PLAYRS:
A Player Rating System for Soccer

Jack Remmert

MSc in Sports Performance Analysis

University of Chichester

Department of Sport & Exercise Sciences

August 2019

Project Supervisor: Dr. Matt Robins

Acknowledgements

Over the course of writing this, the support of my family, friends, and professor were always garnered and appreciated. Their insight and reflections elevated the work herein to its highest standard.

Of most importance, I would like to thank Dr. Matt Robins at the University of Chichester for acting as my academic advisor and guiding me on this journey undertaken, but also challenging me to think differently. That guidance is impossible to translate into words on a page but is always in my mind.

Next, thank you to Southampton F.C. and especially the people within the Scouting and Recruitment Department for providing integral insights and perspectives.

As always though, my family supported me the most in providing me with the exogenous tools I needed to succeed and extremely respectful in trusting the process to pursue my dream of winning championships. But also, my friends who were always there to give me encouragement when I needed it the most.

Lastly, I would like to thank the reader of this dissertation, for without your future praises, criticisms, and reflections, it will not further the work of those to come.

Enjoy.

Table of Contents

PLAYRS: A Player Rating System for Soccer

Jack Remmert

Chichester Centre of Applied Sport and Exercise Sciences, University of Chichester, Chichester, Bishop Otter Campus, Chichester, UK.

Abstract

The aims of this study were to construct a model that predicts match outcome as a function of team performance and calculate objective player ratings based on their performance and position in context of a game. The model prediction was achieved through a Support Vector Machine with an AUC = 0.907 ± 0.01 and Accuracy = 78.7% ± 0.82%. Once the overall model was built, each player's performance, their rating, was calculated based on their positional key performance indicators given past research and their performance relative to the game. From these ratings, the offensive players, such as Centre Forwards, Attacking/Wide Midfielders, and Centre Midfielders, were rated higher than Defensive Midfielders, Centre Backs, and Full-Backs, because those offensive players executed more of the successful actions that led to winning. Comparatively, the ratings of players in this study exhibited a wider distribution versus other methodologies, but this is because this is the first study to model ratings based on a player's positional demands rather than every possible performance indicator. Next, the practical benefits and implications of this methodology were reviewed to show how it solves learning issues, such as highlighting and leniency errors, by removing the human element and thus objectively identifying the positives and negatives of performance as required by Performance Analysis research (Mackenzie and Cushion, 2013). Furthermore, coaches and scouts can implement this model in their respective processes, such as in pre-match analysis of opposition players or in the recruitment process, thus aiding in their decision making and feedback (Hughes and Franks, 2008).

1. Introduction

Currently, soccer is one of the most popular and oldest sports in the world given there exists a professional soccer league for almost every country and has been around since 1863 (FIFA, 2019). However, even though it is one of the longest running professional sports, soccer analytics lags behind other sports due to the difficulty in quantifying complex events and team dynamics (Duch *et al.*, 2010; Reiland, 2017). Like many sports, soccer analytics recorded events or performance indicators such as passes, shots, and goals dating back to 1910, similar to box scores in baseball or basketball (Eaves, 2015). This notational analysis then provided researchers a plethora of data to analyze and contributed to the growth of a novel industry: Performance Analysis (Lewis, 2014). Performance Analysis (PA) is the science relation to the collection, synthesis, interpretation, and communication of real-world sports data into relevant information (Robins and Hughes, 2015). Within soccer, PA has covered three key topics: the individual (Wiemeyer, 2003; Taylor *et al.*, 2004, 2005; Dellal *et al.*, 2011; Hughes *et al.*, 2012; Bradley *et al.*, 2013), the unit (Kim, 2004; Redwood-Brown, 2008; Oberstone, 2011; Collet, 2013), and team strategy (Hughes and Franks, 2005; Tenga *et al.*, 2010; Barreria *et al.*, 2013). Each one of these aspects of PA is vital because a coach or scout needs to understand the holistic integration of them rather than focus solely on one.

However, given the unit and team strategy is largely composed of the individual's role, a major topic within individual research is assessing their performance (Gréhaigne *et al.*, 1999; Borden, 2016; Brooks *et al.*, 2016; Pappalardo *et al.*, 2019), of which each sport has their own methodology. In baseball, professionals utilize statistics such as Wins Above Replacement (Slowinski, 2010), On Base Percentage (Lewis, 2013), Runs Created (Furtado *et al.*, 1999), etc. to search for objective knowledge about players, also defined as Sabermetrics (James, 1981). Basketball has APBRmetrics, which utilizes player's Offensive and Defensive Efficiency Ratings, Effective Technical Shooting Percentage, Plus/Minus ratio etc. (Kubatko *et al.*, 2007; Righini, 2013). Even American Football employs a quarterback rating to analyze the efficiency and productivity of the quarterback (Johnson, 1997). However, in soccer there exists few mainstream and objective methods which quantify how well players perform.

Today many sports data companies such as Opta and Statsbomb, along with media companies (e.g. Sky Sports, NBC Sports Network) and online platforms use simple performance indicators (PIs) to compare players. These PIs include passes, shots, tackles, etc., which are then plotted for comparison (Dewar, 2014) or used to produce numerical grades (WhoScored.com) of a player's performance. However, many of these media methods are based on the judge's personal bias (Pappalardo *et al.*, 2017). Secondly, the majority of research investigating player ratings, compares players on single PIs (Duch *et al.,* 2010; Peña and Touchette, 2012; Schultze and Wellbrock, 2018) or all the same PIs (Brooks *et al*., 2016; Pappalardo *et al.*, 2019) and thus disregards positional demands research. Positional demands research attempts to define the role of each position (Wiemeyer, 2003; Hughes *et al.*, 2012), then compares a range of PIs using averages with minor results like Forwards take the most shots or Midfielders run the most (Taylor *et al.*, 2004, 2005; Bradley *et al*., 2013). However, it is essential to grade players on the role they were assigned because coaches assign roles to players, who then perform certain actions relevant to this role, which then determines their performance (Gréhaigne *et al.*, 1999; James *et al.*, 2002). For example, Mauro Sandreani will ask his forward "*to 'dance' between the defenders*" (Zauli and Milan, 2002). While Carlo Ancelotti will ask one forward to get close to goal and another to be creative (Teja, 2017). Therefore, evaluating those forwards, and consequently other positions, on irrelevant PIs would be nonsensical because they did not fit within the coaches demands. Also, research has not properly evaluated player's performance in relationship to outcome (Hughes *et al.*, 2012) nor contextualized those indicators within the game. Thus, it is imperative to model each PIs impact on winning (e.g. Shots on Target, Penalty Area Entries, Ball Recoveries, etc.), then deduce which player contributed those to determine their rating.

Therefore, to address these gaps of subjective player ratings, positional demands research, and contextualizing the PI, the aims of this study were to 1) construct a model that predicts match outcome as a function of a team's performance relative to the opposition, and 2) provide a player rating based on their performance and position in context of a game. Coaches and scouts should utilize this methodology to monitor player's performances, support training regimes, and aide in the recruitment process, but ultimately define their own requirements for each position to ensure the player is being graded on the coaches' or scout's criteria (Butterworth *et al*., 2012; Lombardi, 2019).

