UPPSALA
UNIVERSITET

IT 23 051

Degree project 30 credits

June 2023

Potential Pass Networks

Aleksander Wojciech Andrzejewski

Master’s Programme in Data Science

UPPSALA
UNIVERSITET

Potential Pass Networks

Aleksander Wojciech Andrzejewski

Abstract

Passes are the events that happen the most often in football. Therefore, it is important for scouting and situation analysis to recognize the most valuable passes in crucial situations. Aggregating the most optimal passes during the game can provide insights to the coaching staff regarding players' availability to receive dangerous passes, missed dangerous passing opportunities, and pass combinations that should have occurred more or less frequently. In the study, we introduced potential pass networks, which visualize the most valuable passes in the form of previously used pass networks. We implemented these networks with ORTEC event data and TRACAB tracking data using a physics-based approach, incorporating a closest defender model and the concept of error associated with the pass, and convolutional neural networks. Additionally, we compared them between each other. The physics-based approach yielded a more accurate approximation of the most valuable pass in various scenarios, while maintaining the explainability of the outcome. Machine learning method tended to overestimate the value of long passes close to the sideline and failed to provide reasons behind the bias. Overall, using the physics-based approach, we produced a useful tool which can be used in football clubs.

Faculty of Science and Technology

Uppsala University, Uppsala

Supervisor: Mirjam Bruinsma Subject reader: Robin Strand

Examiner: Matteo Magnani

Acknowledgements

First, I would like to thank David J. T. Sumpter for introducing me to the fascinating world of football analytics and providing me with multiple opportunities in the field, including this project. Also, I would like to thank you for constant content support and countless discussions. Your help was invaluable in the process of creating this thesis and my personal development in the area of football data science.

I would also like to thank data scientists at AFC Ajax, Mirjam Bruinsma and Ya'gel Schoonderbeek for providing me with the project idea, data, match videos and synchronization algorithm, which made my work easier. Moreover, I would like to thank you for all the discussions and feedback, which helped me develop the project and find practical use cases of my work. I hope that it can help AFC Ajax achieve their football goals in the future.

Last but not least, I would like to thank my family - my mother, my father, my brother and my grandmother - for their continuous support and trust in me during the process of completing the thesis. This accomplishment would not be possible without you.

Table of Contents

List of Tables

List of Figures

1 Introduction

Data are changing the world of football at an enormous pace. Football clubs from all over the world are using artificial intelligence and data analysis to analyze critical situations from the game, scout players, and prevent injuries. Multiple advanced metrics, such as Expected Goals and Expected Threat, are gaining popularity in the discussions around the games. Moreover, the rise of computer vision algorithms allowed clubs to gain greater access to information concerning player positioning during games.

Passes are the events that happen the most often during a football game. On average, there are approximately 900 passes in a match. The quality of passes significantly influences the creation of dangerous scoring opportunities. Notably, creative midfielders, such as Pedri or Kevin de Bruyne, are often key players of their teams. Therefore, there is a lot of discussion among football fans, pundits, coaches, and players concerning the correctness of passing decisions in specific situations. To help the debate, scientists introduced multiple approaches to the approximation of pass success probability and pass value, using both physics-based approach and machine learning methods, to provide an unbiased answer to the question of the most valuable pass in a specific situation.

Pass networks serve as valuable tools to analyze team shape and passing patterns during the game. They provide information about players who passed between each other most frequently and the approximation of the attacking formation during the game. However, these networks do not provide information about missed opportunities, unexplored passing options, and the potential value that could have been created by different passes and player combinations. In this work, we introduce the potential pass networks. By aggregating all the most valuable passes, this tool allows us to glean information about the availability of players to receive the pass, identify which players should have passed between each other more or less frequently, and assess the potential danger they could have created.

