Data Science Research Final

Grace Flitsch

December 15, 2022

Abstract

The low scoring nature of soccer has led to a lack of sophisticated analytics for the world's most popular sport. Recently, there has been a push for more specific and helpful statistics for fans and soccer clubs alike. I have compiled four papers from this field of research that I feel represent the work being done at this time. For each paper, I provide a summary and my personal assessment and recommendations. Through my research, I find that a major issue in soccer analytics is the lack of methods for valuing defensive or off-the-ball player actions. I propose ways to address this problem, and recommend a player vector of action values for each position on the field as the most valuable player statistic for soccer clubs and fans. This "threat vector" could be applied in game analysis, scouting, recruitment, and game preparation.

Contents

1 Introduction

1.1 Problem

One of the most well known applications of data science is to professional sports. Teams in sports like baseball, football, and basketball all rely on sophisticated statistical analysis for recruiting, strategy, and recommendations. Many sports have a large network of research to rely on to produce models that accurately rate players' ability, especially sports that are high scoring, have distinct positive and negative actions for every position, or are composed of discrete events. Given these criteria, soccer is a sport that is especially difficult to evaluate. Soccer is a continuous, low scoring game where actions like passing, dribbling, and defending do not always immediately result in exclusively positive or negative implications. In the past, statisticians have attempted little to no further analysis than assist and goal counting for soccer player valuation. There is a strong need in professional leagues to be able to rank recruits and current players. This could make for more accurate scouting reports, league rankings, player compensation, and player style classification. Recently, several papers have been published suggesting models to address this problem in soccer and other similarly low scoring sports. In this paper, I will summarize and evaluate some outstanding research that has been published for valuing and classifying player actions in sports and make a recommendation for the best machine learning model to apply for soccer evaluation.

1.2 Soccer Basics

In order to understand the content of the papers evaluated in my research, it is important to have a basic knowledge of soccer. Soccer is a sport played by 22 players, 11 on each team.
The positions are:

1. Goalkeeper
2. Right Full-back (or Wingback)
3. Left Full-back (or Wingback)
4. Center-back
5. Center back (or sweeper)
6. Defensive Midfielder
7. Right Midfielder (or Winger)
8. Center Midfielder
9. Center Forward (or Striker)
10. Attacking Midfielder (or Center Forward)
11. Left Midfielder (or Winger)

The only player permitted to touch the ball with their hands is the goalkeeper. Other players can use their feet, head, or chest to control the ball. There are three groups of positions; defenders, midfielders, and forwards. Forwards are the offensive players, midfielders play both offense and defense, while defensive players, as the name suggests, stay back and play defense. In a typical game, forwards will take most of the shots on goal, and the backs will not often cross the half-field line towards their own goal. Soccer games are generally low scoring, with teams scoring 1-2 goals per game on average. Each half of a game is 45 minutes, and the most common action taken is passing. The flow of the game involves teams setting up plays in the midfield, then moving forward to try to create an opportunity for a shot on goal. There are many possible penalties, some of which result in a free shot on goal by the offense. The most important points to understand for this paper are the difference between offensive and defensive players and positions, and what passes and shots are.

1.3 Historical Approaches

A common approach to player valuation in soccer has been rating players' total chance to make a goal in a game. This has been done in the past by simply adding up the players' historical goal contributions to games and their accuracy rate when shooting on goal, or by calculating the chance that any shot on goal taken by a player goes in, then adding together those probabilities for each game. Approaches along these lines have been common in soccer research, focusing on expected goals as a rating system.[1][2] This method forgoes considering any contribution except shots on goal, and sometimes assists. This is not a good representation of the game as a whole, as shots on goal make up a very small portion of the total on-the-ball actions taken during a game. It also results in defensive players receiving very low player ratings, as the amount of shots on goal by these players is much lower than offensive players. Another factor that these approaches neglect to consider is assigning negative values in any form. Soccer players often make mistakes that may lead to a goal for the other team. This is not addressed if it is not possible for a player to lose value once a positive action has been taken. Another common approach has been the introduction of context dependent valuation by splitting the field into certain zones, with players receiving varying ratings based on the previously decided difficulty of making a play in their respective zone.[8][7] This is a step towards the research done in the papers evaluated below, but it could be more specific to provide better context, as Decroos et. all (2019) discuss.[4]

1.4 Relevant Papers

In 2019 Decroos, Bransen, Haaren, and Davis published a paper, *Actions Speak Louder than Goals: Valuing Player Actions in Soccer*, that introduced research that attempted to address problems in soccer player valuation that previous research had neglected. In their model, they consider many possible player actions in many different contexts in order to assign to each player a value that represents the amount of goals they contribute per game on average. This method synthesized previous research in soccer and hockey analytics to create what was at the time one of the most sophisticated soccer valuation machine learning models to exist. There are new methods introduced in the paper, as well as a simple yet effective framework for describing player actions. We begin with this paper as it has been the foundation for many other papers that have built upon these authors' original research.[4]

Next, we will look at Liu and Shulte's research on player evaluation in ice hockey. Through my research, I have found that soccer and ice hockey are extremely similar sports that present the same problems when trying to evaluate players. *Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation* uses the Decroos et. all 2019 paper as a starting point to apply relevant research to ice hockey, as well as adding new methods and a different modeling technique to the conversation on player valuation for low scoring sports.[6]

The next paper that will be evaluated is *Measuring Football Players' On-the-Ball Contributions from Passes During Games.* This paper builds on the soccer analytics conversation by using a similar on-the-ball valuation idea as previous work, but only considering player actions that involve passing. Their aim is to focus on the most common action taken in soccer, passing, to value each action. Their algorithm is similar to that of Decroos et. all's 2019 paper, but bypasses some error in that paper by not valuing shots.[4]

Finally, I will look at a later paper by the same authors as the first paper. This paper, *Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams* follows up on some of the suggested research from *Actions Speak Louder than Goals: Valuing Player Actions in Soccer* and makes it possible for clubs to have some readily applicable information on players.[5]

1.5 Evaluation Criteria

The goal of this paper is to evaluate which, if any, existing research best addresses the problem of the lack of existing statistical analysis of soccer player's contribution to the total goals scored in a game. The criteria on which the papers will be assessed include ease of implementation, complexity of the algorithm, accessibility of data, relevance to soccer, ability to adapt and evaluate other sports, and accuracy. A model must be easy to implement to be truly accessible to the statistics community, if it requires excessive data or time to implement it is not an effective way for soccer clubs to spend their money. The algorithm should not be extremely complicated so that it is intelligible to a common data scientist. This allows for any club to have access to the same resources and implement the model in a reasonable amount of time. A confusing and murky algorithm limits the ability to tweak and change certain elements to better fit a team's specific analytics desires. It is essential that these models are relevant to soccer, as the goal of this paper is to find the best model for evaluating soccer specifically. Research in other sports is included in my collection of papers, so the ability for a paper to be applied to soccer is an important component of research evaluation. Conversely, if a model is created for soccer, it may be useful in other sports, like hockey, so its ability to adapt to other input data types is relevant. Finally, a good method must accurately describe the correct outcome for its inputs. For example, the correct ranking of professional soccer players according to the real-world rankings of those players must be observed in a paper that aims to rank players.