2. Literary Review

2.1 Performance Analysis

In the beginning, PA largely developed as a necessity to understand the positive and negatives of performances to support the feedback process by coaches (Hughes and Franks, 2008). Specifically, PA was designed to provide objective and reliable feedback to support decision making for coaches and athletes (Bartlett, 2001; O'Donoghue, 2006), who were affected by the volume of events and subjectivity such as the halo effect (Iramaneerat and Yudkowsky, 2007), memory overload (Laird and Waters, 2008), highlighting (Franks and Miller, 1986), and leniency errors (Myford and Wolfe, 2003). However, it was vital to understand the critical components of performance to ensure appropriate feedback. Therefore, a variety of research was undertaken to determine the technical, tactical, and physical elements of success (see Mackenize and Cushion, 2013 for a review). From these results, PA has become an integral component of the coaching process (Carling *et al*., 2005; Drust, 2010) by providing augmented and objective feedback (McGarry *et al.*, 2014), which is tailored around positions, tactics, or player evaluation, focused on actual sporting events (O'Donoghue, 2010), and relative PIs or action variables which define the performance (Hughes and Bartlett, 2002).

2.2 Performance Analysis in Soccer

Within soccer, a dominant theme of research studies the individual, taking both a qualitative (Wiemeyer, 2003; Hughes *et al*., 2012) and quantitative approach (Taylor *et al.*, 2004; 2005; Barron *et al*., 2018). Another dominant theme analyzes the unit by comparing passing, space control, or shots between successful and unsuccessful teams (Oberstone, 2011; Timmaraju *et al.*, 2013). Thirdly, studies have analyzed the strategy of successful teams through transition play, ball recovery patterns, and set pieces (Tenga *et al.*, 2010; Barreira *et al*., 2013; Pulling *et al*., 2013). Although each theme is important to understand the holistic system, coaches assign roles to players who implement the strategy and through the combination of roles, the team dynamics can be distinguished (James *et al.*, 2002). Therefore, the evaluation of individual performances sets the foundation for other areas of PA research.

2.3 Positional Analysis Research

Thus, since players determine the team's performance, it is necessary to analyze each role through their own technical, tactical, and physical indicators (Gréhaigne *et al*., 1999; James *et al*., 2002). Research (reviewed in Table 1) has reviewed these indicators in both a qualitative and quantitative manner to define the requirement of each position (Gréhaigne *et al.* 1999; Wiemeyer, 2003; Taylor *et al.*, 2004; 2005; Bloomfield *et al.* 2007; Dellal *et al.*, 2011; Boone *et al.*, 2012; Bradley *et al.*, 2013; Hughes *et al.* 2012). Specifically, the qualitative research subjectively defines each position by their PI, but there is no clear consensus to the number of positions or unique PIs for each role. The quantitative research compares the average technical, tactical, and physical characteristics of a position's performance of each position to distinguish differences but utilize a limited number of PIs without operational definitions and their results provide little impact into the sporting world. Given the complexity of positions, any evaluation method needs to account for those differences by only grading players based on their role, otherwise any method will bias players and inaccurately grade them given their role.

Table 1. Summary of Qualitative and Quantitative Literature

Study	Positions Identified	Key Results and Limitations
Qualitative: Who should play in which position in soccer? Empirical evidence and unconventional modelling. (Wiemeyer, 2003)	Goalkeeper, Sweeper, Central Defender, Central MF, Wide MF, Offensive MF, Defensive MF, and Striker.	Provide a list of criteria for each position as well as a degree of agreement between coaches. However, the coaches report a variety of features and there is no complete agreement amongst coaches of each position's requirements or the number of positions that may exist.
Qualitative: Moneyball and soccer - an analysis of the key performance indicators of elite male soccer players by position. (Hughes *et al.* 2012)	Goalkeeper, Full-Back, Centre-Back, Holding Midfield, Attacking Midfield, Wide Midfield, and Striker	Based on 66 analysts (15 "*experts*" and 51 "*students*") opinions, they provided PIs ranging across Physiological, Tactical, Technical, and Psychological attributes for each position, many of which are the same for each position, but also failed to define these PIs with operational definitions.
Quantitative: Behavioural comparisons of positional demands in professional soccer. (Taylor *et al.,* 2004)	Full-Back, Centre-Back, Midfielder, Forward	Constructed performance profiles for Premier League players based on technical KPIs for each position to find significant differences between them, but also within intra-positional behavioral profiles. Thus, mean inter-positional profiles may hide the subtleties of individual player performance.
Quantitative: Comparison of physical and technical performance in European soccer match-play: FA Premier League and La Liga. (Dellal *et al.,* 2011)	Centre-Back, Full-Back, Central Defensive Midfielder, Central Attacking Midfielder, Wide Midfielder, and Forward	Compared the technical and physical indicators of players in La Liga to Premier League by position. Thus, Centre-Backs were compared only to Centre-Backs etc. From there findings they suggest that cultural differences may exist across leagues but do not state the relative impact of PIs on the positions.

Match performance and physical capacity of players in the top three competitive standards of English professional soccer. (Bradley *et al.*, 2013)	Centre-Back, Full-Back, Central Midfielders, Wide Midfielders, and Attackers.	Studied the technical and physical performance across 711 players in the Premier League, Championship, and English League 1 by position. Evidence suggested that players in lower leagues completed more high intensity runs despite a similar physical standard of players, further concluding that league differences may exist beyond technical differences.
Quantitative: Physical demands of different positions in FA Premier League soccer. (Bloomfield *et al.,* 2007)	Defender, Midfielder, and Striker	Compared primarily the physical demands of 55 players in the Premier League, but also the passing and dribbling across different positions. Results showed that position did have an influence on amount of time at different movement speeds.
Quantitative: Physical Fitness of Elite Belgian Soccer Players by Player Position. (Boone *et al.,* 2012)	Goalkeeper, Centre-Back, Full-Back, Midfielders, and Strikers	Studied 289 players from the Belgium League and analyzed their performance in agility drills, height, and weight to provide insights for conditional programs for positions.
Quantitative: The Foundations of Tactics and Strategy in Team Sports. (Gréhaigne *et al.,* 1999)	Centre-Back, Full-Back, Central-Midfielder, Wide-Midfielder, and Striker	Recorded the pitch location of players at 30-second intervals and concluded each player operated in a distinct area but reviewed more general tactics rather than assigned positions to areas on the field.
Quantitative: A Comparison of Individual and Unit Tactical Behavior and Team Strategy in Professional Soccer. (Taylor *et al.,* 2005)	Centre-Back, Full-Back, Midfielder, and Forward.	Attempted to enhance Gréhaigne *et al.*'s (1999) work by analyzing the location of event by position. Specifically, they analyzed 14 players across 22 games, to create a "heat-map" of actions based on 36 zones to illustrate where players performed, thus providing a foundation of which to analyze where positions perform.

