759	736
	64	19,619	31,496,332	(*)21,361

Table 2 - For each competition, described the corresponding geographic area, the total number of seasons, matches, events, and players. (*) 21,361 indicates the number of distinct players, as some players play with their teams in both national and continental/international competitions (Pappalardo, Ferragina, Cintia, & Pedreschi, 2019)

The research of (Pappalardo, Ferragina, Cintia, & Pedreschi, 2019) shows PlayeRank, a data-driven framework that offers a multi-dimensional and role-aware evaluation of the performance of soccer players which also observed that top performances are rare and unevenly distributed since a few top players produce most of the considered excellent performances. Thus, a result that should be the focus is that top players do not always play excellently, they just achieve top performances more frequently than the other players. Besides, PlayeRank should be seen as a valuable tool to support professional football scouts in evaluating, searching, ranking, and recommending soccer players.

Under team performance analysis, the research of (Santos, et al., 2016) focuses on the analysis of match situations that finished in goal, and also 557 goals were analyzed from 10 teams across Portugal, Spain, England, and German. Regarding those only goals that were possible to obtain the sequence of actions from the moment of ball possession were considered. The study used the Football Goal Observation System and found a higher number of goals happens from ball recovered through a lost ball on the offensive zone and offensive midfield areas, where the last pass occurs in offensive sector zones, through counterattack, within the penalty area with the right foot corresponding 53,17% of the sample and left foot 28,83%. These results are shown in the next three figures.
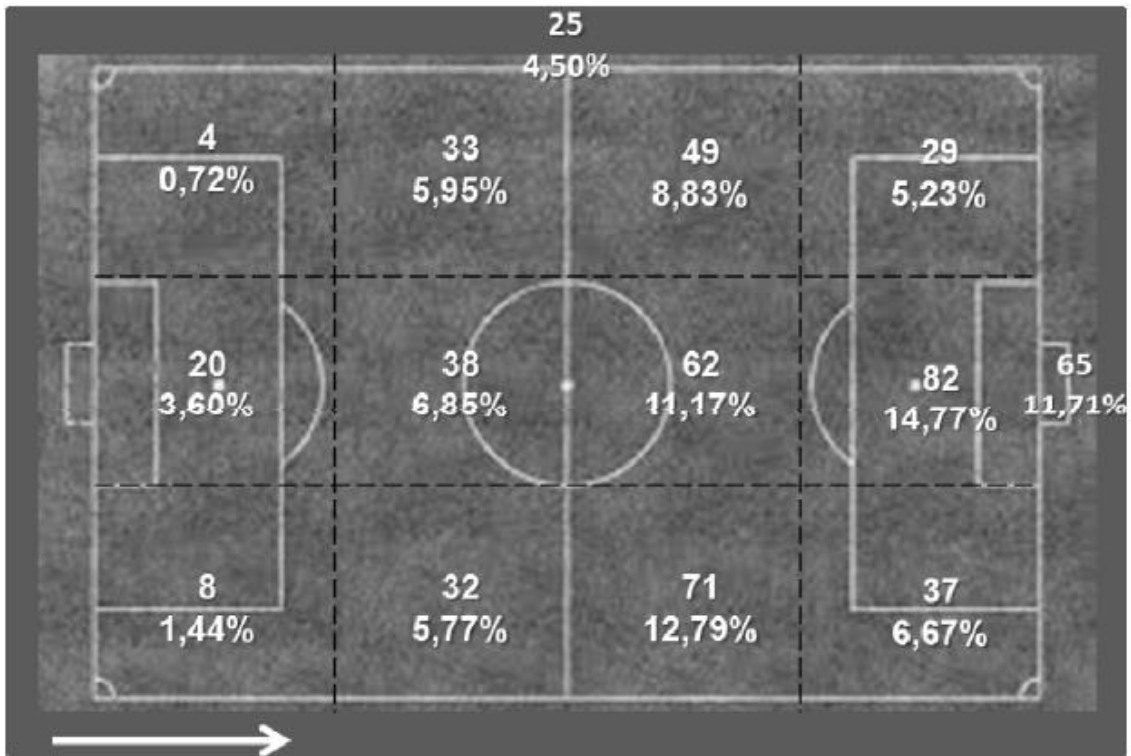
Figure 11 - Zone of ball recovered (Santos, et al., 2016)
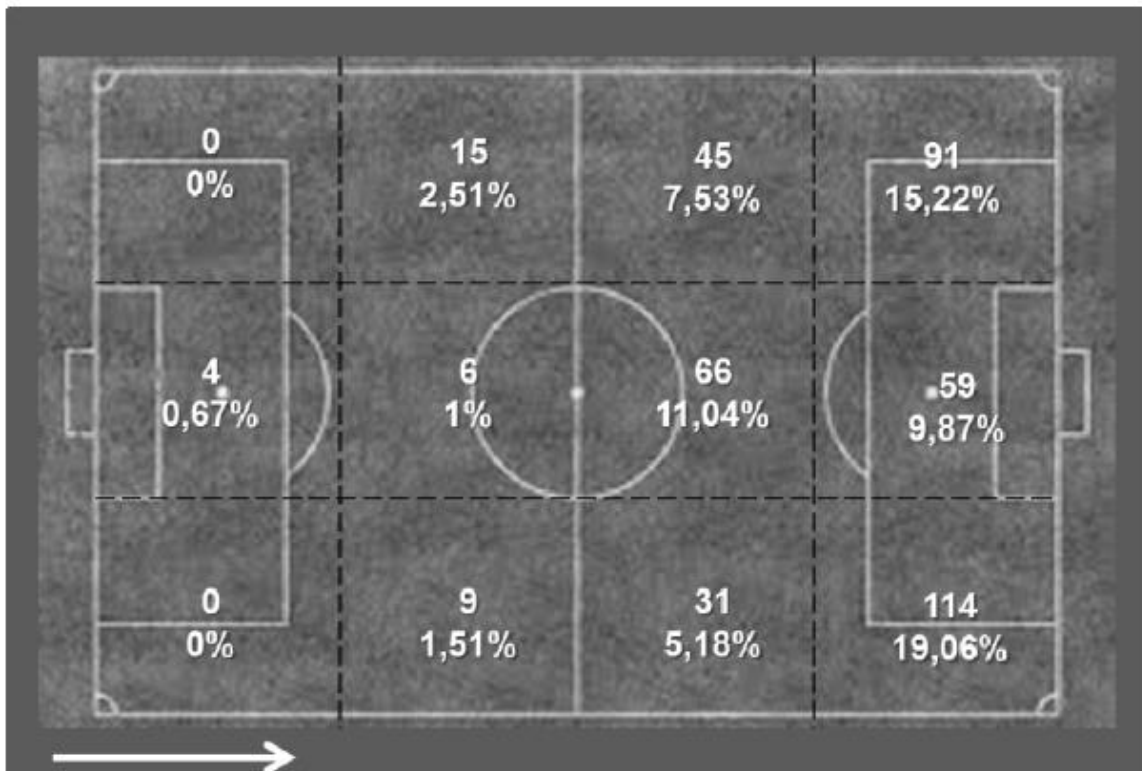
Figure 12 - Zone of the last pass to shot (Santos, et al., 2016)
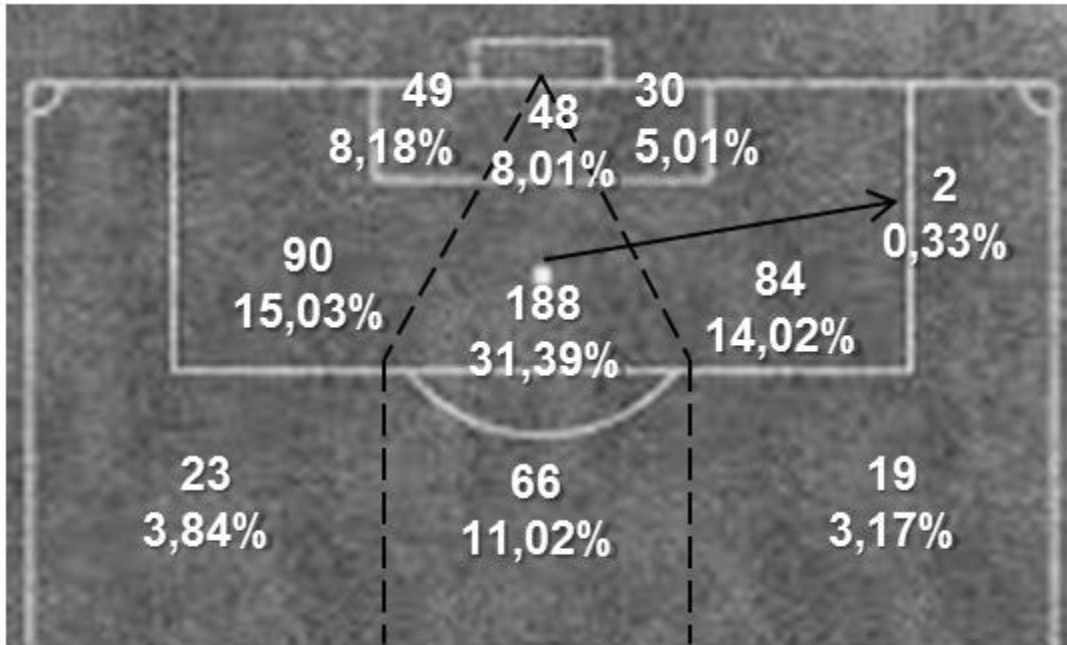
Figure 13 - Shot zone (Santos, et al., 2016)