2 Actions Speak Louder than Goals: Valuing Player Actions in Soccer

2.1 Summary of Research

2.1.1 Research Gap

This paper addresses the previously mentioned problem of the lack of helpful statistics for soccer player valuation. The authors attempt to create a model that scores players' impact on the field by assigning a value to each player based on goals contributed per game. The research in this paper is in high demand, as this score would allow coaches to have more accurate scouting reports, and possibly introduce the opportunity to classify players and teams by playing style. The data comes from the English Premier League's 2016/2017 season. The model has applications in professional soccer, but may be harder to implement in lower-level play because of the detailed statistics it needs as input to the model.

2.1.2 SPADL

An important, non mathematical development of this paper is the introduction of SPADL, "a powerful but flexible language for representing player actions."[4] SPADL stands for Soccer Player Action Description Language. The authors created this language to address the need for a system of discussing soccer that is human interpretable, simple, and complete. The nine attributes of SPADL are[5]:

- StartTime: The exact timestamp for when the action started
- EndTime: The exact timestamp for when the action ended
- StartLocation: The (x,y) location where the action started
- EndLocation: The (x,y) location where the action ended
- Player: The player who performed the action
- Team: The team of the player
- Type: The type of action
- BodyPart: The body part used by the player for the action
- Result: The result of the action

There are twenty one possible action types, each being performed by one of four options for body parts, ending in six possible results. Each of these actions can be included in an action set, and games are composed of action sets. Each action set accounts for all actions taken between two consecutive touches of the ball. SPADL also attempts to unify other existing languages into one common vocabulary.[4]

2.1.3 HATTRICS

The HATTRICS (Honest Attribution of Credit in Soccer) structure works by taking the difference between predicted chances of a goal (or conceding a goal) before and after an action has taken place, and assigning that value to the player who took the action. To estimate the probability of a home goal or visiting goal, the process is the same. If $S_i =< A_1, ..., A_i >$ represents a game state, and F_i^k is the next consecutive action set, these probabilities are calculated as:

$$P_{hg}(S_i) = P(hg \in F_i^k | S_i)$$

$$P_{vg}(S_i) = P(vg \in F_i^k | S_i)$$

Thus, change in probability is calculated from $P_g(S_i) - P_g(S_{i-1})$ The paper also introduces further small notes for dealing with some technicalities of soccer, but those aren't relevant to the goals of my research. The HATTRICS section leads into the HATTRICS-OTB section, which explains the algorithm for combining and assigning all of these values to each player in order to give them an overall score.

2.1.4 HATTRICS-OTB

HATTRICS-OTB (Honest Attribution of Credit in Soccer for On-the-Ball Actions) uses play-by-play data so that it's already split into the necessary subsets of actions between ball touches. The authors aim to:

1. Transform the stream of actions into a feature-vector format,
2. Select and train a probabilistic classifier,
3. Aggregate the individual action values to arrive at a rating for a player.

To prepare the model, the authors require that each game state is approximated by the previous three action sets and the following ten. A positive outcome (goal) in the next ten actions gives a positive value to the current game state. This allows there to be a fixed number of input features to the algorithm and for the immediate (i.e. most important) context to be considered with heavier weight. The authors show later on that three previous actions steps is the most desirable number of actions to be considered for the model to perform well. Each action set should only contain one action, since they are split by touches on the ball. This allows the types of action sets to be split according to SPADL's language of features. There are three subgroups of features that each action has;

1. Those explicitly described by SPADL, including both discrete and continuous feature types.
2. Complex features created by aggregating information from past actions and various SPADL features.
3. Overall game features such as total goals scored at that point, goal difference, goals by other team.

Next, the authors select a machine learning model to use. They trained the options with data and output from the 2012/13 and 2014/15 season to predict output in the 2015/16 season. The three models tested were Logistic Regression, Random Forest, and Neural Network. The Random Forest model ended up being the most accurate compared to the perfect calibration. As mentioned previously, they trained the Random Forest classifier using varying numbers of previous actions to consider, and considering three previous actions resulted in the least logarithmic loss.

Finally, the authors conclude this section by assigning a rating to players. The model assigns values to individual player actions, then can be aggregated to assign those values to the player that took the action, and sum the actions over various desired timeframes. To account for players' difference in time on the pitch, the player values are assigned over a ratio of the actual time played, then multiplied into a 90-minute player value estimation. To account for differences in position, the authors recommend comparing players within the same position rather than across positions. The authors then introduce the topic of their next paper, which will be discussed later. They suggest summing all player values for each action type they take to investigate which action they prove to add more value to the game with when they take it.

2.1.5 Results

The authors discuss various results and use cases at the end of the paper to show the many possible applications of their model. They observe that in the 2012/13 dataset that they used, 54% of the actions were passes, 24% were dribbles, and only 1.4% were shots, with just 11% of those shots going in. This shows how important passing and dribbling is to the game, which supports Bransen and

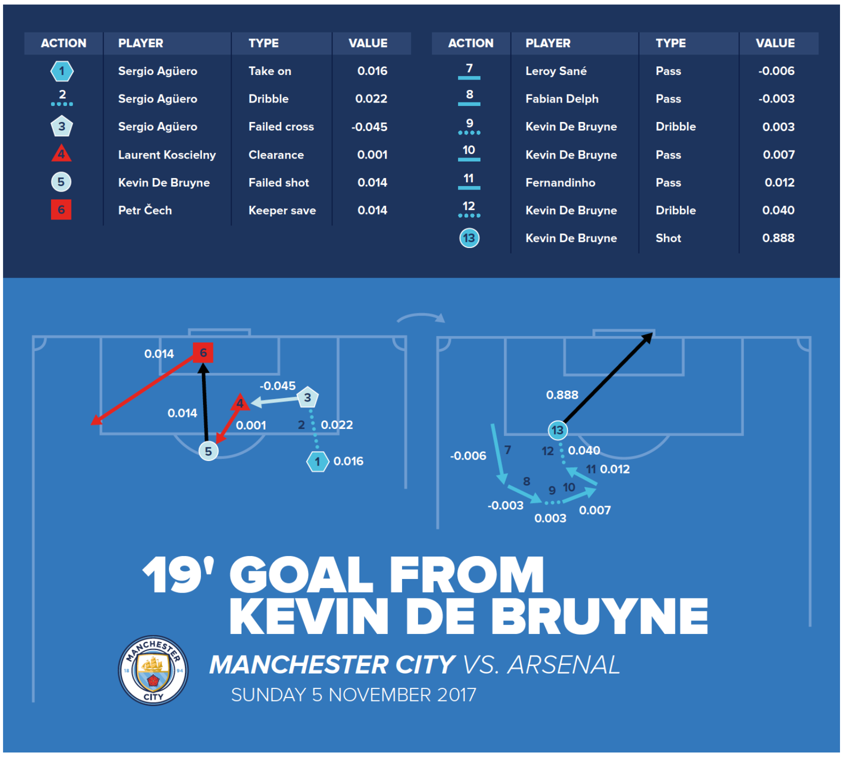


Figure 1: Visualization of De Bruyne's goal from 2017 [4]