2.4 Player Rating Methodologies

Current evaluation methods have also been qualitative and quantitative, but the most popular has been subjective player ratings in newspapers. Only one study modeled these ratings to specifically understand the process behind human evaluation of performance by producing an artificial judge (Pappalardo *et al.*, 2017). The "*judge*" predicted newspaper's ratings of players from 150 PI, (e.g. quality of passes, number of goals, dangerousness of players, etc.) and found these PIs did not fully explain the human evaluation process, only a 55% correlation. Instead, they found the subjective method was based on the halo effect, which is informed when a salient characteristic influences how other traits are judged (Thorndike, 1920). However, when re-training the model on both technical and contextual indicators like outcome of the game, nationality and club of the player, or expected outcome of the game estimated by bookmakers, the new ratings correlated 68%

with the newspapers, implying the presence of subjectivity. Next, Pappalardo *et al.* (2017) computed the importance of each PI for the goalkeeper (GK), defender (DF), midfielder (MF), and forward (FW) to find the PI were "*poorly considered or discarded*", but also noticeable features for GKs and FWs were technical aspects of play (e.g. saves and goals respectively), while noticeable features for DFs and MFs were goal difference, distance covered, and yellow cards, versus the expected PIs. Therefore, given the subjectivity present, media methods are not adequate methods to grade players.

Besides the qualitative attempt to model soccer player ratings, quantitative methods (reviewed in Table 2) have been constructed, but with varying success due to their subjectivity (McHale *et al*., 2012), advanced methods or use of limited PIs, primarily passing (Duch *et al*., 2010; Peña and Touchette, 2012). Some studies have simplified to measure goal difference between the players on the field (Schultze and Wellbrock, 2018), but these results do not fully explain performances, as some players may not have an impact on a goal scored at all. Secondly, all of these methods do not account for positional demands, thus bias offensive players and do not account for defensive actions. Therefore, although these methods produce an objective measure to support coaches and scouts' decisions, by only measuring one aspect of play, players are unfairly assessed on their overall game performance and impact. Thus, any quantitative methodology to create player ratings, must account for positional role differences.

Table 2. Review of Quantitative Player Rating Methodologies

Study	Model Description	Strengths and Weaknesses
On the Development of a Soccer Player Performance Rating System for the English Premier League (McHale *et al*., 2012)	EA Sports Player Performance Index, EASPPI, a weighted formula summed across six indices which measure a player's match contributions, minute played, appearances, goals scored, assists, and clean sheets to calculate a rating across the entire season. Each index was calculated based on a regression predicting points per match (3 for a win, 1 for a draw, or 0 for a loss).	Even though it was one of the earliest rating methods, the weights were based on subjective decisions to account for DFs and GKs being overvalued due to the model subtracting points for shots-off target from FWs and MFs, hence not correcting for position. Thus, the most weight, 37.5%, was given to the index measuring a player's minutes played relative to their teammate's minutes, and the least to the assist and clean-sheet indices, 6.25% for both. Furthermore, every goal was given the same value instead of treating each based on the scoreline of the game. Currently the latest found EASPPI results only exist until the 2015/2016 season showing the team of the season, indicating this method has decreased in popularity (talkSport, 2016).

Quantifying the Performance of Individual Players in a Team Activity (Duch *et al*., 2010)	Network Model using the players as nodes and weighted arcs based on the number of successful passes, but also included two additional nodes to measure Shots On or Off Target. From this network and player's passing and shooting accuracy, betweenness centrality was calculated and used to define their rating.	Only modeled performance based on two variables: passing and shooting. Even though passing is the most frequent event in the game, and shooting is the most important (Goes *et al.*, 2019), by not grading positions on their role, the model is biasing possession orientated teams and positions.
A network theory analysis of football strategies (Peña and Touchette, 2012)	Through Network Models with passes, this study measured various statistics such as Closeness, Betweenness, Pagerank Centrality, and Clustering coefficients to provide different methods on how to rate players	Even though successful teams tend to have more passes than unsuccessful teams (Castellano *et al.*, 2012; Adams *et al.*, 2013) the model is mono-dimensional that biases midfielders and does not take into account defensive actions.
A weighted plus/minus metric for individual soccer player performance (Schultze and Wellbrock, 2018)	An adjusted Plus-Minus (PM) rating, which measures the net score differential for every player on the field when a goal is scored but adjusted for the importance of goals and opponent strength using betting quotas.	The adjusted PM model was able to correct for the over simplification of counting each goal as the same, but by only measuring goals, improper assumptions may be inferred because some players may have little impact on a goal scored for their team, like Goal Keepers.

Besides soccer, baseball and basketball have produced player ratings, but with much more success. Baseball first stated to use Runs Created to estimate how many runs the player created based on their Hits, Walks, Total Bases, and At-Bats (James, 1981). Other statistics have also gained popularity measuring how frequently players get on base (On Base Percentage, OBP) and WAR (Wins Above Replacement Player) all of which measure a player's performance and can be used to build a strategy around to win (Slowinski, 2010; Lewis, 2013). On the other hand, in basketball, players are measured on their positive and negative contributions through Player Efficiency Ratings (PER), which measure a player's points per game, rebounds, assists, steals, turnovers, and missed shots (Heeren, 1988). Research has expanded this metric with Hollinger's PER, which adjusts for the number of possessions and minutes the player participates for, then compares the player's contributions to the league average (Righini, 2013). But these methods, although claim to be holistic in nature, do not measure defensive contributions accurately and have positional bias (S., 2010). Secondly, basketball research has also utilized their own PM model that evaluates the net change in score for each player on the court adjusted for teammates, opposition skill level, match location, and time (Rosenbaum, 2004; Ilardi, 2014). Other similar metrics to PM ratings in the NBA utilize Win Shares, which baseball also exploits, but calculates a player's Offensive and Defensive Rating to determine their marginal wins, which are then summed for each player to estimate a team's wins for the season (Kubatko *et al.*, 2007). Also, a variation of Win Shares is Wins Produced, which calculates a player's production, based on similar PI's to efficiency ratings, then adjusts for teammate's defensive rebounds, assists, the team's defense, and position (Berri *et al.*, 2008). These PM ratings have become more popular instead of efficiency ratings because of their rigorous mathematical nature and more accurate valuation of players (Berri and Schmidt, 2010). Once solved, these extensive formulas provide a simple metric, which coaches and scouts use to measure a player's influence. Therefore, given this success any overall player rating metric for invasion sports must measure two-dimensional (offensive and defensive) PIs, in addition to measuring positions on different requirements.