Moreover, we introduce the concepts of the defender’s influence on the pass success probability and the error associated with the pass. The first model aims to approximate the pass success probability to the zones that could have led to possible block or interception by the pressuring defender. Moreover, this approach can help us identify which components influence pass and pressure success. The concept of error associated with the pass addresses the risk with passes sent to the areas close to the sidelines. Physics-based approaches encountered difficulties in these two areas. Therefore, we claim that by introducing these two models, we can achieve a more precise estimation of pass success probability and the most valuable areas on the pitch in a specific situation. Furthermore, since previous state-of-the-art physics-based approaches only investigated

the reward from making a successful pass, we decided to calculate the value of the pass by including both the reward of making a successful pass and the risk of the ball being regained by the defending team.
Since the pass success probability and pass value concepts were investigated from physics-based and machine learning perspectives, we create and compare the potential pass networks using our physics-based approach with state-of-the-art machine learning. Using the potential pass network, the comprehensive visualization of all potential passes during a football game, we seek to answer the question concerning differences in these two approaches, their advantages and disadvantages.
Therefore, the following research questions are going to be investigated in this work

1. Can introduced potential pass networks provide valuable insights about a game?
2. How can we overcome the problem of overvaluing long passes close to the sideline in a state-of-the-art physics-based approach?
3. Can the inclusion of risk and reward influence the evaluation of the most valuable pass?
4. What are the differences in the outcome between physics-based and machine learning approaches to pass success probability and pass value models?

This paper is structured as follows. Section 2 provides an in-depth literature review of previous research in the areas of pass networks, pass success probability, pass value models, and their combinations. Section 3 presents the necessary mathematical theory from the models used in this thesis. In Section 4, one can find the description of the detailed methodology used in the project. Section 5 provides the results of our work. Section 6 discusses the results, which lead to conclusions presented in Section 7.

2 Literature Review

2.1 Pass networks

The concept of using network science to analyse football was introduced by P. Gould and A. Gatrell [1]. The authors analysed team structures based on a graph created with passers and receivers of the passes. However, pass networks did not become relevant in football until the beginning of the 21st century when Duch et al. presented pass networks as a novel way to visualise players' performance and team structure [2]. The nodes' location accounted for the players' average position on the field, with their size indicating the metric of interest. The presence of edges signified that a pass was made from one player to another, with wider edges showing higher values of the metric of interest. Various football entities adjusted this technique to visualise their specific topics of interest. Since then, multiple concepts from network science have been investigated to explain the success or failure of a team. J. M. Buldu et al. used pass networks to show that the legendary FC Barcelona team led by Pep Guardiola, which won 6 trophies in a year, had the largest value of adjacency matrix, highest centrality acquired by a single player and the lowest shortest-path among La Liga teams in the 2009/2010 La Liga season [3].

2.2 Pass success probability models

Although important in football, the concept of modelling pass success probability was not studied in-depth until a few years ago due to tracking data availability and required computational power. P. Power et al. conducted a breakthrough research in which they modelled the difficulty of a pass using logistic regression for interpretability reasons [4]. Moreover, they introduced the concept of a potential receiver to avoid survivorship bias. They estimated the expected receiver of a pass if it was blocked. Therefore, multiple features, such as the length of a pass, differed from the observed outcomes.

Later, multiple scientists approached the problem of identifying the probability of a successful pass in a physics-based manner. First, W. Spearman et al. introduced the concept of pass success probability based on the self-propelled particles [5]. Given the situation on the pitch, the authors simulated ball trajectory and player movement. Then, they estimated the probability of intercepting and controlling the ball by each player on the field. A year later, W. Spearman introduced the concept of the Potential Pitch Control Field (PPCF) [6]. The author calculated the probability of controlling the ball in each zone by each team, given the starting position of a pass. They accomplished this by simulating players' movement and ball trajectory. However, this approach, while computationally less extensive, did not account for possible pass interceptions. F. Peralta Aguacil et al. managed to overcome this problem by first calculating the equation of motion by each player, then simulating a pass and calculating interception probabilities [7].

Also, multiple machine learning models were introduced to estimate the probability of a successful pass. J. Fernandez and L. Bornn created a neural network architecture to estimate pass success probability [8]. The model aims to produce a 104 x 68 layer with pass success probability, with each pixel corresponding to an area sized 1m x 1m on a football field. The authors accomplished it using convolutional neural networks and the computation of loss only at the area of the pass end location. G. Anzer and P. Bauer presented a feature-based method of estimating pass success probability with hand-crafted features [9]. They used a gradient-boosted algorithm to estimate pass success probability. Moreover, they estimated the receiver of blocked passes by simulating ball trajectory and player movement. P. Robberechts et al., in their research around player creativity, introduced a different gradient-boosted algorithm, which achieved better performance scores than SoccerMap on Statsbomb 360 data [10]. However, due to data limitations, the architecture used by the authors did not contain layers with players' velocities. U. Dick et al. introduced a Gaussian framework to estimate the probability of a player to receive the pass [11]. They created a graph recurrent neural network to estimate the parameters of Gaussian distributions, which are then used to calculate the player's availability to receive a pass.