Van Haaren's idea to assign values solely based on passing success.[3] The authors then dive into a specific use case that demonstrates the correctness of their model. In Figure 1, the action types and features described by SPADL can be seen in action, explaining why each value was assigned to each action. Another way that the authors prove the viability of their model is by predicting the best performers from a season, then comparing this to real rankings. The plot in Figure 2 shows that Messi's ratio of value per action to actions taken is very high, which supports the fact that he is known as the best player in professional soccer. The authors then use the model to select the best line-up based on the 2016/17 season data. They choose the best player from each position, and compile an eleven player roster. Some of the players selected are obvious choices, while others are more surprising. The surprising choices seem to come out more often in defense, since the algorithm neglects to account for off-the-ball actions, which is what defense is all about. It also does not give goalies the most accurate value, since it gives them points for saves while not detracting them for conceded goals. The paper then concludes with more specific applications described, including identification of young talent, players who stand out, characterizing playing style, and team play

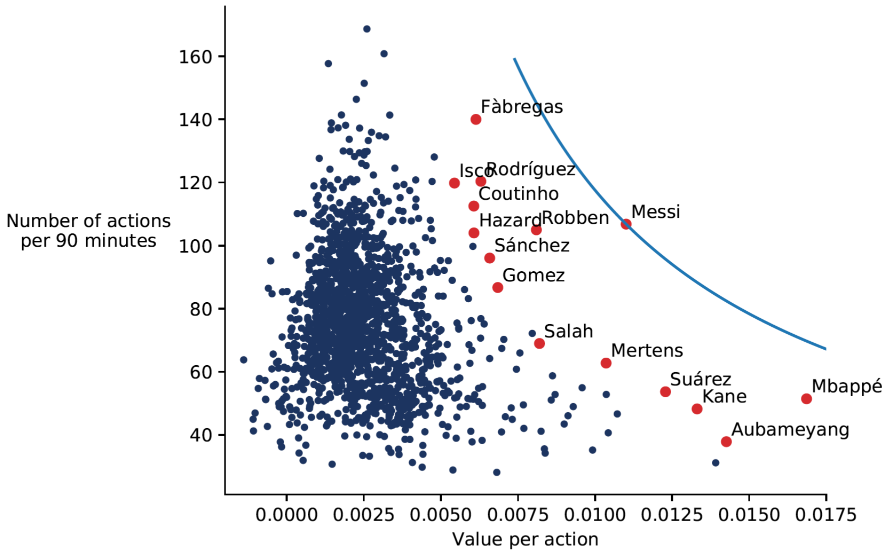


Figure 2: Plot of professional player's average number of actions in 90 minutes vs value per action from the 2016/17 season. [4]

style evolution.

2.2 Assessment of Actions Speak Louder than Goals: Valuing Player Actions in Soccer

In this section, I will evaluate Decroos et. all's model according to the criteria in section 1.5. This model provides the groundwork for much of the research that is happening in the soccer analytics field right now. It synthesizes ideas that existed in other sports and contexts and streamlines a singular approach to player valuation.

2.2.1 Complexity of the Algorithm

Firstly, the algorithm described by HATTRICS-OTB is relatively easy for someone with a math background to understand. This is important to the overall model since the goal is for soccer clubs to use the model themselves. If it isn't easily understood by the club's statisticians, it's not realistic that they could boil it down and make it comprehensible to non-industry coaches and players. A Random Forest probabilistic classifier is a common machine learning model to employ, and most data scientists are familiar with the math behind it. There are also packages that exist to build

Random Forests, so the user doesn't have to fully understand the model to employ it.

2.2.2 Necessary Data

Next, an issue I have with this model is that it requires extremely complex data as input to the model. This is not an issue if one is evaluating professional players, as play-by-play data is widely available online and can be easily transferred into SPADL for input to the model, as described by the authors. The problem with the data for the model is that it makes it difficult to implement, nearly impossible, for a team that does not immediately have access to their play-by-play data. These teams may be non-professional, or not as popular as those in the European League. In order to use the model, the data scientists would have to spend copious amounts of time going through games and filling out nine fields in SPADL for each touch of the ball. This is not reasonable for any team to spend time on, since gathering enough data to create a meaningful model would require about ten games worth of player actions. This is not desireable because the model inherently favors teams with more sponsorship and support, and does not assist teams who are already lacking more than others in statistical analysis.

2.2.3 Applications to Other Sports

Next, I find that this model is extremely relevant to soccer, as that is the sport that the paper is aimed at. The model is made to address statistical analysis in soccer, but is built on models from hockey and basketball as well as soccer, and has been applied to many sports. A strength of the model is that other researchers have found it not extremely difficult to transfer the logic of HATTRICS-OTB to other sports, specifically hockey. This makes the model far more valuable, since it is not only applicable to a niche of the sports industry.

2.2.4 Accuracy of HATTRICS-OTB

Finally, the model's accuracy varies from position to position. It is extremely effective in assigning values to offensive plays, but struggles when it comes to defense. This is my main problem with the model. A method that claims to value players in soccer should not neglect to accurately value half of the players on the field. In the results of the paper, the authors gloss over this weakness by mentioning some "surprising" players who are ranked among the best, but don't necessarily seem that good from their actual play. Rather than these players being great defensive talent that has been overlooked, I believe that these results show the lack of accuracy in ranking defensive players based on HATTRICS-OTB player valuation. This is because good defensive plays often happen off-the-ball. A player positioning themselves correctly can shut down an entire offensive play even if the player doesn't touch the ball. A defender could also challenge an offensive opponent while they are dribbling, ensuring that they can't move closer to a goal. In this model, the defender would not get credit for either of these plays, even though both would decrease the chances of a goal being an element of the following ten action sets. This lack of defensive valuation means that comparisons between players in different positions is not possible, since offensive players are far more likely to have a higher valuation. Another issue with the paper in general is a lack of mathematical assessment of the model. The authors compare their results to European soccer rankings, but not to any competing models. They do not inform the reader how well it performs in comparison to other research in the market, so there is no context or numerical value to evaluate it by. Subjectively saying that it is a good model because it names some talented players as top players does not mean that it is a tried and true trustworthy model. More assessment of the quantitative output needs to be done for the authors to draw a conclusion.