The paper of (Pappalardo, et al., 2019) describes an open collection of soccer-logs that cover seven male football competitions provided by Wyscout and the data were approached on the football Data Challenge initiative (https://sobigdata-soccerchallenge.it/). Regarding that soccer-logs detailed match events, each containing these types of information: pass, shot, foul, tackle, a time-stamp, the player(s), the position on the field, pass accuracy, and other relevant collections.

As well as the other studies presented above, (Pappalardo, et al., 2019) explores football analytics use cases, as the two figures described below, regarding player performance on the team context. The first one presents the distribution of positions per event type, plotting the distribution of the positions of the events during the match. The darker is the green, the higher is the number of events in a specific field zone. Moreover, the same figure also describes the distribution of the passes' position during a match for each player's role. The darker is the color, the higher is the number of passes in a specific field zone. Then, the second figure shows the represent player passing networks of the match Napoli and Juventus, Italian first division, where each node is a player, and edges represent passes between players. Moreover, The size of the nodes reflects the number of ingoing and outgoing passes, while the size of the edges is proportional to the number of passes between the players.
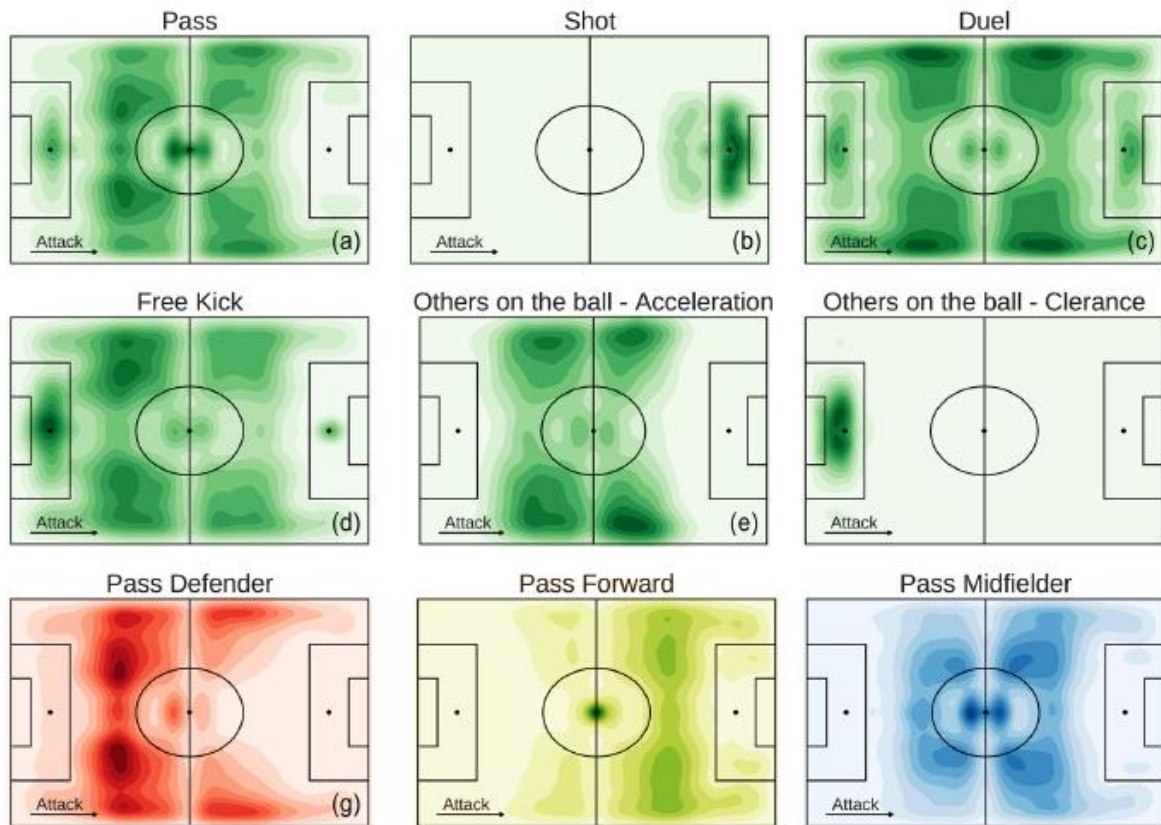
Figure 14 - distribution of positions per event type (Pappalardo, et al., 2019)
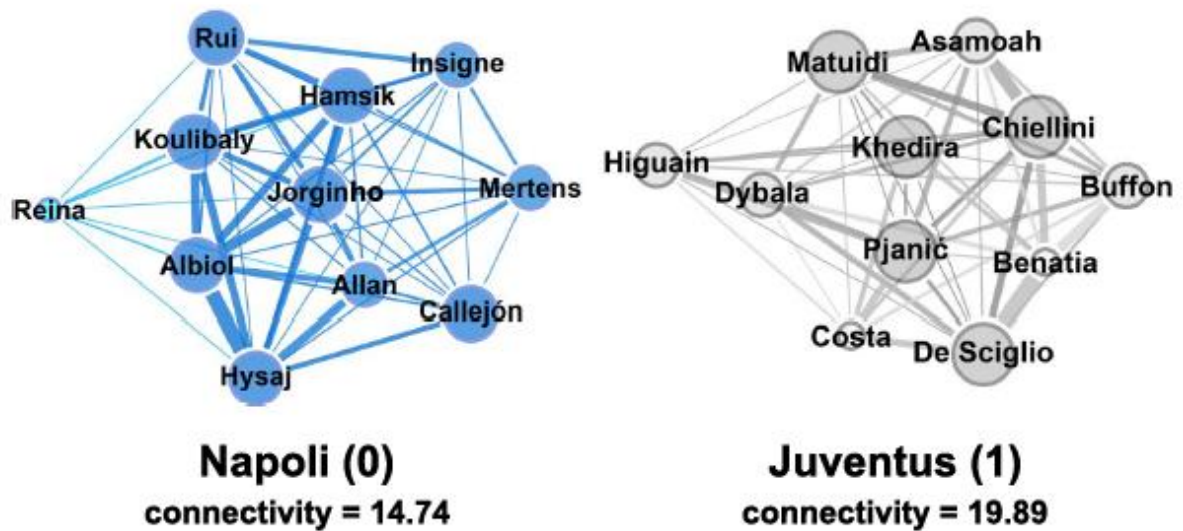
Figure 15 - Player passing networks (Pappalardo, et al., 2019)

3. MACHINE LEARNING MODELS

To evaluate and discuss the relevant ML models per authors, the study provides a summary of all the individual studies reviewed and presented in Table 2, tracking key findings in terms of authors, sample or dataset, methods or Algorithms, and key outputs.