2.3 Pass value models

With the rise of data availability, scientists working in football clubs started to analyse events on the football pitch that contributed to scoring a goal. S. Rudd created the first model to evaluate the offensive contributions of players [12]. The model is based on the Markov chain and the concept of ball transition around the football pitch. K. Singh developed this approach and, using a similar framework, estimated the value of each of 192 equally sized zones on the pitch [13]. He named the created framework Expected Threat (xT).

Data availability and computational power development allowed scientists to use machine learning techniques. T. Decroos et al. introduced the Valuing Actions by Estimating Probabilities (VAEP) framework. Using a gradient-boosting algorithm, the authors estimate the probability of scoring and conceding a goal. Their difference was the value of an action. J. Fernandez et al. introduced the possibility of using SoccerMap architecture to evaluate actions [14]. In comparison to the SoccerMap architecture for pass success probability, the authors used different layers in the input tensor, changed the prediction surface and used a different cost function. Moreover, they trained two different models for successful and unsuccessful passes. This approach does not assume that missed passes produce negative values or that successful passes must provide positive value. Additionally, multiple data providers developed their models for evaluating actions, such as Statbomb's On-Ball Value (OBV) model and Opta's Expected Threat (xT). However, the details of these frameworks are not publicly available.

2.4 Combining pass success probability and pass value models

Knowing both pass probability and pass value, one can easily calculate the expected value of a pass. W. Spearman created the On-Ball Scoring Opportunity framework based on the multiplication of Potential Pitch Control Field, Expected Goals and ball transition probability, which serves as a penalty for playing long passes to opposition [6]. F. Peralta Aguacil et al., similarly to W. Spearman, multiplied different value and probability layers. In this case, the authors multiplied their pass success probability layer with the possession team's space control and pass impact, achieving the expected value of a pass [7]. J. Fernandez et al. multiplied the pass success probability achieved from SoccerMap architecture with pass value and pass selection probability layers, calculated with the architecture mentioned above [14]. By multiplying them, the authors achieved the expected value of a pass. Another approach was suggested by L. Shaw in the Friends of Tracking YouTube video series [15]. They estimated the expected value of a pass by multiplying the Expected Threat with the Potential Pitch Control Field, which allowed for an approximation of the value for each zone. Lastly, P. Robberechts et al. created a gradient-boosted model to calculate pass success and pass selection probabilities together with pass value [10]. They calculated the values for each of 442 (26 x 17) equally sized cells on a football pitch and multiplied them with each other to estimate the expected value of a pass.

3 Theory

3.1 Data

3.1.1 Event data

Football event data represent events, such as passes, shots, carries, and tackles, that happened on the football pitch. The data are collected manually by an unbiased collector and labelled. Since physical actors collect the observations, there is an element of human error associated with the precision of captioned observation.

3.1.2 Tracking data

Tracking data represent the position of players and the ball on the football pitch. It is collected from the cameras located on the field. The players are identified by their shirt numbers, and the dataset consists of their x and y coordinates. Tracking data also provides x, y and z coordinates of the ball. While human error in the data collection is neglected, errors may exist from the imperfection of the data collection algorithms.

3.2 Physics-based approach

3.2.1 Expected Threat

Expected Threat model is a position-based value model introduced by K. Singh [13]. The model is based on the concept of ball transition around the pitch. The author suggested an iterative solution to estimating the expected value of a zone. They initialize the values to zero and calculate them until the convergence of the Equation 3.1.

$$xT_{x,y} = s_{x,y} * g_{x,y} + (m_{x,y} * \sum_{z=1}^{16} \sum_{w=1}^{12} T_{(x,y)\rightarrow(z,w)} * xT_{z,w}) \tag{3.1}$$

In the formula, $xT_{x,y}$ is the value of the zone (x,y), $s_{x,y}$ is the empirical shot probability from this zone, $g_{x,y}$ is the goal probability, $m_{x,y}$ is the empirical probability of moving the ball from this zone and $T_{(x,y)\rightarrow(z,w)}$ yields the transition probability from zone (x,y) to zone (z,w). Notably, one can interpret the value of a zone as the probability of scoring the goal within the following five events if the ball is currently in this area.