2.2.5 Recommendation

Overall, this model is a good start, but has key weakness that would lead me to not suggest the application of the model to a professional team for scouting or team improvement purposes. The ideas in the model are a good start, but not enough to be useful unless applied to only offensive players. Since this paper was published, many authors have built on its research. In the following sections, I will evaluate that future research in three other papers that attempt to create a less errant tactic for player valuation and characterization.

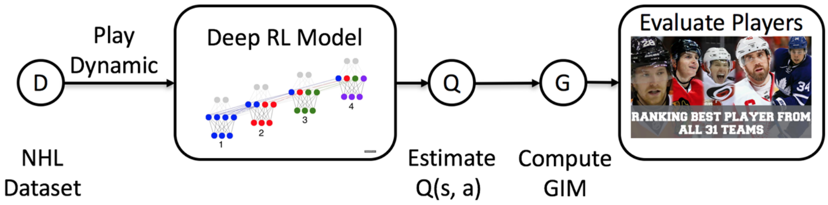


Figure 3: Visualization of the process for hockey player evaluation [6]

3 Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation

3.1 Summary of Research

3.1.1 Context

In researching the papers that model low scoring sports, I have found similarities between soccer and ice hockey. In both sports, the path of the puck or ball is similar, and passing is the most common action. While there are less players on the ice in hockey, there are similar positions in both sports. The goals of attacking and defending are also very similar, hence the common low scoring nature between the two. The shape of the field and ice are similar, and both games enforce offsides rules. Some differences include that hockey is a more stop-and-go sport, with each team playing everyone on the roster consistently and subbing often. Hockey is also a higher contact sport that involves more intermingling of offense and defense. All players in hockey have a higher chance of scoring given the smaller size of the ice compared to the pitch, and having less players in at one time.

3.1.2 Methodology

This paper has similar goals to the last, working to make a model to value player actions in ice hockey in order to evaluate players. The methodology can be visualized in Figure 3. The authors introduce the Goal Impact Metric, which values each player based on "aggregated impact of their actions."[6] The data used in this paper is composed from computer vision, which implements optical spacing, as was recommended by Decroos et. all in their 2019 paper. This paper considers the reward (home goal, visiting goal, or neither) of a possible thirteen action types. There are ten features evaluated for each action, shown in Figure 4. The games evaluated are split into subsections that begin with

Name	Type	Range
X Coordinate of Puck	Continuous	[-100, 100]
Y Coordinate of Puck	Continuous	[-42.5, 42.5]
Velocity of Puck	Continuous	(-inf, +inf)
Game Time Remain	Continuous	[0, 3600]
Score Differential	Discrete	(-inf, +inf)
Manpower Situation	Discrete	{EV, SH, PP}
Event Duration	Continuous	[0, +inf)
Action Outcome	Discrete	{successful, failure}
Angle between puck and goal	Continuous	[-3.14, 3.14]
Home or Away Team	Discrete	{Home, Away}

Figure 4: The features considered for each action for ice hockey player evaluation. [6]

the last goal scored, or the beginning of the game for the first episode. They then evaluate the Q-function:

$$Q_{team}(s, a) = P(goal_{team} = 1 | s_t = s, a_t = a)$$

which calculates the probability that a goal is scored at the end of the current section of the game by "team" (away or home). [6] This Q-Function is selected by the authors because

1. It reveals goal expectation regardless of the spread in score, meaning that a coach could know what performance to expect out of their players even if they are winning or losing by a significant amount. This means that play that is above or below potential can be identified.

2. It is an easily interpretable output, since it predicts upcoming events rather than the entire match outcome.

3. It captures offensive and defensive aspects, since a defensive stop is considered and adds to the probability that the defenses' team will score a goal.

The authors model the Q-function with a function approximation approach given by a neural network. The paper provides more detail on the specifics of the model, but the most important factors are that the output nodes are estimates of the predicted value of the Q-functions for home, away, and neither scoring a goal. The neural network is built with 5 layers, 3 of them hidden. The model is sketched in Figure 5. In the next section, the authors describe their extremely complex deep learning algorithm for making predictions of future goals. The most important parts of this explanation are that the model accounts for change in possession, the time in the play, and that the score probability of one team is sensitive to the score probability of the other.

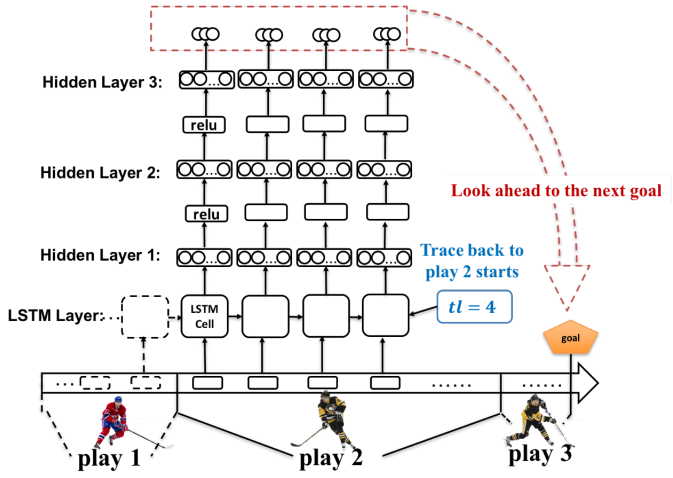


Figure 5: The neural network model for hockey player valuation. [6]

3.1.3 Player Evaluation

Similarly to Decroos et. all, the authors measure the value of an action and then apply that impact to the player's overall value for that game. The scoring probability as a whole represents the chances for the team to score, while the difference between the previous and current scoring probability after an individual action takes place applies to the player that took that action. The goal of compiling these values is to constrict the Goal Impact Metric (GIM) for each player. The below equation shows how the GIM is calculated [6]:

$$impact^{team}(s_t, a_t) = Q^{team}(s_t, a_t) - Q^{team}(s_{t-1}, a_{t-1})$$

$$GIM^i(D) = \sum_{s,a} n_D^i(s, a) \times impact^{team_i}(s, a)$$

3.1.4 Results

The authors then list the top 20 NHL players according to their GIM. They show that the model can not only rank players, but also value certain action types of each player. Other applications include analyzing pay of players and identifying players who are being paid less than the model suggests they are worth. The authors then evaluate their model by comparing the output to the Plus-Minus, Goal Above Replacement, Win Above Replacement, Expected Goal, and Scoring Impact methods. These methods are defined as:

1. Plus-Minus: Measures how presence of a player influences the goals of his team.
2. Goal-Above-Replacement: Estimates the difference of team's scoring chances when the target player plays vs replacing them with an average player.
3. Win-Above-Replacement: Same as above but with regards to winning chances.
4. Expected Goal: Values each shot by the chance it leads to a goal.
5. Scoring Impact: Learned Q-values with spacial regions and game time as inputs.