There was a mix of studies, with different datasets and Algorithms. Each one has a specific approach and achievement, but instead of evaluating isolated variables and accuracy, this chapter has the goal to report useful methods that can bring results based on a previous investigation and may drive a better understand of behaviors and use key reference authors as an influence to drive researches.

Starting with Kumar's research, it helps to find attributes that are relevant to assessing players' performance and that most influence the game. Kumar explored different methods and the analysis found supports as a guide to having a starting point both to analyze specific characteristics as well as different machine learning algorithms that apply to football within a previously explored study.

Following table 2, the second line, the research of (Khan & Kirubanand, 2019), Besides, the performance comparison between two algorithms such as XGBoost and SVM, has the purpose of accuracy in terms of match prediction, this study within the current context of this work can support potential algorithms that can be useful, concerning algorithms that behave well within the scope of football variables.

In the sequence of the table, there is the study of (Boyd-Graber, He, Kwok, & Daumé, 2016), it approaches the use of machine learning to take advantage of the understanding of the opponent's behavior and adapt strategy based on predictions according to the opponent's parameters.

Within the studies presented, (Zawbaa, Hassanien, & El-Bendary, 2011) study shows the potential use of algorithms such as support vector machine (SVM) and neural network (NN) for analysis of football performance. The paper takes advantage of a video summarization system and shows applicability to achieve a result of great accuracy and precision. Another preliminary analysis that can be used as a reference for research within football.

The fifth line of the summary lists the (Kusmakar, et al., 2016) study. It analyzes the pattern-forming dynamics of player interactions that can leverage the understanding of tactical behavior. Also, the study explores quantitative measures of a team's performance, focused on player interactions. Moreover, the research shows a path that machine learning-enabled approach for automated predictive analysis of performance.

Key Outputs	Methods / Algorithms	Sample/Dataset	Study /Author
59 attributes obtained with positive gain ratio are the ones affecting match outcome With the applied algorithms, it concluded top of 34 attributes characterizes the match outcome to a satisfactory extent.	Multilayer Perceptron Functional Trees Sequential Minimal Optimization Naive Bayes Random Forest Decision Table Fuzzy Unordered Rule Induction Algorithm J48Graft J48 Jrip REP Tree LibSVM Kstar AdaBoostM1 with Functional Tree	The dataset of player performances for EPL released by OPTA contained 210 attributes and 10369 instances. Out of the 210 attributes of players, 198 attributes were performance statistics while the others were identifiers for the player for that match. The dataset released by OPTA did not contain match-outcomes. The dataset did not contain Own Goals, the goals scored by a player against his team. The data for Own Goals scored by each player in each match of the tournament was fetched from WhoScored.com	(Kumar, 2013)
An ensemble learning can be a better choice when trying to predict the results in this field than SVM	SVM	The Dataset selected contained features such as the number of goals scored by the home team, the number of goals scored by away team, Shots taken by the home team, Shots taken by away team, home team points, away team points, a variety of betting odds, and finally the Full-time result. The datasets collected were from the year 2000 to 2013.	(Khan & Kirubanand, 2019)

Study / Author	Sample/Dataset	Methods / Algorithms	Key Outputs
(Boyd-Graber, He, Kwok, & Daumé, 2016)	Game multi-agent that is played on a 6x9 grid by two players, which simulates movements and situations of a football environment	DRON (Deep Reinforcement Opponent Network) DQN	DQN-world is confused by the defensive behavior and significantly sacrifices its performance against the offensive opponent; DRON achieves a much better trade-off, retaining rewards close to both upper bounds against the varying opponent.
(Zawbaa, Hassanien, & El-Bendary, 2011)	Five videos for soccer matches from World Cup Championship 2010, Africa Championship League 2010, Africa Championship League 2008, European Championship League 2008, and Euro 2008.	ANN SVM	Compared to the performance results obtained using the SVM classifier, the proposed system attained good NN-based performance results concerning recall ratio, however, it attained poor NN-based performance results concerning precision ratio. Accordingly, it has been concluded that using the SVM classifier is more appropriate for soccer video summarization than the NN classifier.

Study/Author	Sample/Dataset	Methods / Algorithms	Key Outputs
(Kusmakar, et al., 2016)	A dataset from a season of Major League Soccer division of the United States and Canada. The dataset consists of the possession chain data from 13 matches. The interaction information (possession chain) comprises of time and duration of all ball passes and tackles between players. The dataset also includes the nature of the interaction which can be categorized as being between teammates or between opposing players. The positional information includes the x-y position of all individuals throughout the entire match.	NN SVM	machine learning-enabled approach for automated predictive analysis of performance and team's network derived using possession chain data, by quantitatively analyzing measures of performance that have a specific distribution and that can be used to predict the performance of a team.

Table 3 - ML Summary

Regarding the study of (Kumar, 2013), it presented the following table results of algorithms performance, sorted by correctly classified, ROC areas, F-Measure and Kappa statistic, taking into account all the threshold levels. The top four of all the involved are multilayer Perceptron, Functional Trees (FT), AdaboostM1 with FT, and Sequential Minimal Optimization (SMO). It shows the data used for the classification activity is applicable and informative in terms of match outcome.

Algorithm	% Correctly classified	ROC-Area	F-Measure	Kappa
MultiLayer Perceptron	69.4737	0.836	0.693	0.5235
FT	67.8947	0.801	0.683	0.5045
AdaboostM1 with FT	67.6316	0.827	0.672	0.4939
SMO	65.2632	0.754	0.648	0.4529
NaiveBayes	56.5789	0.746	0.571	0.3384
RandomForest	56.8421	0.74	0.552	0.3094
DecisionTable	57.6316	0.724	0.554	0.3196
FURIA	56.8421	0.715	0.551	0.3162
J48Graft	53.9474	0.637	0.527	0.275
J48	52.6316	0.631	0.522	0.2619
Jrip	52.1053	0.622	0.48	0.2174
REP Tree	47.1053	0.609	0.438	0.1342
LibSVM	45	0.5	0.279	0
Kstar	45	0.5	0.279	0

Table 4 - Algorithms performance on match data (number of attributes=336) (Kumar, 2013)

Moreover, the study of (Kumar, 2013), as moving forward on decreasing the number of attributes by setting gain-ratio threshold the performance of some algorithms change, it concludes that the Multilayer perceptron performance decreases as it decreases the number of attributes, but Sequential Minimal Optimization algorithm keeps its good prediction. Also, another interesting behavior that improved with the fewer variables approach was KSTAR.