3.2.2 On-Ball Scoring Opportunity

On-Ball Scoring Opportunity is a framework developed by W. Spearman to evaluate the value of the current game-state for the attacking team. The author defines the value as the probability of scoring a goal given the current situation on the pitch $P(G|D)$. This value is first computed for each zone $r \in R$ using the probability that the ball will be moved to a specific zone - T_r, controlled there by the attacking team - C_r, and the goal will be scored from there - G_r. As the next step, W. Spearman summed together the estimates for each area to evaluate the value of the game-state. This process can be described with Equation 3.2.

$$P(G|D) = \sum_{r \in R} P(G_r \cap C_r \cap T_r|D) \tag{3.2}$$

The author decomposed this sum using the law of probabilities as illustrated by the Equation 3.3.

$$P(G|D) = \sum_{r \in R} P(G_r|C_r, T_r, D) * P(C_r|T_r, D) * P(T_r|D) \tag{3.3}$$

The first factor can be interpreted as the probability of scoring a goal from zone r given that the attacking team manages to move and control the ball in this zone. The second one signifies the probability of controlling the ball in zone r. The latter factor denotes the probability of moving the ball to zone r. D indicates the instantaneous state of the game, i.e., the current situation and position of the players on the pitch.
As the probability of scoring a goal from zone r, W. Spearman suggested his own Expected Goals model. The author used the Potential Pitch Control Field (PPCF), described in Section 3.2.3, as the probability of controlling the ball in zone r given that the ball arrives there. The transition model was created as the combination of the Normal distribution, with the mean being the average length of a pass and the variance - the variance of their length, and the previously created PPCF model. The output is then normalized so that the sum of the layer equals 1.

3.2.3 Potential Pitch Control Field

Potential Pitch Control Field is a model created by W. Spearman that quantifies the probability of controlling the ball by each team given the current location of the ball and the players [6]. This value is estimated by solving the Equation 3.4.

$$\frac{dPPCF_j}{dT}(t, \vec{r}, T|s, \lambda_j) = (1 - \sum_k PPCF_k(t, \vec{r}, T|s, \lambda_j)) * f_j(t, \vec{r}, T|s)\lambda_j \tag{3.4}$$

where $f_j(t, \vec{r}, T|s)$ is the probability that a player j can reach area r within time $t \leq T$ with s being standard deviation of player's arrival time. λ_j is the inverse of the time required to control the ball by a player j, $PPCF_k$ denotes the individual potential pitch control field for each player k. After this computation, the equation is integrated from 0 to T to get each player's individual pitch control values. As a result of this computation, the output consists of three layers with potential pitch control field: one for the attacking team, one for the defending one and individual pitch control values for attacking players.

3.2.4 Off-ball Expected Threat

The concept of Off-ball Expected Threat was developed by L. Shaw in their series of YouTube videos for Friends of Trackings [15]. The author did not give a name to this approach. However, D. Sumpter, co-creator of the channel, named it *Off-ball Expected Threat* on the course webpage for Mathematical Modelling in Football offered at Uppsala University [16].

The approach combines the positional Expected Threat model, discussed in Section 3.2.1, and the Potential Pass Control Field, presented in Section 3.2.3. The value is calculated using the following Equation 3.5.

$$oxT = PPCF * xT \tag{3.5}$$

By multiplying the positional Expected Threat with the Potential Pitch Control Field for the attacking team, the author achieved the Off-ball Expected Threat value. The author then proposed that the pass with the highest metric value would add the most value to the possession.

3.3 Machine Learning approach

3.3.1 SoccerMap

SoccerMap is a deep neural network architecture created by J. Fernandez and L. Bornn [8]. It is a convolutional neural network with input consisting of $l\ x\ h\ x\ c$ matrix, representing a snapshot of a game, which is mapped to $l\ x\ h$ representation of the field for a specific value. The architecture is based on concepts of convolutional layers. Nearby pixels should share similar values due to parameter sharing. Additionally, the network learns from single-location labels, which makes the training of the model a weakly-supervised learning task. The loss is learned only from the pixel with the end location of an event, which allows to shrink the output to a single prediction value. Moreover, the authors suggested the architecture which makes predictions at each of $1x, 1/2x$ and $1/4x$ representations of the field, first downsampling the output, then upsampling predictions to fuse them with higher representations. The architecture is presented in Appendix A.