The authors then performed a significance test to test the difference between these metrics and GIM, and found that GIM was very different from other metrics. The authors go on to find the correlation between ranking metrics and success measures, predict future performance based on past performance, and predict players' salary for future seasons, all based on the GIM. The paper concludes by summarizing the model and its purpose, then suggesting future research.

3.2 Assessment of Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation

I found this paper to be a more sophisticated approach to player valuation than the previous. It builds upon a similar idea of finding the value of each play, then assigning that value to the player that made the play. The difference between the papers is that this paper builds a much more complex model that requires deep learning and does not rely on future plays to assign values. Given the more complex approach, the model is much more difficult to understand and implement than Decroos et. all (2019).

3.2.1 Complexity of the Algorithm

This is the first problem with this model- it has a high barrier to understanding, requiring knowledge of many aspect of data science, deep learning, and machine learning to even read the paper. A typical sports analyst would have to spend significant time working to understand the math in order to implement this model. The implementation would also be difficult because of the detail needed for the input data. A positive to the complexity of the algorithm is that it is able to make predictions without future action sets as was done in the previous article.

3.2.2 Necessary Data

Similarly to the last paper, each play is described by many features and possible action types. The definition of a play is a bit broader in the hockey context, so there are less plays to examine, but still too many for one person to realistically compile the data. Sample data is available from the NHL, but again this means it is not available to more minor teams.

3.2.3 Applications to Other Sports and Accuracy

This algorithm has the potential to be applied to other sports, but the authors did not make an effort to encourage this. A few factors are more specified to hockey, but I believe that this algorithm could be generalized without too much issue. The accuracy of this model seems to be high according to the paper. It recommends the expected players as the top 20 in the NHL, and performs well compared to other models in the field. These predictions prove to be very accurate compared to other modeling techniques, and the authors did exceptionally well in assessing their model compared to other common models. This makes it easier to know which model for the reader to select and implement. The paper doesn't directly address the major problem of defensive undervaluing mainly because this is less of a problem given the nature of hockey. Players are more likely to play the whole ice, which is different from play in soccer. The model doesn't do anything to account for position of the player, which is acceptable for hockey, but still leaves the problem to be solved in soccer.

3.2.4 Recommendation

In a business context, I would recommend this model depending on the time and statistical knowledge available to my team. Since it is such a sophisticated model, I would be hesitant to recommend it to a team with a low budget, as the authors mention a high computing time, and a lot of time would be put into data prep and understanding the model. If a professional team with ample resources had the opportunity to use a model, I would recommend this one because of the end accuracy. I may also be hesitant to recommend the transfer of this model to soccer applications since soccer involves more players and plays, which would only add to the complexity and computing speed problem. This is a good model, but presents some issues that may make it inaccessible to many.

4 Measuring Football Players' On-the-Ball Contributions from Passes During Games

4.1 Research Summary

4.1.1 Setup

In this paper, the authors aim to assign values to players, but to only generate those values based on passes. This methodology partially addresses an issue introduced earlier, that defensive players are harder to evaluate. Since all players are more similarly likely to pass the ball, the playing field for earning value is evened a bit, but not all the way. Actions like dribbling and shooting are not included in this framework, so the data collection and modeling is a little simpler.

Just like the other models we have looked at so far, the authors of this paper choose to value actions by computing the difference in the probability of a future goal before and after an action takes place. A small difference in the assignment of value across sequences is that the value given is not the actual reward of the goal but rather the "number of goals expected to arise from a given pass if that pass were repeated many times."[3] The process to measure expected contributions from passes is broken down into four steps; constructing possession sequences, labeling possession sequences, valuing passes, and finally rating players.

4.1.2 Constructing and Labeling Possession Sequences

Possession sequences are a more broad encapsulation of player actions than we have seen before. Each sequence consists of what one would normally think of as a "possession." These sequences begin at the start of the game and can be ended, thus start a new one, when the ball changes possession for any reason. This could be due to the loss of the ball to the defense during play, a scored goal, a foul, or the ball going out of bounds. Each possession sequence is labeled by its expected reward. As can be predicted, the reward across a sequence that does not end in a shot is zero. A sequence that does end in a shot is given the expected-goals value as defined above. Each shot is represented by its (x,y) location, distance from the goal, and the angle between the ball and the goal posts.

	Train set	Validation set	Test set
Games	4,253	2,404	2,404
Possession sequences	1,878,593	972,526	970,303
Passes	3,425,285	1,998,533	2,023,730
Shots	95,381	53,617	54,311
Goals	9,853	5,868	5,762

Figure 6: The characteristics of the data used. [3]

4.1.3 Valuing Passes

Each possession sequence is further split into possession subsequences. The subsequences consist first of just one pass, then each following subsequence is equivalent to the previous plus the value of the next pass. A previously mentioned, passes receive a value based on the difference between the expected reward of the possession subsequence after and before the pass. An assumption made here is that teams can only earn reward when they have the ball. The authors then use a k-nearest neighbors search to "average the labels of the k most-similar possession subsequences."[3] They then use Dynamic Time Warping and interpolation on the x and y coordinates to account for field position context. They split the field into cells and cluster all possible origin-destination pairs (where the pass could begin and end) and then assign each possession subsequence to one cluster. This allows for the assigning of values to each pass.

4.1.4 Rating Players

Rating each player is very simple after each pass has been valued. Pass values can be summed over any desired period of time to obtain a player evaluation score, then normalizing it over a 90-minute game period.

4.1.5 Results and Experimental Evaluation

The authors train and test their data on games from the 2014, 2015, 2016, 2017, and 2018 seasons. Figure 6 shows a breakdown of the data. In this section, the paper examines in more detail the methodology outlined in the previous subsection. The exact mathematics used are not directly helpful to this paper, so I omit the most complex computations for the sake of highlighting what is relevant. The authors found that clustering possession sequences by (x, y) location was the best

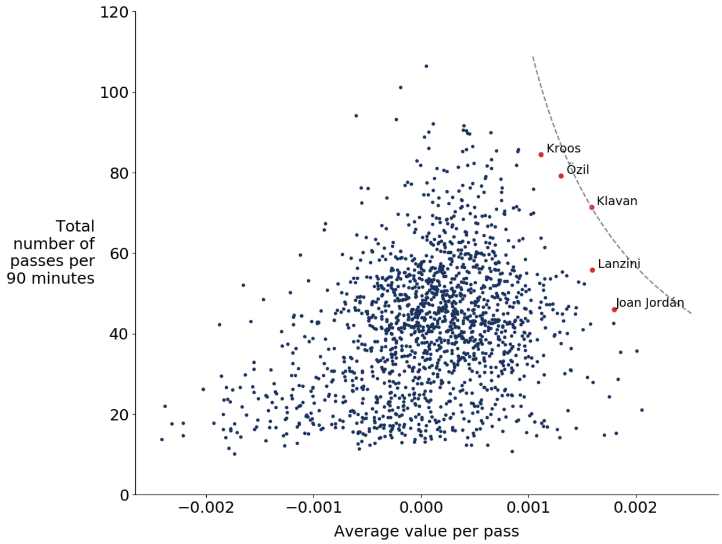


Figure 7: A plot of the players in the data, showing their average value per pass vs the average number of passes per 90 minutes. [3]

way to cut down on costly computing time. They tested various amounts of clusters vs no clusters and found that clusters significantly reduced the runtime. They then ran a similar experiment to find the optimal amount of k nearest neighbors, and found that setting k to ten was most desireable. The authors then tested their model, comparing it to two common approaches; pass accuracy and prior distribution over possible game outcomes. The approach described in this paper out-preforms both other approaches.