Algorithm	% Correctly classified	ROC-Area	F-Measure	Kappa
MultiLayer Perceptron	59.4737	0.783	0.593	0.3694
FT	62.1053	0.736	0.619	0.4098
SMO	63.4211	0.74	0.621	0.4206
NaiveBayes	54.7368	0.761	0.555	0.3154
RandomForest	56.3158	0.743	0.549	0.3064
DecisionTable	58.6842	0.747	0.566	0.3385
FURIA	56.8421	0.501	0.501	0.287
J48Graft	52.6316	0.635	0.522	0.2591
J48	52.8947	0.644	0.527	0.2659
Jrip	54.2105	0.636	0.493	0.2399
REP Tree	55	0.668	0.513	0.2708
LibSVM	45	0.5	0.279	0.3667
KStar	51.5789	0.7	0.512	0.2438

Table 5 - Algorithms performance with selected attributes (threshold=0.07, number

of attributes=34) on match data (Kumar, 2013)

Besides presenting insights in terms of algorithms and the respective performance, the study of (Kumar, 2013) brings a setlist of highest gain ratio team attributes that can support the exploration of the relevance of football attributes. The table below is sorted by the top gain-ratio found in the research.

Gain-Ratio	Attribute
0.2423	SA Through Ball
0.2252	SH Successful Crosses Corners in the air
0.1483	AvgH Time Played
0.1448	SH Successful Crosses Left
0.1282	SH Shots On Conceded Inside Box
0.1173	SA Shots On from Inside Box
0.1099	SH Successful Short Passes
0.1048	SH Big Chances Faced
0.1048	SA Big Chances
0.1048	SA Shots On Target inc goals
0.1043	SH Total Successful Passes Excluding Crosses Corners
0.1019	SH Touches
0.0981	SH Total Successful Passes All
0.095	SA Shots On Conceded
0.0947	SH Successful Passes Own Half
0.0927	SA Shots On Conceded Inside Box
0.0922	SH Successful Passes Middle third
0.0914	SH Successful Passes Opposition Half
0.0891	SH Right Foot Shots On Target
0.0888	SH Shots On Target inc goals
0.0879	SH Shots On from Inside Box
0.0879	SH Big Chances
0.0879	SA Big Chances Faced
0.087	SH Pass Backward
0.0868	SH Shots On Conceded
0.0839	SH Pass Left
0.0811	SA Touches open play opp box
0.0807	SH Total Unsuccessful Passes All
0.0802	SH Red Cards
0.0796	SH Unsuccessful Passes Final third
0.0769	SH Left Foot Shots On Target
0.0736	SA Total Clearances
0.0723	SA Headed Clearances
0.0719	SH Successful Passes Defensive third

Table 6 - Top 34 highest gain- ratio team attributes (Kumar, 2013)

Additionally, the algorithms approached Kumar's study, the research of (Khan & Kirubanand, 2019) tested the performance of the highest gain-ratio team attributes applied to football. The proof followed the figure below, regarding that the flow starts on the training and test set. The process triggers the data clean up and the features are computed to the current data set. Moving forward the flow, selected features are put into the SVM model with the RBF kernel. Finally, the data is fed into the XGBoost model and the outcome from both the methods are compared.
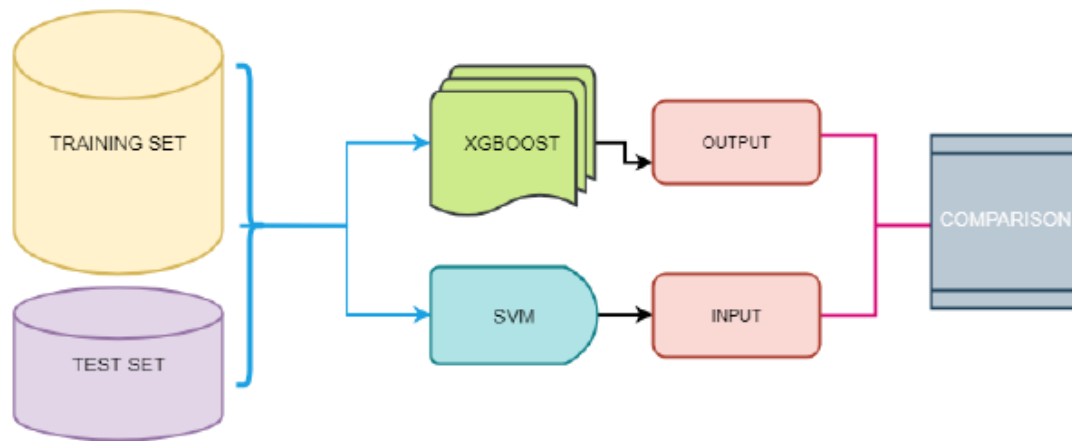
Figure 16 - Model used for comparison (Khan & Kirubanand, 2019)

Extrapolating the team attribute analysis, the study of (Boyd-Graber, He, Kwok, & Daumé, 2016) explores the concept of simulate movements and situations in a football environment, taking advantage of a multi-agent. Developing probabilistic models or parameterized strategies for specific applications, encoding observing through a deep Q-Network (DQN) and Deep Reinforcement Opponent Network (DRON) variations. The following figure illustrates how the game situations and the response against different behaviors.
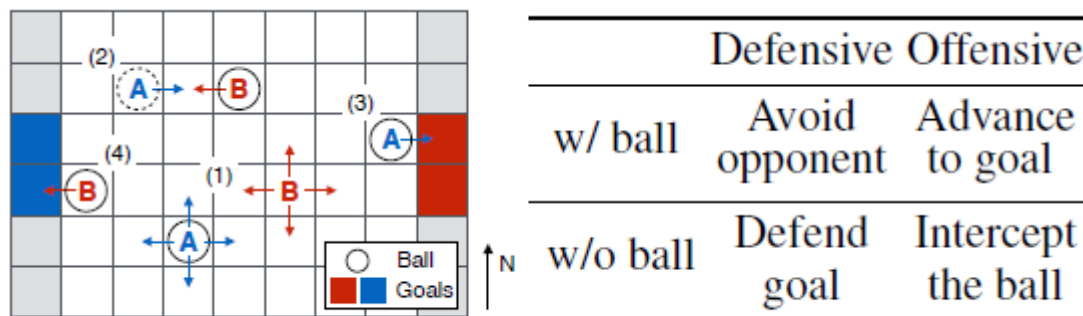
Figure 17 - Left: Illustration of the football game. Right: Strategies of the hand-crafted rule-based agent (Boyd-Graber, He, Kwok, & Daumé, 2016)

So far, these listed studies in this chapter can leverage the different types of analysis and cover together potential gaps within tested algorithms and outcomes. Furthermore, the findings and experimentation of (Zawbaa, Hassanien, & El-Bendary, 2011), using video as a source, adopting neural network (NN) and SVM, can underline the key events during the match.

According to (Zawbaa, Hassanien, & El-Bendary, 2011) and following the findings, table 6 describes an overview of the comparison between SVM and NN, and, as reported, the SVM shows a better Performance, once the precision detected was worth 20% higher than NN. Meanwhile, Table 7 shows the precision in terms of event detection and all the results are above 89%, which means the proposed system achieved high accuracy.

Factors	SVM	NN
Duration (hh:mm:ss)	1:53:39	1:53:39
Correct	103	98
False	8	43
Miss	2	7
Recall	98.1%	93.3%
Precision	92.8%	69.5 %

Table 7 - Evaluation of logo based replay using SVM and NN (Zawbaa, Hassanien, & El-Bendary, 2011)

Event Detection	Goal	Attack	Other events
Goal	57	3	0
Attack	6	176	8
Other events	0	18	283
Recall	95%	92.6%	94%
Precision	90.5%	89%	97.3%

Table 8 - Confusion matrix for event detection and summarization (Zawbaa, Hassanien, & El-Bendary, 2011)

After high accuracy achieved in the last study and (Boyd-Graber, He, Kwok, & Daumé, 2016) that shows an approach in terms of behavior and movements, there is additional research that copes with an analysis chain during the match. The findings of (Kusmakar, et al., 2016) can unlock the value in terms of this type of approach.