2.4.1 Machine Learning and Player Ranking Methodologies

However, with the evolution of big data, Machine Learning techniques have been applied to produce soccer player ratings with more success versus qualitative and quantitative methods (Brooks *et al.*, 2016; Memmert, 2018; Pappalardo *et al.*, 2019). The field of Machine Learning (ML) tries to build a system that automatically improves with experience and understands the fundamental components of that learning process (Mitchell, 2006). ML techniques are either supervised, as in the outcome is known like a Classification or Regression Model, or unsupervised as in Clustering (Bunker and Thabtah, 2019). Classification, the most popular (Domingos, 2012), and Regression models are similar to methods outside of ML, but Clustering is a method to detect sub-groups from a sample (Wilmink and Uytterschaut, 1984). To perform ML, a machine will learn and generalize a certain task (e.g. how to decide if emails are spam, the outcome of a game, or if patients will respond to a treatment) with respect to a performance metric and an abundance of data (Bunker and Thabtah, 2019). Also, ML techniques correct for overfitting through cross-validation methods and/or penalizing models with more features (Domingos, 2012). Currently, teams such as Manchester City and Barcelona utilize ML to support decision making both on and off the field (Fernández *et al.,* 2016; Tobin, 2018). Outside of soccer, ML techniques are implemented in a wide range of industries from the music industry by predicting songs people enjoy (Mandel *et al.,* 2006), to the medical industry with gene coding and medical imaging

(Wernick *et al.*, 2010; Yennamalli, 2019), and with technology companies such as Google to provide relevant search results and smart replies for emails (Le, 2016).

Brooks *et al*. (2016) utilized a classification ML technique called a Support Vector Machine (SVM), which works well with complex problems in high dimensions, has less risk for overfitting, and generally gives better results to other classification techniques such as Artificial Neural Networks (Agresti, 2013). The model, called Pass Shot Value (PSV), classified each pass, in a possession sequence of three or more passes, from a starting point *i* to a destination point *j* to determine if the possession ended in a shot and found an Area Under the Receiver Operating Curve (*AUC*) value of 0.79 indicating the model had a 79% chance of distinguishing a random pass that would lead to a shot than a randomly chosen pass which does not (Narkhede, 2018). From the research, evidence suggested long passes were less effective than short passes to produce a shot, hence it does matter how the ball is moved around the field. Also, although PSV did bias offensive players as the authors stated, the authors compared players by their line: offense, midfield, or defense rather than grading all players within the same space to account for the positional bias. Thus, by incorporating location, utilizing ML techniques, and correcting for positional bias, PSV was an objective method which coaches and scouts can utilize to monitor players, but only if their role was attacking orientated. If the role of the player was not attacking orientated, then PSV would unfairly bias them, yet it is a novel process with promising results.

Another player rating method, which uses a SVM, is PlayeRank by Pappalardo *et al*. (2019). Instead of only evaluating passes and shots, PlayeRank utilized event data from 18 competitions and over 20,000 matches. Each event was either a pass, foul, shot, duel, free kick, offside, or touch that was given a subtype (e.g. what type of pass, either a cross or simple, was it an air duel, dribble, tackle, etc.) and a tag (e.g. accurate, no card, yellow card, counter attack, etc.) for a total of 76 different features. These events were then used to predict the outcome of a match as either a Win or No Win with an *AUC* of 0.89. Other classifications were tested such as Loss or No Loss and Win, Draw, or Loss, but the model's accuracy did not significantly improve. Once each PI's weight was determined, PlayeRank calculated a player's game rating as their production multiplied by the respective PI's weight. Next PlayeRank classified each player's position to compare the ratings using *K*-means Clustering to find 8 different positions, while other research has historically classified players into only 3 or 4 positions: Goalkeeper, Defender, Midfielder, and Forward (Soto-Valero *et al.*, 2017; Brecque, 2018; van de Ven, 2018). However, coaches already provide players with a role before the match, hence a methodology does not necessarily need to assign a position. Secondly, by not normalizing the data to account for the frequency of their occurrence within the game, the results may have been misinterpreted within the prediction of match outcome (Hughes and Franks, 2005). Therefore, although PlayeRank is a robust ML method with greater accuracy than PSV, a player rating methodology needs to contextualize events within the game.

2.5 Modeling Outcome

Given past research success with modeling match outcome to produce player ratings, any player ranking methodology should first start there. Research to date has focused on studying match outcome by analyzing the technical, tactical, and physical factors with significance tests and correlations. Technical research has shown successful teams tend to possess the ball for longer (McGarry and Franks, 2003; Jones *et al.*, 2004; Redwood-Brown, 2008), have more passes per possession (Hook and Hughes, 2001; Hughes and Franks, 2005), more penalty area entries (Ruiz-Ruiz *et al.*, 2013), convert more dribbles (Rampinini *et al.*, 2009) and win the ball further up field from either tackles, interceptions, or ball recoveries (Barreira *et al.*, 2013; Almedia *et al.*, 2014). Also, technical research has studied shot location and its impact on scoring goals (Olsen, 1988; Dufour, 1993; Yiannakos and Armatas, 2006; Wright *et al.,* 2011). Secondly, the tactical elements that contribute to success have been studied through space control. These studies have employed the use of Voroni Diagrams (Figure 1) to calculate the total space a team employed to find that successful teams tend to control more space (Kim, 2004; Fonseca, 2012).

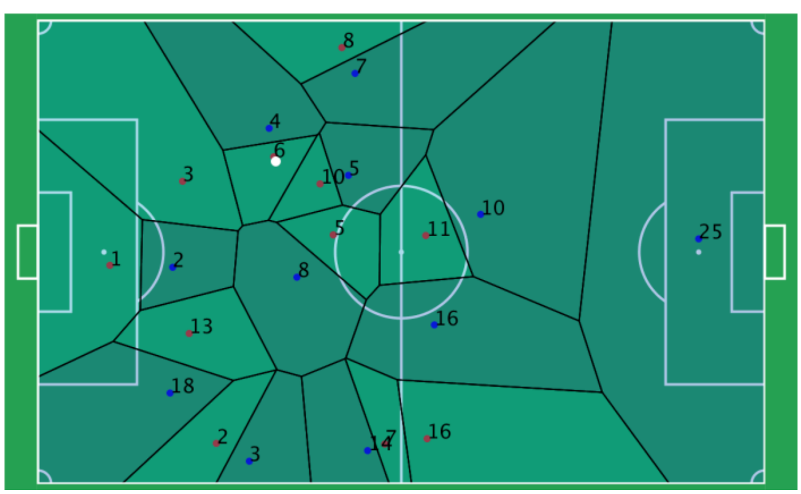

Figure 1. Example of a Vornoi Diagram

Thirdly, research has examined the physical demands of the game to find evidence which suggests players cover between 10-13km per game, primarily from walking and jogging (Mohr *et al.*, 2003; Bradley *et al.*, 2010) and players run further and faster against higher quality opponents (Castellano *et al.*, 2011), but current research to date has not yet modeled the physical factors of successful teams. Lastly, past research has also shown teams are more likely to win at home rather than away, hence situational factors such as location have an impact on outcome (Tucker *et al.*, 2005; Moskowitz and Wertheim, 2011; García *et al.*, 2015; Marek and Vavra, 2017).