The authors then present their results, which is a list of the top ranked players under 21. They observe that goalkeepers rank highly as far as completed passes, since they aren't often contested. They also note that midfielders make the most passes per game. Figure 7 shows that the model can identify talented players who make a significant amount of passes while still retaining a high average pass value. They then go on to demonstrate a use case in which a player needs to be replaced by a similar player.

To conclude, the authors summarize their work, then introduce the shortcomings they hope to address in future research. These include accounting for the strength of the opponent, taking more context into consideration, and trying different mathematical modeling methods.

4.2 Assessment of Measuring Football Players' On-the-Ball Contributions from Passes During Games

This model builds on both the Decroos and Liu models and adds important new insights to player valuation[4][6]. It also addresses some shortcomings that I mentioned in my assessment of both previous models. Specifically, it decreases the input data needed, evens out values across positions, and seems to be accurate. Unfortunately, some of the solutions led to more problems.

4.2.1 Complexity of the Algorithm

Similarly to Liu's model [6], the modeling technique is complex. This again creates a barrier to use for a soccer club that may not have a full data science team. The mathematics seem to create an accurate model, but it is hard to tell since the authors do not provide an extensive testing and verification section. The authors could have perhaps provided a more in depth explanation of the machine learning applied, but again, this means that statisticians attempting to implement this model would have to commit quite a but of time to understanding the model. If the team working on implementing a model has time, or a strong data science background, then this would not be an issue.

4.2.2 Necessary Data

An improvement that this model makes to previous ones is that it requires less input data, since the only actions that the model looks at are passes. It is surprising, but the benefits of using less data are twofold to the consumer of this paper. It removes an element of unfairness between the way offense and defense is evaluated. Since the offense cannot gain extra points for shots, and the defense is credited if they are involved in the lead-up to a goal, value is more evenly distributed.

4.2.3 Application to Other Sports

Another downside that I find with this research is that it is less easily applied to other sports, since soccer uniquely relies on passing from all players. The research could possibly be generalized to be used in ice hockey, but shots happen more often in hockey, so omitting credit for them may cause more of a loss of information than in soccer.

4.2.4 Accuracy

Finally, looking at accuracy, the model seems to perform well enough to make productive recommendations. The authors presented several use cases, none of which hold solid proof of concept, but it seems to perform at least as well as the other two models, if not better in the case of defense. If I were asked to provide a recommendation on this model alone, I would say that it is worth implementing if you want to perform comparisons across a full roster.

4.2.5 Recommendation

Although the idea of only valuing passes sounds more simple, this research is actually among the more complex soccer evaluation methods. Since Liu and Schulte's model already works well enough for valuing both offense and defense in hockey, I would not recommend that a hockey team use this model. The most important development made was the evening of scores between offense and defense, so it is not important to apply this model to a sport where that is not a problem. If the business or soccer team that I worked with did not have many resources for statistics, I may also recommend staying away from the model since from what I can tell it is quite complex. For a simple valuation of a single player, the original Decroos paper still seems to be the frontrunner for understanding the model. As I said earlier, I would also not recommend this model to other sports, since it has a heavy reliance on specific, unique aspects of soccer.

5 Player vectors: Characterizing Soccer Players' Playing Style from Match Event Streams

5.1 Research Summary

5.1.1 Applications of Player Evaluation

This paper differs from the previous three in that it dives into a use case previously suggested for further research, and uses player values obtained through similar methods to apply machine learning to concrete problems that soccer clubs have. The specific problem addressed here is player characterization to reduce scouting mistakes. Scouting in the past has been subjective and performed by humans who are susceptible to bias and confusion. The authors aim to remove human bias from the problem and create a model that can characterize players simply using raw input data. The authors discuss how raw data does not reveal much about actual players' ability, and can be wrongly interpreted when left up to human intuition. The authors present three reasons why characterizing player style is important to the research field:

1. Scouting: Being able to typecast and typematch players could simplify scouting for clubs.

2. Monitoring Player Development: The research in this paper outputs a human interpretable player-vector that a coach can inspect to suggest play style changes, or encourage the continuation of habits.

3. Match Preparation: Clubs can use characterization to prepare certain defense and offensive attacks that work better against specific play styles.

The authors conclude the introduction by explaining that the goal of the paper is to characterize players by analyzing event stream or play-by-play data. They also define playing style as "manifest[ing] itself as where on the pitch a player tends to perform specific actions on the ball."[4]

5.1.2 Data and Challenges

The data used is in the "event stream" style. This means that it specifies the time and location of events as they occur. The data is formatted in the SPADL style, described in detail in the second section of this paper.[5] This paper relies heavily on SPADL to describe events, which is how it relates to Decroos' original research from 2019.[5]

The authors then introduce four challenges that they face as they attempt to meet the goals of their research:

1. Players rarely repeat the same action in the exact same location, making it difficult to characterize plays in similar locations.

2. Actions have both continuous and discrete features, which doesn't work for most machine learning algorithms.

3. The size of the data varies, which also does not work well with most machine learning algorithms.

4. Off-the-ball actions are not considered even though off-the-ball actions often characterize playing style.

5.1.3 Assumptions and Definitions

In the third section of their report, the authors assume that a player's playing style arises from the interplay between his skills and the tactics employed by the team. This means that the playing style will show in the player's behavior during the game.
The first term defined is playing style: "A player's playing style can be characterized by his preferred area(s) on the field to occupy and which actions he tends to perform in each of these locations."[5]
Two further assumptions are made to create a backdrop for the model:

1. Most players have different playing styles and can be characterized by these different styles.

2. A player's playing style will remain consistent over a reasonable amount of time.

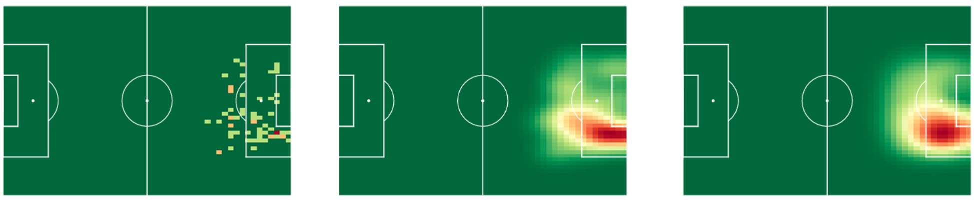
Given these assumptions and the definition, once the model is made, the authors predict that player identities can be retrieved from anonymous data.