Regarding table 8, presented by (Kusmakar, et al., 2016), shows the type of events that can be leveraged according to a sequence during the match.

Team*	Type	SubType	Period	StartTime [s]	From*	To*	Start X (GPS)	Start Y (GPS)
team-1	Set Piece	Kick Off	1	47.4	Player 1		nan	nan
team-1	Pass		1	48.64	Player 1	Player 2	0.5	0.5
team-1	Pass		1	50.81	Player 2	Player 3	0.67	0.49
team-2	Challenge	Aerial Fault Won	1	52.68	Player 4		0.32	0.22
team-1	Challenge	Aerial Fault Lost	1	52.84	Player 3		0.32	0.23
team-1	Ball Lost	Forced	1	54.8	Player 3		0.32	0.23
team-2	Set Piece	Free Kick	1	55.96	Player 5		nan	nan
team-2	Pass		1	57.52	Player 5	Player 6	0.27	0.3
team-2	Pass		1	59.96	Player 6	Player 7	0.3	0.66
team-2	Pass		1	64.44	Player 7	Player 6	0.43	0.9

Table 9 - An example of ball possession chain data. The table shows a part of a ball possession chain dataset, which represents events in the 1st half of a match (Kusmakar, et al., 2016)

Nevertheless, predictive models developed by (Kusmakar, et al., 2016), show a mean accuracy of up to 75% in predicting the segmental outcome of the likelihood of a team making a successful attempt

to score. The following table presents the segments predicted in favor of a team with the overall prediction accuracy.

Team	Segments won - team 1 (Predicted)	Segments won - team 2 (Predicted)	Classification Accuracy	Predicted Winner*	Match Winner
G_1	5	7	0.75	team 2	team 2
G_2	10	11	0.62	team 2	team 1
G_3	13	9	0.77	team 1	team 1
G_4	7	7	0.71	team 1	team 1
G_5	12	7	0.78	team 1	team 1
G_6	8	12	0.70	team 2	team 1
G_7	11	14	0.72	team 2	team 2
G_8	9	5	0.71	team 1	team 1
G_9	17	10	0.74	team 1	team 1
G_{10}	6	15	0.71	team 2	team 1
G_{11}	8	5	0.69	team 1	team 2
G_{12}	7	7	0.71	team 1	team 1
G_{13}	8	11	0.84	team 2	DRAW

Table 10 - The predictive performance of the developed machine learning models. Shown are the segments predicted in favor of a team with the overall prediction accuracy, the predicted winner, and the true match results (Kusmakar, et al., 2016)

4. AUTHORS PER FIELD: PLAYER AND TEAM

Five different authors are reviewed and presented in table 10, tracking key findings in terms of authors, sample, field, and key outputs. The summary is distributed on 40% focused on the player, 40% focused on the team and 20% focused on player and team.

The first study listed in table 10, from the top to the bottom, study of (Decroos, Haaren, Bransen, & Davis, 2019) shows a framework for valuing player actions, regarding the outcome, taking into account the context. Thus, the player contributions can be measured according to offensive and defensive performance.

Shifting to the next row of table 10, the article of (Hughes, et al., 2012) approaches the player perspective as well, regarding performance indicators per position and category sets that may support the priorities of skills driven by the game needs.

On the other hand, some authors oversee both perspectives on the same study: team and player, such as (Kröckel, 2019). The outcome of this work is a valuable deep dive among different approaches regarding player and team performance. It's possible to extract insights on metrics and methods. Also, it supports an extended overview of other topics that are not the focus of this research, but it should be considered in terms of a reference to good references.

Moving forward with table 10, (Kumar, 2013) research presents a reference in terms of team perspective. Besides the last chapter, where this author was explored, the study can leverage specifically the team performance and attributes that may influence the game.

Last but not least, the study of (Kusmakar S., et al., 2020) shows the potential for uncovering local numerical markers of team performance. Regarding the topic ML was driven in chapter 9, it's important to underline that this research also has a machine learning-enabled approach. Moreover, it has valuable insights into team analysis.

Study / Author	Sample / Dataset	Field	Key Outputs
(Decroos, Haaren, Bransen, & Davis, 2019)	Wyscout data for the English, Spanish, German, Italian, French, Dutch, and Belgian top divisions. It considered 11,565 games played in the 2012/2013 through 2017/2018 seasons.	Player	Values all action types as passes, crosses, dribbles, and shots Reasons about an action's possible effects on the subsequent actions. The player actions that increase and decrease a team's chance of scoring
(Hughes, et al., 2012)	European Football Championships of 2004. The measure was based on a subjectively drawn continuum that analyses a player's technical movement throughout the game	Player	Key performance indicators per position and category sets

Study / Author	Sample / Dataset	Field	Key Outputs
(Kröckel, 2019)	Euro 2016 dataset among 51 games from OPTA database, not all games are used for each approach during the study. A selection was performed based on the aim of the analysis and the required amount of data.	Player and Team	Social network analysis (SNA) Dynamic network analysis (DNA) Overview of data and tools in SNA research in football Social network metrics used in football performance analysis SOM network be applied for team performance analysis Comparison of mining algorithms, regarding football Clustering algorithms Data and information useful for real-time decision support

Study / Author	Sample / Dataset	Field	Key Outputs
(Kumar, 2013)	The dataset of player performances for EPL released by OPTA contained 210 attributes and 10369 instances. Out of the 210 attributes of players, 198 attributes were performance statistics while the others were identifiers for the player for that match. The dataset released by OPTA did not contain match-outcomes. The dataset did not contain Own Goals, the goals scored by a player against his team. The data for Own Goals scored by each player in each match of the tournament was fetched from WhoScored.com	Team	Top of 34 team attributes that affect the match outcome
(Kusmakar S., et al., 2020)	A dataset from a season of Major League Soccer division of the United States and Canada. The dataset consists of the possession chain data from 13 matches. The interaction information (possession chain) comprises of time and duration of all ball passes and tackles between players. The dataset also includes the nature of the interaction which can be categorized as being between teammates or between opposing players. The positional information includes the x-y position of all individuals throughout the entire match.	Team	Potential for uncovering local numerical markers of team performance.

Table 11 – Authors Summary per field: Player and Team

Moving back to (Decroos, Haaren, Bransen, & Davis, 2019) research, the framework studied has different ratings, regarding actions of players' offensive and events influence. As an example of this kind of approach, there is the figure below.