With the determinants of success outlined, it is important to understand how accurate those variables are in predicting outcome to ensure comparable research (Mackenzie and Cushion, 2013). From these studies (reviewed in Table 3), which use ML or linear and logistic regression techniques, the results showed varying levels of accuracy, but most fail to state operational definitions and discuss practical conclusions that would impact the game.

Table 3. Review of Models Predicting Match Outcome

Study	Model	Summary
Comparing Team Performance of the English Premier League, Serie A, and La Liga for the 2008-2009 Season (Oberstone, 2011)	Linear Regression	Predicted each team's season's total points as a function of technical PIs such as Total Crosses, Average Goals Conceded, Tackles, etc. Overall accuracy, measured as $Adjusted\ R^2$, was greater than > 0.97 in each model. Lastly, performed an ANOVA test between different leagues to discover which league produces the most crosses or tackles or which has better passing success rates with a wide range of results, for example teams in La Liga and Serie A take more shots than teams in the Premier League.
The possession game? A comparative analysis of ball retention and team success in European and international football, 2007–2010 (Collet, 2013)	Linear Regression	Using data from 5 European leagues, UEFA and FIFA tournaments, this study predicted mean points per match as a function of possession time, total passes, and passes to shots on goals ratio for a total of 3 models with the best $Adjusted\ R^2$ from the total pass model of 0.64. They found passes were strong predictors, but once team quality and home advantage were taken into account, the power of prediction declined. Specifically, in league play, greater possession was negative, while in UEFA and FIFA, there was no significance.
Predicting Football Match Results with Logistic Regression (Prasetio and Harlili, 2016)	Logistic Regression	Constructed a model predicting Home or Away wins from the EPL 2015/16 season based on Home Offense, Home Defense, Away Offense, and Away Defense, but fail to state operational definitions. Only were able to achieve an accuracy of 69.5% and concluded the significant variables were Home Defense and Away Defense, but no practical conclusions were made.

Game ON! Predicting English Premier League Match Outcomes (Timmaraju *et al.*, 2013)	Multinomial Logistic Regression and Support Vector Machines	Through a Multinomial Logistic Regression, since there were more than 2 outcomes, and Radial based SVMs on EPL data from the 2013/14 season, the study predicted match outcome based on a team's Goals, Corners, Shots on Target in both the game and their average from *k*-past performances. The highest accuracy was found when $k = 7$ on the 2-Predictor Radial SVM of 83.78, but the 1-Predictor Radial SVM exhibited an Accuracy of 66.67%. However, there was little discussion of their results as to which past performances and variables contributed the most to the results.
Predicting Football Results Using Machine Learning Techniques (Herbinet, 2018)	Various ML techniques such as Cluster Analysis, SVM, and Decision Trees	Tested a wide variety of ML models on the Top 5 European Leagues to predict match outcome based on a team's ELO ratings, which were projected from both a shot and non-shot expected goal model, which was predicted from an unvalidated dataset and with no operational definitions. The model with the highest accuracy was the Linear SVM, 51%, closely followed by Neural Networks and Radial SVM, while the lowest model was a Random Forest with 45%. These results primarily indicate the need for reliable datasets and that SVMs tend to be the optimal model.
Analysis of football game-related statistics using multivariate techniques (Moura *et al.*, 2014)	Cluster Analysis	Based on game related statistics such as shots, fouls, yellow cards, possession, etc., from 2006 World Cup data, the study clustered teams into winning and non-winning groups with an accuracy of 70.3% for winning and 67.8% for non-winning teams, then showed which variables were key to winning such as Shots on Target, Fouls Won, Corner Kicks, and losing such as Cards.
Applying Machine Learning to Event Data in Soccer (Kerr, 2015)	Ridge Regression (Type of ML which penalizes coefficients with a minor contribution to the outcome (Hoerl, 1978))	Created a Ridge Regression model to predict the outcome of a game as either a win or non-win using PIs from La Liga. These PIs were pass percentage, shot distance, and tackle percentage, but also were contextualized on location (e.g. passes were split between Attacking Third and Overall, while tackles were split between Attacking and Defensive Halves). Lastly, within the model, each team's values, as well as the difference, instead of only analyzing the value of each variable per team, as Pappalardo *et al.*, (2019), thus providing some context of the game within the model.

The research on player rating systems is vast, but very few methods are objective, provide reliability tests on data, and operational definitions as PA warrants (Mackenzie and Cushion, 2013). Furthermore, these player rating systems do not factor in positional analysis research or only measure performance on few PIs. However, ML methods have shown promising results with high levels of accuracy through incorporating location data and modeling outcome using event level data. But research needs to contextualize the events within a game otherwise results may be misinterpreted. Hence, to address these gaps, the aims of this study were twofold. First construct a ML model that predicts the outcome of each game as a function of a team's performance relative to the opposition. Then, from the model's output, provide an objective player rating based on their performance and position.

3. Methods

3.1 Participants

A total of 11,862 games from the English Premier League (EPL), Spanish La Liga (ESP), French Ligue 1 (FRA), German Bundesliga (GER), Italian Serie A (ITA), English Football League Championship (EFL), Belgian First Division (BEL), and the Dutch Eredivisie (NLD) were sampled from the 2015/2016 to 2018/2019 season (Table 4). These leagues were selected given their primary use in research (see Mackenize and Cushion, 2013 for a review) and the global financial impact of those leagues (Green and Jones, 2019). In addition, a total of 5,907 unique players were ranked given they played a minimum of 45 minutes in a game. This 45-minute minimum, which was developed in consultation with four professional scouts with 24 years of cumulative experience, was defined to account for players coming off too early, as in from an injury, or on too late, as in being substituted. Also, this filtering process is similar to current media methods not providing a rating for players introduced after a certain time within the game (ESPN FC). Lastly, ethical approval was granted by the University of Chichester Ethics Committee and informed consent provide by an elite soccer team to perform and utilize data for research purposes.

Table 4. Participants

Competition	Region	Number of Seasons	Number of Games	Number of Players
English Premier League	England	4	1,520	827
English Championship	England	4	2,207	1,228
Italian Serie A	Italy	4	1,520	870
German Bundesliga	Germany	4	1,224	733
Spanish La Liga	Spain	4	1,520	931
French Ligue 1	France	4	1,519	974
Belgian First Division A	Belgium	4	1,104	794
Dutch Eredivisie	Netherlands	4	1,248	813
	Total		**11,862**	**5,936***

Table 2: List of Participants used for the study
** total number of unique players*

3.2 Procedures

Data for each game was collected from Opta (London, UK), which has been shown to provide strong levels of reliability (Liu *et al.*, 2013), using Structured Query Language (Microsoft, USA) to summarize the events on a team and player level (Figure 2).