5.1.4 Methodology

In this section, the authors build a model that can return a human-interpretable "player vector" that accurately characterizes the player's style. The methodology for this goal is such: select relevant action types, overlay a grid on the field and count the times that each type of action occurs in each location then input this into a matrix, reshape this matrix into a vector to be combined with all other vectors of the same action type, and finally construct the player vector by concatenating all compressed vectors of each action type.[5]
The first action item is to select action types that are relevant to player characterization. Only offensive actions will be selected due to the weak nature of the defensive analysis in SPADL. The action types that are performed based on position on the field are also removed (eg throw-ins, free kicks) because specialists typically do these actions, which means that it more depends on position of the player than personal play style. The remaining types to be considered are passes, dribbles, crosses, and shots. Figure 8 shows both the eliminated and selected player actions for this model.
After selecting the action types, the authors construct heat maps to indicate where player p performs actions of type t.[5] Constructing the heat maps involves three steps: counting the actions in each cell of the grid, normalizing over amount of minutes player, smoothing areas with high granularity in the grid. Figure 9 shows this process in images. The authors next took these heatmaps and compressed them to produce small player vectors. The process here is complex, but follows the

Action type	Frequency	Offensive	Open play
pass	53.1%	✓	✓
dribble	25.2%	✓	✓
clearance	3.8%		✓
throw_in	2.8%	✓	
interception	2.6%		✓
tackle	2.3%		✓
cross	1.8%	✓	✓
shot	1.5%	✓	✓
bad_touch	1.4%		✓
foul	1.3%		✓
freekick_short	1.3%	✓	
keeper_pick_up	0.8%		✓
keeper_save	0.8%		✓
corner_crossed	0.6%	✓	
freekick_crossed	0.2%	✓	
keeper_claim	0.2%		✓
corner_short	0.1%	✓	
shot_freekick	0.1%	✓	
keeper_punch	0.1%		✓

Figure 8: Possible player actions to be considered, with chosen actions in bold. [5]



(a) The raw shot heatmap obtained by overlaying a grid and counting shot locations.

(b) The smoothed shot heatmap obtained after normalizing and smoothing the raw shot heatmap.

(c) The same shot heatmap reconstructed from a length-4 feature vector.

Figure 9: A heatmap undergoing smoothing and reconstruction. [5]

basic outline of constructing individual vectors from subsection 5.1.4. Similarly, player vectors are constructed according to previously mentioned steps. The authors note that differences between player vectors can be found by computing the "Manhattan distance," or the distance between two points measured along axes at right angles, between them.

5.1.5 Results

Since there is no "objective ground truth" for player characterization, analyzing correctness is difficult.[5] The authors investigate intuitive outcomes, claims in media about players, reports of player development, and anonymous player retrieval to evaluate their model.

The authors selected various players of different positions and created player vectors for them. In each case the vector matched assumed information about the player given their position, then provided additional insight, mostly about position on the field that actions take place.

Given claims in the media about players that have similar characteristics, the authors tested the output of the player vectors. The model performed exceptionally well, assigning players who are hailed as similar according to soccer fans very similar vectors. Offensive players who are similar to the naked eye ended up being ranked as extremely similar by the model. The model was less accurate for defensive players, but this is an anticipated weakness.

Next, the authors produced player vectors from different years for Cristiano Ronaldo, who many claim has evolved over the years. The vectors confirmed two very different styles of play, showing that they can reveal changes in playing style for individual players.

Finally, the paper addresses the anonymous player retrieval problem. This is the most unbiased test of the model. The model is trained on data where the player is not hidden, then tested by producing a ranking of the most similar players to the anonymous given player vector, with ranking representing confidence in the player chosen to match the vector. The model was able to successfully retrieve 38.2% of the players with one attempt, and 64.4% of the players within the top ten estimation.

5.2 Assessment of Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams

The nature of this research is quite different than others I have evaluated, but my evaluation criteria will remain the same. I feel that this paper is relevant to the player evaluation conversation even though it evaluates qualitative characteristics of gameplay rather than quantitative. It is interesting to see a different approach to soccer analytics than valuing players based on expected goals, or some similar statistic.

5.2.1 Complexity of the Algorithm

The algorithm presented in this paper is not overly complex. There are some more difficult parts to understand, but I do not think that a company or team would have to devote a costly amount of time to learning the algorithm. I find the transition from heat maps to matrices then vectors to be a simple yet effective methodology. The authors stated that part of their goal was to make the player vectors interpretable, and I think they accomplished that.

5.2.2 Necessary Data

The data used is similar to that of the first paper since they both use SPADL. Applying this model may be fractionally easier than the first since only four player action types are considered rather than nineteen. I find both the algorithm and data required to not be overly complicated for a professional data scientist to implement.

5.2.3 Application to Other Sports

As far as the flexibility and ability to apply to other sports, I feel that this could easily be used for hockey. Since passing, dribbling, cross, and shot actions all have a mirror in hockey, it would not be difficult to transfer the model over to hockey. It may be harder to apply to other sports like basketball where actions are different and certain positions are far more likely to dribble or shoot than others. The algorithm would still be interesting in basketball, so perhaps it may be worth the data science team's time to tweak it to their respective sport.

5.2.4 Accuracy

Finally, the biggest problems from this research show in the accuracy of the model. Just like the first model, *Actions Speak Louder than Goals: Valuing Player Actions in Soccer*, the defense is not given correct characterization since most of their actions are taken off the ball. This results in the inability for the model to anonymously select the correct players given a player vector, which was the main test for the model. The accuracy for offensive players is not high, but it is statistically significant. It can choose the correct player in one try based on the player vector 38.2% of the time, then the correct player is in the top ten estimation 64.4% of the time. The model also makes several simplifying assumptions to work, and though I feel that they are reasonable assumptions, they may reduce the accuracy. Without them, the model would not be possible to build, so I understand why the authors had to make them.

5.2.5 Recommendation

I would not recommend this model for the purpose of anonymous selection, but I don't think that is the main application. The descriptions of play from the vectors tended to match the known cultural thought of the players, so these vectors could be used to give basic outlines of player characterization to fans during games. They also could supplement human scouting, but I would not recommend the complete replacement of scouts with this model. The same can be said of player development and replacement. I find this model to be the most unique out of the four, and that the research is more cutting edge than the others. This does not mean that it is better, but I appreciate the authors' attempt to expand the field of soccer analytics.