		TIME	PLAYER	ACTION	P_{scores}	VALUE
	1	92m4s	S. Busquets	pass	0.03	0.00
	2	92m6s	L. Messi	pass	0.02	- 0.01
	3	92m8s	S. Busquets	pass	0.03	+ 0.01
	4	92m11s	L. Messi	take on	0.08	+ 0.05
	5	92m12s	L. Messi	pass	0.17	+ 0.09
	6	92m14s	A. Vidal	shot	1.00	+ 0.83
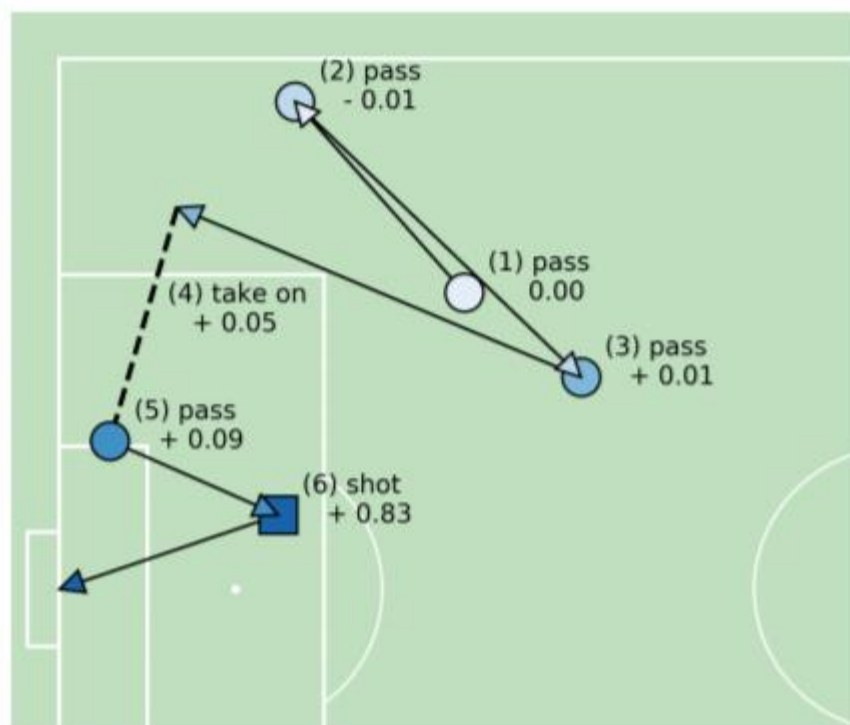
Figure 18 - The attack leading up to Barcelona's final goal in their 3-0 win against Real Madrid on December 23, 2017 (Decroos, Haaren, Bransen, & Davis, 2019)

Following the research of (Decroos, Haaren, Bransen, & Davis, 2019), figure 18 illustrates exactly what is stated in the title of the article: actions speak louder than goals. And connecting the next figure 19 shows a correlation between the number of actions per 90 minutes with the related value per action, the result of the finding demonstrates that Lionel Messi is a world-class player.

Figure 19 – Value per action (Decroos, Haaren, Bransen, & Davis, 2019)

Keeping in mind the player perspective, the study of (Hughes, et al., 2012) presents the following figure, indicating key performance Indicators (KPI) that best fit per position. Thus, it's a valuable reference to have a deep dive and tackle the important KPI regarding the player's objective during a match.

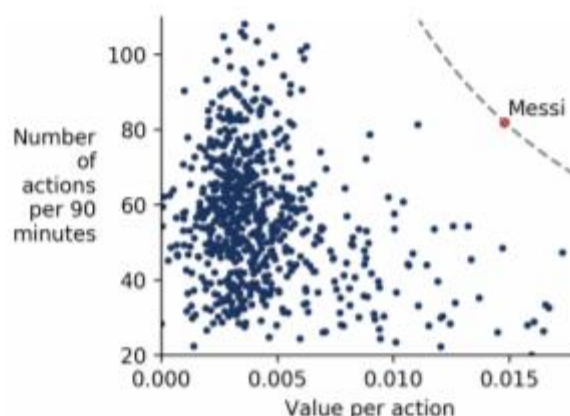| PERFORMANCE INDICATORS	GK	Full Backs	Centre Backs	HM	AM	WM	Strikers
Physiological	Height Strength Power Agility Coordination Reaction Time	Speed Power Stamina	Height Strength Speed Power Stamina	Stamina Speed Power Strength	Stamina Speed Power Strength	Speed Stamina Power Strength	Speed Agility Power Strength Stamina
Tactical	Vision Organisation Communication Distribution	Support play When to cross Passing Running off the ball Forcing offside	Vision Organisation Communication Passing	Vision Organisation Communication	Vision Organisation Communication	Vision Organisation Communication	Vision – awareness of space Anticipation Organisation Communication
Technical – Def	Shot stopping Coordination Recovery speed Save Punch	Tackle Pressing opposition Interception – anticipation Clearance Defensive header	Tackle Defensive header Pressing opposition Interception – anticipation Clearance	Tackle Pressing opposition Interception – anticipation Heading	Tackle Pressing opposition Interception – anticipation Heading	Tackle Pressing opposition Cover full-back Interception – anticipation Heading	Tackle Pressing opposition Interception – anticipation Heading
Technical – Att	Passing Throw Ball control with feet Kick Tackle	Tackle Interception – anticipation Dribbling Running with the ball Clearance Defensive header	Passing Heading Running with the ball Support play Dribbling Crossing Shooting	Passing Running with the ball Dribbling Support play Crossing Shooting Heading	Passing Running with the ball Dribbling Support play Crossing Shooting Heading	Passing Running with the ball Dribbling Support play Crossing Shooting Heading	Shooting Heading Reception Dribbling Passing Running with the ball Support play Crossing
Psychological	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language	Concentration Motivation Attitude Body language

Figure 20 - The skill requirements (Key Performance Indicators) for the different positions in soccer (Hughes, et al., 2012)

Shifting to another author, the research of (Kröckel, 2019) approaches some relevant topics among football analytics, on the player and team perspective. The results of the study present insights on metrics and methods, catching how coaches can take advantage of decision support during live matches.

One way to structure data, regarding (Kröckel, 2019), is to track a single event., for instance, shot on goal, in a manner that it's possible to define the actor (who), the type of action (what), time (when) and the pitch position (where). The figure below shows the file format that aggregates this kind of information.
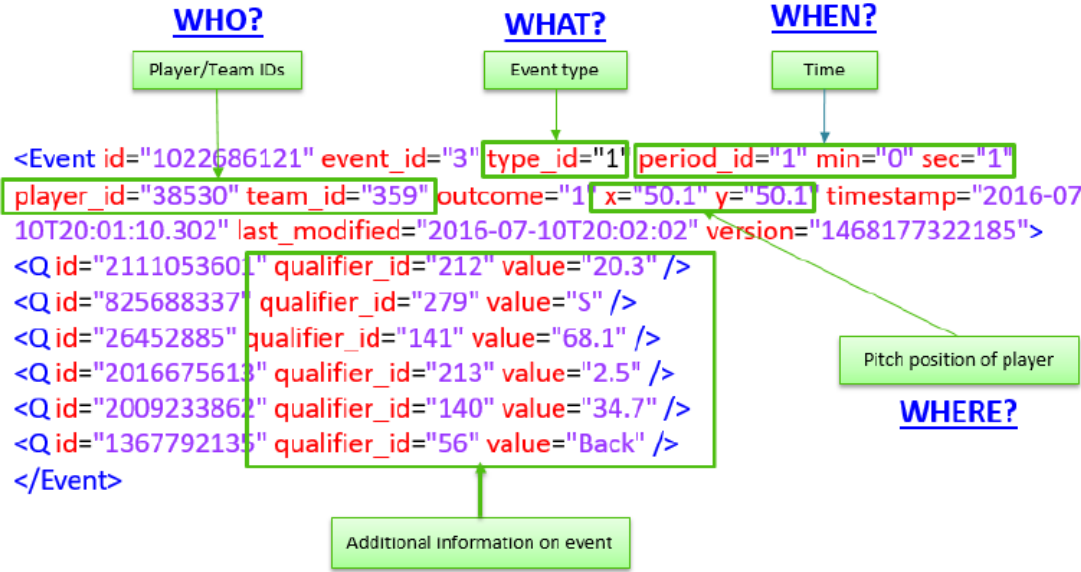
Figure 21 - Example of a passing event as recorded by OPTA (Kröckel, 2019)

Continuing on the same study, it shows a long list of authors and depicts player performance analysis and team performance, in a specific chapter, largely exploring social network analysis and dynamic network analysis.