3.3.2 Framework for evaluating passes

Originally, the authors presented the architecture to predict pass success probability in football. They optimized the negative log loss, presented in Equation 3.6

$$\mathcal{L}(\hat{y},y) = -\frac{1}{N}\sum_{i} y_i * log(\hat{y}_i) + (1 - y_i) * log(1 - \hat{y}_i) \tag{3.6}$$

and achieved a surface l x h with the metric.

J. Fernandez et al. [14] suggested a framework to evaluate a possession based on this deep learning model. The authors achieved the pass selection probability by changing the final activation function from sigmoid to softmax. Then, by adding layers with contextual features - the number of attacking and defending players behind the ball, the number of dynamic lines a pass would break and the number of players between the goal and location - together with pass success probability, the authors approximated the value of a pass. To do so, they suggested training separate models for successful and unsuccessful passes by minimizing the mean squared error, defined in Equation 3.7. As the target variable, the authors suggested 1 if a goal was scored within 15 seconds from the pass, -1 if the goal was conceded, 0 otherwise.

$$\mathcal{L}(\hat{y},y) = \frac{1}{N}\sum_{i}(y_i - \hat{y}_i)^2 \tag{3.7}$$

After creating the probability surface, the authors suggested the following way to estimate the value of making a pass. First, they multiplied the pass success probability with the value model for successful passes and the pass failure probability with the value model for unsuccessful ones, as presented in Equation 3.8.

$$\begin{aligned}\mathbb{E}[G|A=\rho,D_t,T_t] = \mathbb{E}[G|A=\rho,O_\rho=1,D_t,T_t]\mathbb{P}(O_\rho=1|A=\rho,D_t,T_t)\\ +\mathbb{E}[G|A=\rho,O_\rho=0,D_t,T_t]\mathbb{P}(O_\rho=0|A=\rho,D_t,T_t)\end{aligned} \tag{3.8}$$

In this equation, A signifies the action, ρ is the notation for a pass, D stands for the location in the field, T_t represents the spatiotemporal information at time t, and O corresponds to the outcome of the pass. G is the outcome of possession.

Then, they multiplied element-wise the outcome of this operation together with the pass selection probability and summed the outcome for each zone l in all possible locations L, achieving the expected possession value given a pass is made, as described in Equation 3.9.

$$\mathbb{E}[G|A=\rho,T_t] = \sum_{l\in L}\mathbb{E}[G|A=\rho,D_t=l,T_t]\mathbb{P}(D_t=l|A=\rho,D_t,T_t) \tag{3.9}$$

4 Implementation

4.1 Data

4.1.1 Event data

In this project, we used ORTEC data provided by AFC Ajax. The data were provided in the *JSON* format. Data provided information about events that happened on the pitch, their timestamps and the players who performed them. Moreover, information retrieved from the data contained the x and y coordinate of an event, collected on 100x100 space, where the top left corner of the field is the point $(0,0)$. The coordinates were stored so that the team on the ball constantly attacked from left to right.

4.1.2 Tracking data

For this work, we used TRACAB data. TRACAB collects the data with the frequency 25 *Hz*, corresponding to a collection every 40 milliseconds. The provided data were saved in the *xml* format. The coordinates were stored on an interval $[-1,1]$ with the bottom left corner representing the point $(-1,-1)$.

4.1.3 Data processing

Since the event data was provided in the *JSON* format, using *pandas* package in Python, we extracted the information stored in a nested structure and stored it in a new *JSON* file. For the tracking data, we first extracted the data from the *xml* file using the Python *xml* library. Then, we created a *pandas* dataframe. Each row corresponded to a different frame, and each column corresponded to a different player coordinate. For each player, we created two columns, one with their x and one with their y coordinates, normalized them to the [-52.5, 52.5] interval for x and [-34, 34] interval for the y coordinate. Therefore, the point $(0,0)$ corresponded to the middle of the pitch. We also kept the x,y and z coordinates of the ball in separate columns. As the next step, we calculated each player's velocities and smoothed them with the Savitzky-Golay filter. For each player, we stored their velocities in x and y directions, along with their speed (i. e. total displacement over frame length) as new columns. Moreover, we changed the velocities exceeding 12 m/s to *NaN* since they are impossible to achieve by humans and most likely occurred due to possible data collection errors. In the end, we saved the dataframe in a *JSON* file.