6 Summary and Conclusions from Research in Soccer Analytics

In this paper, I have summarized the research done in four papers from the soccer analytics field. I provided my personal assessment and recommendation for each paper.

The first paper, *Actions Speak Louder than Goals: Valuing Player Actions in Soccer* introduced some foundational work to the player valuation project. It required extensive input data, but also had a new, simple language for describing player actions. The model was quite accurate for offense, but showed weaknesses in valuing defensive players. It is most applicable to soccer and hockey, and may be difficult to transition to sports that are more different in nature than soccer and hockey. The algorithm behind the model was not overly complex, and I feel that data scientists would not struggle to implement it.

The next paper, *Deep Reinforcement Learning in Ice Hockey for Context- Aware Player Evaluation* had a similar goal to the first, but applied to hockey with a different algorithm. The data needed to make the model is quite detailed, which would present problems to hockey teams who don't have computer generated game data like they do in the NHL. The algorithm in this paper was complex, and requires knowledge of deep learning. Depending on the analysts on the data science team, a hockey program may not have the resources to apply the algorithm. The upside of such a complex algorithm is that it proved to be more accurate than the last. This could also be because hockey has less defined positions, so defenders find themselves all over the ice during the game with opportunities to take shots. The authors tested the algorithm extensively, which I appreciated because that led me to trust it more. The work in this paper, if applied to soccer, would probably have a similarly inaccurate outcome as the first model, since the only work done to address the defensive problem is applying it to a sport where the problem is smaller by nature.

The third paper I discussed was *Measuring Football Players' On-the-Ball Contributions from Passes During Games*. This paper did productive work in soccer player valuation because it addressed the defensive valuing inaccuracy problem. By only valuing passes and not shots, defenders get a more fair shot at earning value. Since the model only considers one action type, the data needed is far less specific and difficult to compile. The algorithm is extremely complex, which again means more time spent learning than applying for a data science team that may apply it. As I mentioned in my original evaluation of the paper, I don't feel that this could easily apply to other sports, since soccer uniquely relies on passing, even more than hockey. Other sports also do not show as much of a problem of overvaluing offense, so they may not even have a need to address the problem that this paper solves. The authors do not do a lot of model confirmation, but it seems to be accurate enough as far as they did test it. Testing for this use case in general is difficult since no ground truth exists for valuing players. This model makes progress towards a better player valuation framework, but I would not say it is far better than other options out there.

Finally, I looked at Decroos' research in *Player vectors: Characterizing Soccer Players' Playing Style from Match Event Streams*. This paper does not have the same goal as the others. It aims to characterize playing style rather than provide a quantitative ranking of professional soccer players. This is still a valid soccer analytics question, and I found it easier to see the benefits of having a model that produces player characterizations rather than overall scores. The algorithm was not overly complex and the data needed is less bulky than for a couple of the papers I have discussed. A data science team would not struggle to implement this model if they wanted to characterize players. Several simplifying assumptions are made to create the model, but I don't feel that this degrades its quality enough to be a significant factor in my analysis of the model. This model could be easily applied to hockey, and again probably not as easily applied to other sports. I find the applications of it (scouting, monitoring player development, and match preparation) to be so compelling that it may be worth making the necessary changes to apply it to other sports. Finally, the accuracy is hard to quantify in this paper. They perform many tests for each different use case, and again identify the lack of accuracy for defensive evaluation. The authors brush over this a bit, but I would like to see this problem addressed further if I were to recommend the model.

If a soccer club were to ask me to make a recommendation of the best model to apply for player evaluation based on my personal research, I would recommend the application of *Player vectors: Characterizing Soccer Players' Playing Style from Match Event Streams.* This model, while imperfect, has direct and obvious application to any soccer teams' goals. The other models would be helpful, but give less detail about the individual player than this one. This model does not provide a strong characterization of defensive players, but I feel that some improvements could be made to address that. The Bransen and Haaren approach of only looking at passes could be applied, which would simplify the input data and allow defenders to be valued more fairly.[3] Characterizing players' passing habits is still helpful to soccer clubs. This recommendation does not account for all research in soccer analytics, but just the four papers summarized in my research. Even given this fact, this paper is helpful, feasible to implement, and has satisfactory accuracy.

If I were tasked to make a new, improved, model for soccer evaluation, I would combine factors from each of the papers I looked at in my research. It would be valuable to a soccer club to have a player vector that indicates threat level for that player at each position on the field. Teams could specify defensive strategy before games based on where opponents' actions have the most value. In my model, I would implement the heatmap to vector approach as in Decroos et. all's Player Vector paper.[5] Instead of producing vectors that simply tell how often actions are taken at each position on the field, I would use elements from the other papers to assign an average valuation per player for places on the heatmap where their actions are taken. I might specifically apply the passing valuation technique from Bransen and Van Haaren since it best confronts the defensive valuation problem. I would use SPADL to format the data because this was a clean and straightforward way to break down complex data. This model would require a lot of data, but I cut down on this a bit by only valuing passes. The loss of information from not considering dribbling and shots is worth the increased defensive analytics accuracy in my opinion. This method would give soccer clubs a "threat index" for any player they want to evaluate before a game. The applications of this research would be mostly to game preparation and scouting.

In the future, I hope to see the defensive issue addressed in a way that does not result in the loss of consideration of certain action types, and for the data to be simplified so that non-professional teams can easily collect their own data and apply machine learning models to improve their teams' chances. I believe that accessibility is important so that sports analytics are less of a feature of solely professional teams. I would find it exciting to see researchers make some models that can be quickly and easily applied so that there is equal opportunity in the world of sports.

References

[1] Bart Aalbers. Expected goals model 2.0. In *SciSports*, 2016.

[2] Daniel Altman. Beyond shots: A new approach to quantifying scoring opportunities.(2015), 2015.

[3] Lotte Bransen and Jan Van Haaren. Measuring football players' on-the-ball contributions from passes during games. In *International Workshop on machine learning and data mining for sports analytics*, pages 3–15. Springer, 2018.

[4] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1851–1861, 2019.

[5] Tom Decroos and Jesse Davis. Player vectors: Characterizing soccer players' playing style from match event streams. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 569–584. Springer, 2019.

[6] Guiliang Liu and Oliver Schulte. Deep reinforcement learning in ice hockey for context-aware player evaluation. *arXiv preprint arXiv:1805.11088*, 2018.

[7] Olav Nørstebø, Vegard Rødseth Bjertnes, and Eirik Vabo. Valuing individual player involvements in norwegian association football. Master's thesis, NTNU, 2016.

[8] Kurt Douglas Routley. *A markov game model for valuing player actions in ice hockey.* PhD thesis, Applied Sciences:, 2015.