Another interesting highlight on the (Kröckel, 2019) research is the chapter that deeply dives on the tactical behavior and the individual actions performed, regarding mining algorithms. Also, it explores a sequence of game steps ending on a final action target, as demonstrated in the following figure.

Figure 22 -England's offensive sequences ending in a shooting attempt (Kröckel, 2019)

Changing and looking forward to a reference about team perspective, there is the research of (Kumar, 2013), mentioned also on the last chapter, also (Kusmakar S., et al., 2020) analyzes football as a dynamic system, regarding player actions and the context among events as it demonstrates on the following figure.
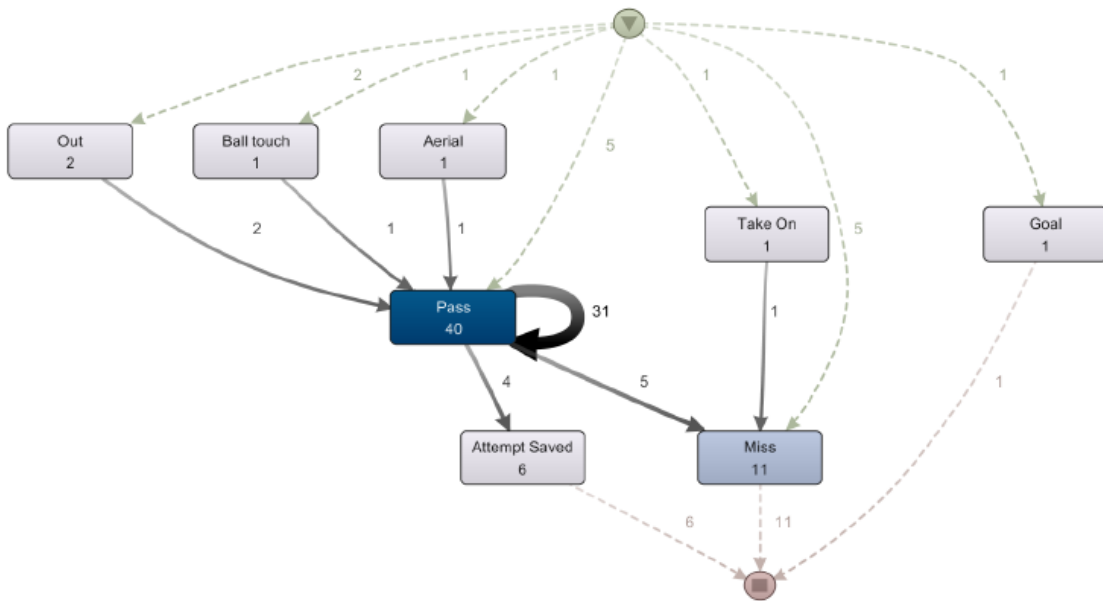
Team*	Type	SubType	Period	StartTime [s]	From*	To*	Start X (GPS)	Start Y (GPS)
team-1	Set Piece	Kick Off	1	47.4	Player 1		nan	nan
team-1	Pass		1	48.64	Player 1	Player 2	0.5	0.5
team-1	Pass		1	50.81	Player 2	Player 3	0.67	0.49
team-2	Challenge	Aerial Fault Won	1	52.68	Player 4		0.32	0.22
team-1	Challenge	Aerial Fault Lost	1	52.84	Player 3		0.32	0.23
team-1	Ball Lost	Forced	1	54.8	Player 3		0.32	0.23
team-2	Set Piece	Free Kick	1	55.96	Player 5		nan	nan
team-2	Pass		1	57.52	Player 5	Player 6	0.27	0.3
team-2	Pass		1	59.96	Player 6	Player 7	0.3	0.66
team-2	Pass		1	64.44	Player 7	Player 6	0.43	0.9

Figure 23 - An example of ball possession chain data. The table shows a part of a ball possession chain dataset, which represents events in the 1st half of a match (Kusmakar S., et al., 2016)

Keeping on the same study of   (Kusmakar S., et al., 2016), another key finding on it, is the combination of machine learning with team performance, achieving an automated machine learning model to predict the outcome of a game segment, comparing another authors and researches as illustrated on the next figure. It can drive results between different approaches.

Competition	Segmental Performance (AUC)[†]		Correctly Predicted Match Outcome[†]	
	Proposed Approach	Decross et al. [21]	Proposed Approach	Pappalardo et al. [20]
European Cup 2016	0.82 (0.79-0.84)	0.81 (0.78-0.85)	72.5% (29/40)	67.5% (27/40)
French First Division	0.83 (0.82-0.8)	0.83 (0.82-0.85)	80.4% (227/282)	81.5% (230/282)
German First Division	0.84 (0.83-0.85)	0.83 (0.81-0.85)	81.9% (181/221)	80.5% (178/221)
Italian First Division	0.85 (0.84-0.86)	0.85 (0.83-0.86)	85.5% (254/297)	79.4% (236/297)
Spanish First Division	0.84 (0.83-0.85)	0.83 (0.82-0.85)	78.1% (232/293)	81.5% (239/293)
World Cup 2018	0.83 (0.81-0.85)	0.82 (0.77-0.89)	70.1% (39/55)	70.1% (39/55)
English First Division	0.83 (0.82-0.84)	0.84 (0.83-0.85)	86.1% (240/279)	81.3% (227/279)
Overall	0.84	0.83	81.9% (1202/1467)	77.5% (1176/1467)

†

Figure 24 - Validation of the proposed approach on the largest open collection of soccer logs from 7 major competitions (Kusmakar S., et al., 2020)

5. CONCLUSIONS

5.1. SYNTHESIS OF THE DEVELOPED WORK

The objective of this study was to investigate and explore relevant academic contributions toward machine learning methods applied for football analytics and the respective used fields were the following: player individual technical skills and team performance. Furthermore, assess the most relevant ML models per authors on the football analytics field to deliver a summary. Besides, the aim was to evaluate the most relevant authors per chosen fields to deliver a summary as well. Finally, in this final chapter, give future directions for the next steps using this study.

In the first part of the research, the context of football analytics background was approached on the literature review. The boundaries and fundamentals of the study were established to drive the path for the next steps.

After that, analyzing the literature, the summary of machine learning methods among football analytics had five relevant findings which had a concentration on SVM and different kinds of neural network algorithms.

After comparing machine learning methods, the analysis went down on authors per the defined boundaries of scope perspective: team and player. Regarding results, there were five satisfactory studies in which it's possible to explore valuable player and team metrics, the importance of actions, and the sequence itself. Thus, the applicability of football analytics-driven on the specific fields approached on this research.

Despite the sample size found in the study, the results from the literature can drive other researchers to have a guide and, through compiled summaries, pave new further in-depth and targeted investigation toward football analytics.

Besides, it was concluded that football analytics has broadly vast opportunities in terms of research, regarding machine learning methods and a high potential to have a deep exploration of team and player perspective.

5.2. FUTURE WORK

As the next step, I would explore deeply samples and experiments in the field of machine learning application, as long as use cases taking advantage of team and player as the target. Nevertheless, it would be great to apply in a real-life world and follow up results, supporting football decisions and tracking achievements.

Second, future research might also further investigate the usage of the SVM, regarding that this model was found in most of the studies showed in the present research, and compare the results obtained to other machine learning algorithms, applying football datasets and targeting a player and teams analysis, identifying reliably and accurately.

Last but not least, football has alternative perspectives that influence team results and should be explored, regarding player recovery in terms of fatigue, psychological mindset, and other relevant aspects that may be addressed. Thus, another further investigation would be the correlation

between team results and other potential perspectives taking advantage of previous studies, and the list of researches already touched on the presented work.