After prepossessing these data, we extracted the information about the attacking direction from the tracking data, i.e. we checked the coordinate of the home team goalkeeper in the first frame. If it was smaller than zero, the home team attacked from left to right. Then, according to this information, we adequately flipped the event data coordinates so that they represented the actual location of the pass.

4.1.4 Data synchronization

Since different data providers collected the data, the timestamps of events did not correspond perfectly to the situation on the pitch in the tracking data. Therefore, we used a synchronization algorithm provided by AFC Ajax, which was based on investigating frames when the ball was close to the passer and receiver. Information about them was available in the event data. First, we created a dictionary with player keys and their team keys from the event data together with their shirt numbers, which allowed us to map their ORTEC identifiers to the tracking data ones. Then, using the dictionary, we found the positions of these players. As the next step, we applied the provided synchronizer. Using this algorithm, we managed to synchronize (i.e. estimate the frame with the start of the contact with the ball of the passer, the frame when the pass occurred and the frame when the receiver controlled the ball) most of the passes. After the synchronization, we stored the dictionary as a *pickle* file and the event dataframe, with the frame identifiers corresponding to frames in the tracking data when the event happened, in a *csv* file.

4.2 Physics-based approach

Due to the computational efficiency and interpretability of the method, for the physics-based approach, we decided to use the framework by L. Shaw [15] described in Section 3.2.4, which uses the concepts from W. Spearman's *On-Ball Scoring Opprotunity* framework presented in Section 3.2.2. Similarly to L. Shaw, we decided to omit the transition probability model since we claim it serves as a punishment for long passes. It would be useful to evaluate the possession rather than to find the pass that would have provided the highest value. To investigate the most optimal passes, we decided that we want the pass with the highest expected value to be chosen with probability 1 rather than with its transition probability. Also, we tried to overcome the potential issue of overvalued long passes close to the sideline. Importantly, in this approach, we prioritize the value of explainability while acknowledging the significance of mathematical correctness.

4.2.1 Pass success probability

Potential Pitch Control Field

For the Potential Pitch Control implementation, we based on the *Python* code provided by L. Shaw in their series of YouTube tutorials on the Friends of Tracking channel [15]. We also used suggested modifications by J. Michalak, who proposed using *Numba* to improve the performance of the code [17]. However, in contrast to both authors, we decided to adhere to the parameters from the original paper by W. Spearman [6].

Moreover, contrary to all the authors, we decided to set the player's maximum speed to 7.8 m/s and their acceleration to 10.24 m/s^2. Initially, the values were set to appropriately 5 m/s and 7 m/s^2, which, we claim, are too low for a professional athlete. Therefore, we used values suggested by A. Fujimura and K. Sugihara [18] as first suggested by G. Rolland [19]. We expected that this change would overcome the problem of too high Potential Pitch Control Field values for the attacking team on the wings, which would lead to overvaluing long passes close to the sideline.

Closest defender influence

Potential Pitch Control Field does not account for a possible interception or block of a pass. However, from the empirical data, we found that the distance between the passer and the closest defender is essential to pass success probability. Therefore, we decided to build a probabilistic model of a pass being intercepted by the closest defender. For the reason of interpretability, we utilized a logistic regression model to predict the pass success probability. Moreover, other than distance, we decided to use features related to the relative position of both players, their velocities and the nonlinear transformations of these features. The features are presented in Table 4.1. The angles used as features are visualised in Figure 4.1.

Feature	**Value**	**Description**
distance	float	Distance of the passer and the closest defender (the closest opponent) in meters
defender_speed	float	Moving speed of the closest defender (m/s)
ball_speed	float	Moving speed of the ball (m/s)
passer_speed	float	Moving speed of the passer (m/s)
time_to_pass	float	The ball controlling the time of the passer in seconds
distance_to_x	float	Relative distance of the passer to the top and right size of the football pitch
distance_to_y	float	
defender_angle	float, $[0, \pi]$	angle θ_1(*)
passer_angle	float, $[0, \pi]$	angle θ_2(*)
ball_distance_angle	float, $[0, \pi]$	angle θ_3(*)
ball_passer_angle	float, $[0, \pi]$	angle θ_4(*)