Team Data Example								
Comp	Season	Game Id	Team	Opposition	Outcome	V1	V2	V3
EPL	2018/19	987619	Leicester City	Southampton F.C.	W	83	25	197
EPL	2018/19	987619	Southampton F.C.	Leicester City	L	46	21	226

Player Data Example								
Comp	Season	Game Id	Player's Team	Player	Position	Mins. Played	V1	V2
EPL	2018/19	987619	Leicester City	Wes Morgan	CB	95	13	1
EPL	2018/19	987619	Southampton F.C.	Nathan Redmond	WM	95	1	4

Figure 2. Example of Data

Example of event data summarized in tabular form from SQL corresponding to Southampton F.C.'s home game against Leicester City in the EPL during the 2018/19 Season and two players who played in the game, Wes Morgan at Centre-Back and Nathan Redmond at Wide-Midfielder. V1, V2, and V3 refer to variables analyzed.

These events are the PIs for each team. When defining these PIs, analysts may have different definitions (Carling *et al.*, 2005; Williams, 2012). Due to this subjectivity and importance of achieving consensus, all 48 operational definitions are defined in Table 5 and visuals are provided in Appendix A (Williams, 2012). These operational definitions and figures have been cross-validated with Opta, an elite scouting panel, and past research. Also, some variables are based on location and divided into thirds, unless noted otherwise, as a sensible compromise between precision and accuracy (Casal *et al.*, 2017).

Table 5. Definitions

KPI	Conditions & Restrictions	Number of KPIs	Definition
Pass[ς] (Redwood-Brown, 2008)	1. Location: Defensive, Middle, or Attacking Third. 2. Outcome: Successful or Unsuccessful. 3. No Free Kicks, Crosses, Corner Kicks, Throw-ins and Goal Kicks.	6	Any intentional played ball with any part of the body from one player to another. If received by a teammate it was successful, otherwise the pass was deemed unsuccessful.
Cross[ς]	1. Outcome. 2. Does include Free Kicks, Crosses, and Corner Kicks.	2	Any intentional played ball from a wide position intending to reach a teammate in a specific area in front of the goal.

Penalty Area Entries[ς] (Ruiz-Ruiz *et al*., 2013)	1. Outcome.	2	Any intentionally played ball into the Attacking 18-yard box.
Shots[ς] (Ensum *et al*., 2005)	1. Location: Inside or Outside the Box. 2. Outcome: On-Target or Off-Target.	4	Any intentional goal attempt. On-Target is defined as any goal that goes in regardless of intent or a clear attempt to score that would have gone in except for being saved by the keeper or last-man. Off-Target is defined as a shot that goes over or wide of the goal without making contact with another player, a hot that would have gone over or wide of the goal but for being stopped by a goalkeeper's save or outfield player, or lastly, hits the frame of the goal and not scored.
Aerials[ς]	1. Location. 2. Outcome.	6	This is where two players challenge in the air against each other for the ball.
Ground Duels[ς]	1. Location. 2. Outcome: Won or Lost.	6	This is an attempt by a player to beat an opponent when they have possession of the ball. or when a player is tackled and dispossessed.
Tackles[ς] (Hargreaves and Bate, 2010)	1. Location. 2. Outcome.	6	A tackle is defined as a player connects with the ball in a ground challenge where he successfully takes the ball away from the player in possession. The tackled player must clearly be in possession of the ball before the tackle is made. A tackle won is deemed to be where the tackler or a teammate regains possession from the challenge. A tackle lost is where a tackle is made, but the ball goes to an opposition player.
Interceptions[ς]	1. Location	3	This is where a player reads an opponent's pass and intercepts the ball by moving into the line of the intended pass.
Ball Recoveries[ς]	1. Location	3	This is where a player recovers the ball in a situation where neither team has possession or where the ball has been played directly to him by an opponent, thus securing possession for their team.
Clearances[ς]	NA	1	This is a defensive action where a player kicks the ball away from his own goal with no intended recipient.
Fouls[ς]	1. Location. 2. Outcome: Won or Committed.	6	Any instance in which a team wins a free kick or penalty after a violation of the rules as deemed by the referee. Offside Is not given as a foul won or conceded.

Cards[ς]	1. Type: Yellow or Red 2. No difference was made between Double Yellow and Red.	2	Any instance in which a player receives a card for a violation of the rules as deemed by the referee.
Space Control	NA	1	This is measured as the Ellipsoidal Area, $Area = \sigma_1 * \sigma_2 * \pi$, based on the player's average position determined through Principal Component Analysis as outlined by Moura *et al.* (2015). See Appendix B for explanation.
Situation[ς]	1. Match Location: Home or Away	2	Observation if the game was Home or Away.

[ς] *indicates validated by Opta Research and Scouting Panel. Validated by other research when source present.*

3.3 Reliability

To ensure the data was reliable and objective, both inter-observer and intra-observer reliability tests were performed, as differences in measurements can influence external validity (Cooper *et al.*, 2007). The inter-observer tests were conducted by the researcher using SportsCode (Version 10.3.36; Hudl, Lincoln, NE), who has ten months experience in this software. The intra-observer tests were conducted within a six-week gap given the general practice (Jones *et al.*, 2004; Scoulding *et al.*, 2004; James *et al.*, 2005; Worsfold and Macbeth, 2009). Previous research utilizing Opta data did not test for reliability (Oberstone, 2011; Collet, 2013; Beato *et al.*, 2018), but Opta data has shown to be reliable between one game and one player in Liu *et al.* (2013). Also, other research has reviewed 1-22% of their data, but the number of matches analyzed ranges from one half (Turner and Sayers, 2010), two twenty-minute games (Eldridge *et al.*, 2013), to five matches (Taylor *et al.*, 2004). Given one match is 0.00843% of the data and consists of 1,463.9 ± 103.0 events, two random matches were tested. The reliability of the data was validated using Absolute and Relative measures of agreement (von Eye and von Eye, 2008; Hallgren, 2012; Eldridge *et al.*, 2013) and found to exhibit strong to perfect levels of agreement (Table 6) (McHugh, 2012). Cohen's Kappa was used to check Absolute Agreement, which corrects for agreement by chance (von Eye and von Eye, 2008; Hallgren, 2012), and Percentage Error was used to check for Relative Agreement, as similar to past research (Mitchell, 1979; Stemler, 2004; Eldridge *et al.*, 2013).

Table 6. Reliability Testing Results

Reliability	Variable	Kappa Value	P-Value	Relative Percentage Agreement
Inter-Reliability	Passing	0.921	<0.001	4.274%
	Defending	0.863	<0.001	14.563%
	Shooting	0.940	<0.001	3.030%
	Fouls	0.900	<0.001	11.905%
	Aerials	0.857	<0.001	9.524%
	Ground Duels	0.838	<0.001	10.435%
	Space Control	0.813	<0.001	0.272%
	Positions	1.000	<0.001	0.00%
Intra-Reliability	Passing	0.964	<0.001	2.142%
	Defending	0.912	<0.001	8.020%
	Shooting	1.000	<0.001	0.000%
	Fouls	0.942	<0.001	4.762%
	Aerials	0.892	<0.001	2.395%
	Ground Duels	0.880	<0.001	10.909%
	Space Control	0.826	<0.001	0.139%
	Positions	1.000	<0.001	0.000%

3.4 Data Analysis and Model Construction

Figure 3 describes the design of the Player Rating System (PLAYRS), which was designed in three phases. Furthermore, all calculations and analysis were performed using the R statistical computing software (Version 3.6.0).