6. BIBLIOGRAPHY

Américo, J. (2013). *Sistema de Seguimento de Jogadores de Futebol baseado em Vídeo de Baixa Qualidade.* Porto: Faculdade de Ciências da Universidade do Porto em Ciência de Computadores.

Anderson, C., & Sally, D. (2013). *The Numbers Game: Why Everything You Know About Soccer Is Wrong.* London: Penguin Group.

Aquino, R., Alves, I., Fuini, E., & Garganta, J. (2017, June). Skill-Related Performance in Soccer: A Systematic Review. *Human Movement*.

Babbar, M. (2019). A systematic review of sports analytics.

Borges, P., Garganta, J., Guilherme, J., & Jaime, M. (2019, August). Tactical efficacy and offensive game processes adopted by Italian and Brazilian youth soccer players. *Motriz. Revista de Educação Física*.

Boyd-Graber, J., He, H., Kwok, K., & Daumé, H. I. (2016). Opponent Modeling in Deep Reinforcement Learning. *Proceedings of the 33rd International Conference on Machine.* New York.

Castelão, D., Garganta, J., Afonso, José, & Costa, I. (2015, June). Sequential analytsis of attacking bahaviors performed by top-level national soccer teams. *Revista brasileira de ciência do esporte*, pp. 230-236.

Caya, O., & Bourdon, A. (2016). A framework of value creation from business intelligence and analytics in competitive sports. *49th Hawaii International Conference on System Sciences (HICSS)* (pp. 1061-1071). IEEE.

Clarke, N., & Noon, M. (2019). *Editorial: Fatigue and Recovery in Football.* Coventry University.

Collignon, H. (2019). *Winning in the Business of Sports.* Retrieved from Atkearney: https://www.atkearney.com/communications-media-technology/article?/a/winning-in-the-business-of-sports

Constantinou, A., & Fenton, N. (2017). Towards Smart-Data: Improving predictive accuracy in a long-term football team. *Knowledge-Based Systems*.

De Silva, V., Caine, M., Skinner, J., Dogan, S., Kondoz, A., Peter, T., . . . Smith, B. (2018, October 26). Player Tracking Data Analytics as a Tool for Physical Performance Management in Football: A Case Study from Chelsea Football Club Academy.

Decroos, T., Haaren, J. V., Bransen, L., & Davis, J. (2019, August). Actions Speak Louder than Goals: Valuing Player Actions in Soccer. *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19)*.

Doidge, M., Claus, R., Gabler, J., Irving, R., Millward, P., & Silvério, J. (2019). The impact of international football events on local, national, and transnational fan cultures: a critical overview. *Soccer & Society*.

FIFA. (2019, September 29). *FIFA*. Retrieved from https://www.fifa.com/livingfootball

Frias, T. (2012). *Changes in defensive playing methods influence the collective behavior of association football teams.* Universidade Técnica de Lisboa.

Hughes, M., Caudrelier, T., James, N., Redwood-brown, A., Donnelly, I., Kirkbride, A., & Duschesne, C. (2012). Moneyball and soccer - An analysis of the key performance indicators of elite male soccer players by position. *Journal of Human Sport and Exercise, 7*(2).

Khan, S., & Kirubanand, V. (2019). Comparing machine learning and ensemble learning in the field of football. *International Journal of Electrical and Computer Engineering (IJECE)*, 4321-4325.

Klyuchka, Y. A., Cherednichenko, O. Y., Vasylenko, A. V., & Yakovleva, O. V. (2015). Forecasting the results of football matches on the internet-based information. *Bulletin of NTU "KPI"*.

Korte, T. (2014, June 19). *Datainnovation.org.* Retrieved from Datainnovation.org: https://www.datainnovation.org/2014/06/how-data-and-analytics-have-changed-the-beautiful-game/

Kröckel, P. (2019, July). Big Data Event Analytics in Football for Tactical Decision Support.

Kumar, G. (2013). *Machine Learning for Soccer Analytics.* Dublin.

Kusmakar, S., Shelyag, S., Zhu, Y., Dwyer, D., Gastin, P., & Angelova, M. (2016). Machine learning-enabled team performance analysis in the dynamical environment of soccer.

Kusmakar, S., Shelyag, S., Zhu, Y., Dwyer, D., Gastin, P., & Angelova, M. (2020, March). Machine learning-enabled team performance analysis in the dynamical environment of soccer. *DSI Collaborative Research (Intelligent Sensor Processing for Enhancing Defence Decision Support)*.

Lepschy, H., Woll, A., & Wäsche, H. (2018). How to be Successful in Football: A Systematic Review. *The Open Sports Sciences Journal*.

Memmert, D., Bischof, J., Endler, S., Grunz, A., Schmid, M., Schmidt, A., & Perl, J. (2011). World-Level Analysis in Top Level Football Analysis and Simulation of Football Specific Group Tactics by Means of Adaptive Neural Networks. *Artificial Neural Networks*.

Mohammadi, M., & Sorour, S. (2018, June 5). Deep Learning for IoT Big Data and Streaming Analytics: A Survey. *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*.

Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*.

Novatchkov, H., & Baca, A. (2013, December). Artificial Intelligence in Sports on the Example of Weight Training. *Journal of Sports Science and Medicine*, pp. 27-37.

Pappalardo, L., Cintia, P., Pedreschi, D., Giannotti, F., & Barabási, A.-L. (2017, December). Human Perception of Performance. *arXiv on Physics and Society*.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*.

Pappalardo, L., Ferragina, P., Cintia, P., & Pedreschi, D. (2019, September). Player Rank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach. *ACM Transactions on Intelligent Systems and Technology, 10*.

Pires, M., & Santos, V. (2018). Assessing the Impact of the Internet of Everything Technologies in Football. *Journal of Sports Science 6*, pp. 36-55.

Pratas, J. M., Volossovitch, A., & Carita, A. I. (2018). Goal scoring in elite male football: A systematic review. *Journal of Human Sport and Exercise*.

Qing, Y., Ruano, M.-Á., Hongyou, L., Zhang, S., Gao, B., Wunderlich, F., & Memmert, D. (2020). Evaluation of the Technical Performance of Football Players in the UEFA Champions League. *International Journal of Environmental Research and Public Health*.

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*.

Santos, F., Mendes, B., Maurício, N., Furtado, B., Sousa, P., & Pinheiro, V. (2016). Análise do Golo em Equipas de Elite de futebol na época 2013-2014. *Revista de Desporto e Actividade Física, 8*, 11-22.

Sarmento, H., Campanico, J., & Marcelino, R. (2014). Match analysis in football: a systematic review. *Journal of Sports Sciences*.

Schoenfeld, B. (2019, May 22). *Nytimes.* Retrieved from Nytimes: https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html

Schulenkorf, N., & Frawley, S. (2017). *Critical Issues in Global Sport Management.* London: Routledge.

Spearman, W., & Basye, A. T. (2017). Physics-Based Modeling of Pass Probabilities in Soccer. *Sports Analytics Conference.* MIT Sloan.

Tax, N., & Joustra, Y. (2015). Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. *TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. X, NO. X, MONTH YEAR 1*.

Vanoye, J., Penna, A., Parra, O., & Díaz, D. (2017). Motivation Index to Improve Soccer Performance. *International Journal of Combinatorial Optimization Problems and Informatics, 8*, 45-57.

Zawbaa, H. M., Hassanien, A. E., & El-Bendary, N. (2011). Machine Learning-Based Soccer Video Summarization System. *Communications in Computer and Information Science*, (pp. 19-28).