Table 4.1: Features used for the closest defender influence model

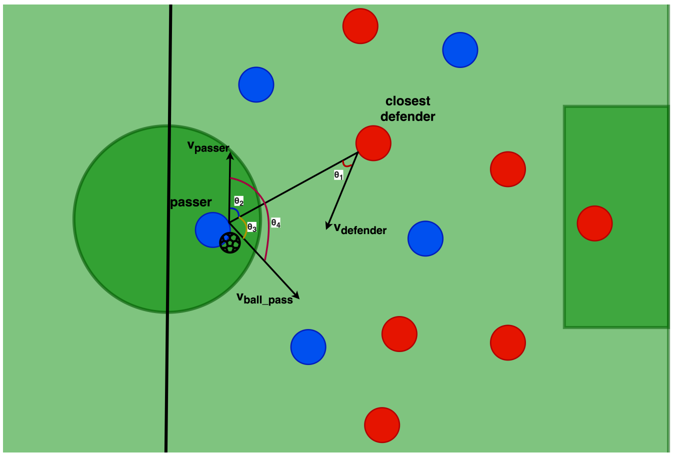


Figure 4.1: Visualised angle features

We implemented the model in *Python* using the *scikit-learn* package using 90% of correctly synchronized passes and evaluated it on 10% of data unseen in the training process. As the next step, we compared the model with the baseline - average pass probability assigned to every pass in the test dataset.

Then, we approximated the pass success probability to each zone, given the closest defender influence for each of 1600 equally (52 x 30) sized zones on the football pitch, creating a layer for a specific situation on the pitch.

However, in various situations, multiple defenders were close enough to the passer to influence pass success. We claimed that the pass success could be affected by the defenders no more than 7 meters away from the passer, as showed by M. Bruinsma [20]. We decided to treat each defender in this area independently and created a layer for each. Then, we multiplied the layers element-wise. If one or no defenders were in the 7-meter radius, the layer was computed only for the closest defender. While not perfectly correct mathematically, we claim this approach provided a better approximation of probability from a football perspective.

Probability of ball staying in the field

There is always some error associated with a pass. Players do not always manage to play the ball perfectly to the zone they aim for. Therefore, the ball sometimes goes out of the play. We decided to quantify the probability of the ball staying on the pitch. However, no data is available on where precisely the player aimed the ball. Therefore, attempting to quantify this probability from the historical data was impossible. After consultation with football specialists working at AFC Ajax, we decided to assume that if a player aims the pass to a specific location (x,y), the pass end location is normally distributed, with the mean in the end location and standard deviation corresponding to $10 * \frac{length}{120}$. We decided to use 120 since it is (in meters) the pass length that was not exceeded in the 2021/2022 Eredivisie season.

This approach allowed us to create a layer for the probability of the ball staying in the

field. First, given the starting position, we calculated the length of a potential pass to each of 1600 (52x30) equally sized zones. Then, for each zone, we sampled 1000 passes from the two-variate normal distribution, with the mean as mentioned earlier and the covariance matrix with the previously described standard deviations squared on the diagonal. The proportion of passes which stayed inside $[0,105]$ interval for the x-coordinate and $[0,68]$ interval for the y-coordinate is the estimated probability.
We implemented this approach in *Python* using *scipy* package.

Pass error and Potential Pitch Control Field

In the previous section, we introduced the concept of an error associated with each pass. Even if the ball stayed in the field, the pass may have ended in a different zone. Therefore, we simulated the pass to a specific location. Then, we calculated the probability of a ball ending in each of 1600 (52x30) equally sized zones. As the next step, for each zone, we multiplied the probability of the ball ending there P_k with the previously calculated potential pitch control value of the attacking team $PPCFa_k$, resulting in $sPPCFa_k$, as presented in Equation 4.1.

$$sPPCFa_k = \sum_k P_k * PPCFa_k \tag{4.1}$$

Subsequently, we summed up all the products. By calculating the value for each zone, we achieved the Potential Pitch Control Field with possible error incorporation, which will be later called *smoothed Potential Pitch Control Field* in this paper.
We applied the procedure in *Python* using *scipy* package.

Combining layers into pass probability

After we calculated the layers with the probability of pass success given the closest defender influence and the probability of the ball staying in the field given pass start and smooth Potential Pitch Control Field, we combined them to achieve the pass success probability. We decided to treat these layers as independent and multiply them with each other element-wise. The approach of assuming independence and multiplying various layers was used before by W. Spearman [6] and F. Peralta Aguacil [7], among others.