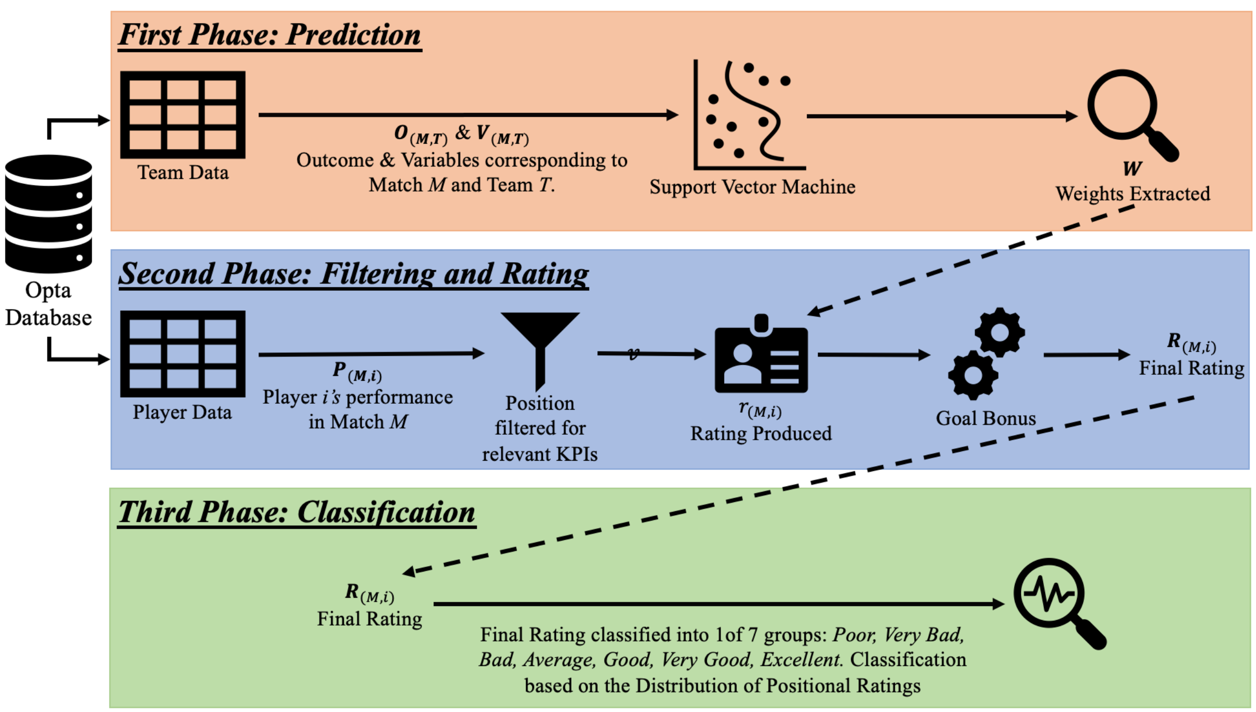

Figure 3. Player Rating System Model Design

3.4.1 First Phase: Prediction

In the first phase, each outcome, $\boldsymbol{O}_{(\boldsymbol{M},\boldsymbol{T})}$, for a match M, and team, T, was predicted based on the variables, $\boldsymbol{V}_{(\boldsymbol{M},\boldsymbol{T})}$ utilizing a Support Vector Machine (SVM). However, to contextualize each variable, v, instead of using the frequency within the game, each variable was normalized based on the amount in each game as follows:

$$V_{(M,T=1)} = \frac{\sum_{i=1}^{n} v_{(i,M,T=1)}}{\sum_{i=1}^{n} v_{(i,M,T=1)} + \sum_{j=1}^{k} v_{(j,M,T=2)}}$$

Therefore, each variable, besides Situation, was measured as its percentage within the game. Also, the outcome variable was predicted using the classification of Win and No Win as similar to past research (Herbinet, 2018; Pappalardo *et al.*, 2019). The predication model was a classification model utilizing a SVM, which has previously shown the highest level of accuracy relative to other ML models (Pappalardo *et al.*, 2019). A SVM is a discriminate classifier defined by a separating hyperplane (Zhang, 2012; Patel, 2017). Essentially, given a labeled data set, the algorithm will output an optimal hyperplane, a divider, which is able to categorize new data and tries to maximize the space between each point and the divider (see Figure 4 for examples).

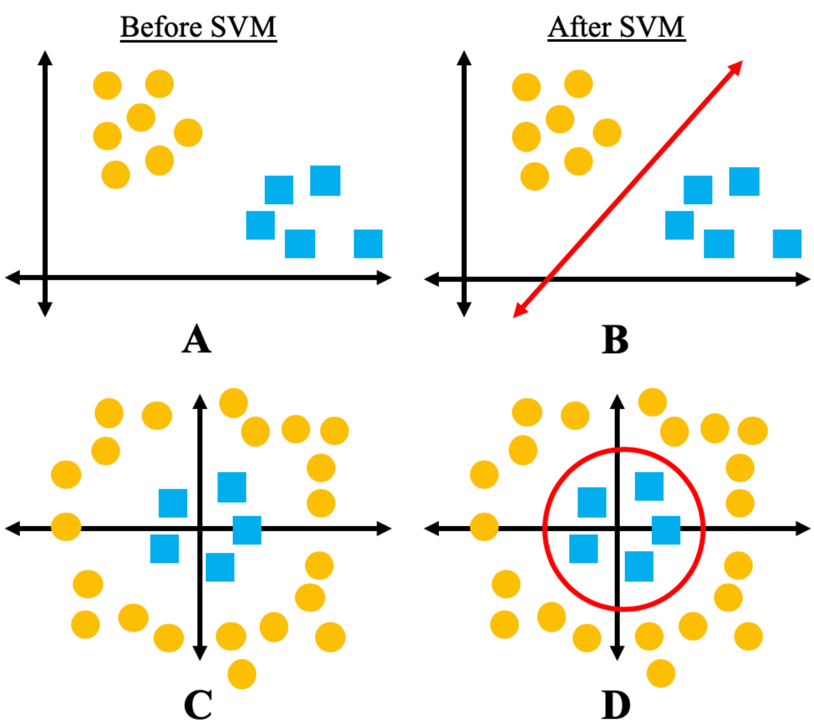

Figure 4. Example of How a Support Vector Machines Classify Data

Example of how a Support Vector Machine Classified Data with a Hyperplane, in red. For A→B the SVM is able to draw a simple linear hyperplane. For C→D the SVM can draw the hyperplane by transforming the data into a new dimension (not shown), then back into the original dimensions preserving the space.

These hyperplanes can be linear (**A→B**), radial (**C→D**), polynomial, or sigmoid (see Torgo, 2017 for a review). Brooks *et al*., (2016) and Pappalardo *et al*., (2019) use a linear classifier, but given the outcome of a game is based on numerous variables, a radial classifier is recommended (Hsu *et al.,* 2016). Furthermore, there are parameters within the radial SVM, Cost, C, and Gamma, γ, which must be optimized for classification (Eitrich and Lang, 2006). For large values of C, there are less training errors with a smaller margin between the hyperplane and data, while smaller values of C lead to more training errors with a larger margin. Gamma defines the influence of each datapoint on the hyperplane. For small values, datapoints further away have more influence, while for large values, datapoints very close to the hyperplane have more influence. Therefore, a range of values were tested with exponentially growing sequences of C, ($C = 2^{-5}, 2^{-3}, \dots, 2^{11}$), and γ, ($\gamma = 2^{-15}, 2^{-13}, \dots, 2^{3}$) as suggested by Hsu *et al*., (2016). Next to account for overfitting, the SVM utilized *K*-fold Cross-Validation (K-CV) (Refaeilzadeh *et al*., 2009). K-CV performs k different iterations of prediction by splitting the data into k equally sized folds, then trains the model on different combinations of $k - 1$ folds and tests on the leftover fold (James *et al.*, 2017). There is no formal rule for selecting the optimal value for k, but there is a bias-variance trade-off associated with the choice such that for small values of k, there is lots of variance with little bias and vice versa for large values of k (Fortmann-Roe, 2012). In this case $k = 5$ was implemented, but $k = 10$ would also have been acceptable given these values have been shown to yield test error rate estimates which do not suffer from high bias or variance (James *et al.*, 2017). Thirdly, given 45.6% of the dataset consisted of Home Wins and 29.2% were Away Wins, the *AUC* performance metric was evaluated for performance (Bunker and Thabtah, 2019). Lastly, given the possibility of league differences, individual models for each league were created, in addition to an all-inclusive model (*ALL-M*), to compare *AUC* and *Accuracy*, for a total of nine prediction models.

Finally, the SVMs were designed using the caret package (Version 6.0-84; Kuhn, *et al.,* 2019), while all graphics were made with the ggplot2 package (Version 3.1.1; Wickham *et al*., 2019), jrrnoldmisc (Version 0.2.0.9000; Arnold, 2019), MASS (Version 7.3-51.4; Venables and Ripley, 2002), ggrepel (Version 0.8.1; Slowikowski, 2019) and scales package (Version 1.0.0; Wickham, 2018) from R.

3.4.2 Second Phase: Filtering and Rating

Once each player's performance was extracted, $\boldsymbol{P}_{(M,i)} = [v_1, v_2, v_3, \dots, v_n]$, it was contextualized in the same manner as the team's performance. Next, the player's performance was filtered based on their relevant position. Before every match, Opta assigns each player a position within a formation, of which there are a total of 19 different positions and 23 formations (see Tables A.1 & A.2 in Appendix A), which can be surmised as Goalkeeper (GK), Full-Back (FB), Wing-Back (WB), Centre-Back (CB), Defensive Midfielder (DM), Central Midfielder (CM), Attacking Midfielder (AM), Winger (WG), and Forward (FW). However, research primarily identifies eight positions: GK, CB, FB, DM, CM, AM, Wide Midfielder (WM), and FW (Wiemeyer *et al.*, 2003; Taylor *et al.*, 2004; 2005; Hughes *et al.*, 2012). Given the disagreement, six positions were examined and rated given the relevant PIs (Table 7). GK were omitted given their vast specialization (Soto-Valero *et al.*, 2017) and this omission is present in other quantitative positional research (Taylor *et al.,* 2004; 2005; Dellal *et al.,* 2011).

Table 7. Position Matrix

PI	Full Back	Centre Back	Defensive Midfielder	Central Midfielder	Attacking/Wide Midfielder	Centre Forward
Successful Defensive Passes	x	x	x	x		
Unsuccessful Defensive Passes	x	x	x	x		
Successful Midfield Passes	x	x	x	x	x	x
Unsuccessful Midfield Passes	x	x	x	x	x	x
Successful Attacking Passes	x		x	x	x	x
Unsuccessful Attacking Passes	x		x	x	x	x
Successful Cross	x				x	x
Unsuccessful Cross	x				x	x
Successful Penalty Area Entries	x		x	x	x	x
Unsuccessful Penalty Area Entries	x		x	x	x	x
Successful Defensive Tackles	x	x	x	x		
Unsuccessful Defensive Tackles	x	x	x	x		
Successful Midfield Tackles	x	x	x	x	x	x
Unsuccessful Midfield Tackles	x	x	x	x	x	x
Successful Attacking Tackles			x	x	x	x
Unsuccessful Attacking Tackles			x	x	x	x
Defensive Interceptions	x	x	x	x		
Midfield Interceptions	x	x	x	x	x	x
Attacking Interceptions			x	x	x	x
Defensive Ball Recoveries	x	x	x	x		
Midfield Ball Recoveries	x	x	x	x	x	x
Attacking Ball Recoveries			x	x	x	x

Clearances	x	x	x			
Shots On-Target Inside the Box				x	x	x
Shots Off-Target Inside the Box				x	x	x
Shots On-Target Outside the Box				x	x	x
Shots Off-Target Outside the Box				x	x	x
Defensive Fouls Won	x	x	x	x		
Midfield Fouls Won	x	x	x	x	x	x
Attacking Fouls Won	x		x	x	x	x
Defensive Fouls Committed	x	x	x	x		
Midfield Fouls Committed	x	x	x	x	x	x
Attacking Fouls Committed			x	x	x	x
Yellow Cards	x	x	x	x	x	x
Red Cards	x	x	x	x	x	x
Successful Defensive Ground Duels	x	x	x	x	x	
Unsuccessful Defensive Ground Duels	x	x	x	x	x	
Successful Midfield Ground Duels	x	x	x	x	x	x
Unsuccessful Midfield Ground Duels	x	x	x	x	x	x
Successful Attacking Ground Duels	x		x	x	x	x
Unsuccessful Attacking Ground Duels	x		x	x	x	x
Successful Defensive Aerials	x	x	x	x	x	
Unsuccessful Defensive Aerials	x	x	x	x	x	
Successful Midfield Aerials	x	x	x	x	x	x
Unsuccessful Midfield Aerials	x	x	x	x	x	x
Successful Attacking Aerials		x	x	x	x	x
Unsuccessful Attacking Aerials		x	x	x	x	x
Space Control	x	x	x	x	x	x
Total KPIs Evaluated On	37	30	42	45	39	35

Furthermore, each location for players was based on Taylor *et al*., (2005) which showed Forwards performed 60% in the Attacking and 37% in the Middle Third, Midfielders 36%, 45%, and 19% in the Attacking, Middle, and Defensive Thirds respectively, and Defenders performed 44% in both the Defensive and Middle Thirds. Next, players were assigned a position based on formation and position by Opta, which was developed in consultation with the same Scouting Panel as before (Table 8).