4.2.2 Pass value

As a pass value function, we decided to use K. Singh's Expected Threat model presented in Section 3.2.1. We used a value matrix provided by L. Shaw in their series of YouTube tutorials on the Friends of Tracking channel [15], in which they calculated the values using event data from multiple leagues and seasons. The matrix provides the value of 1600 (52x30) equally sized zones on the football pitch. While some of the previous authors [6], [7], [15] only considered reward when estimating the value of a pass, we chose to approach the expected value of a pass using both reward and risk. First, we computed the reward in the following manner. Let

- $sPPCFa$ be the smoothed Potential Pitch Control Field for the attacking team
- CD be the probability of a successful pass given the closest defender influence
- SD be the probability of the ball staying in the field
- xT be the Expected Threat matrix

then, we can calculate the reward using Equation 4.2.

$$Reward = sPPCFa * CD * SD * xT \tag{4.2}$$

The first three factors represent the probability of a successful pass to each zone, and the last one is its value without spatio-temporal context.

Moreover, we calculated the risk of the pass using the following approach. We considered the opponent's Expected Threat matrix as a flipped Expected Threat matrix along both x and y axis for the attacking team. Let

- $sPPCFa$ be the smoothed Potential Pitch Control Field for the attacking team
- CD_k be the probability of a successful pass given the kth defender, whom we previously considered in the closest defender influence calculations
- SD be the probability of the ball staying in the field
- xT_d be the Expected Threat matrix for the opposing team
- xT_k be the Expected Threat matrix for the opposing team in the zone where kth defender is located
- $sPPCF_k$ be the smoothed Potential Pitch Control Field for the defensive team kth defender in the zone they are located
- xT_s be the Expected Threat of the closest zone to where the ball possibly went out

then, we can calculate the reward using Equation 4.3.

$$Risk = (1 - sPPCFa) * xT_d + (1 - SD) * xT_s + \sum_{k=1}^{n} (1 - CD_k) * xT_k * sPPCF_k \tag{4.3}$$

The first element represents the value the opposing team gains when controlling the ball in the zone it was aimed at. The second component denotes the value the opposing team gains when the ball gets kicked out of the pitch. The closing element signifies the value the opposite team gains if any of k defenders intercept a pass we created the closest

defender influence layer for. To account for the possibility that the defenders may not manage to control the ball, we used the defender's individual potential pitch control in this calculation.
After calculating the risk and reward components described above, we can subtract the risk from the reward to obtain the expected value of a pass, as presented in Equation 4.4.

$$EV = Reward - Risk \tag{4.4}$$

4.2.3 Finding potential passes

Once we computed the value of a pass, we can find the potential pass. Due to computational limitations, we computed the value for every fifth frame from the start of the contact to the moment of pass. We claim that the situation on the pitch is not dynamic enough to encounter significant changes if the values were computed for every 0.2s instead of every 0.04s. The potential pass is selected as the one which generated the highest value across all computed frames. The start location of a potential pass is assigned as the ball location in the frame when the value is the highest, while the end location is determined as the coordinates of the zone with the highest pass value. To estimate the potential receiver, we chose the player whose individual pitch control in the potential pass end location was the highest. We set a threshold of 0.1 for the Potential Pitch Control Field value for the attacking team in the end area of a potential pass. We claimed we could not determine the potential pass if the $PPCF$ value was lower than the threshold. However, such a situation did not happen more than four times in any of the analyzed games.

4.2.4 Creating potential pass network

To compute the pass network, we found the potential pass for every situation when players made a pass. Then, we made a network. We calculated the location of nodes as the average location of potential passes and their receptions for each player. We assigned the node size as the number of potential passes normalized per 90 minutes played. Then, we estimated the node's colour as the average value of potential passes. The vertices between two nodes represented the presence of more than three passes from one player to another. We calculated their width as an average number of passes from one player to another and assigned their colour as the average value added by passes between 2 players.

4.3 Machine learning approach

We decided to follow the framework introduced by J. Fernandez et al. as described in Section 3.3.2 for the machine learning approach. However, during this procedure, we decided not to use pass selection probability, which we claim served a similar purpose as

W. Spearman's transition probability [6]. While estimating the expected value of a pass, we wanted to consider only the probability of completing it together with its value. The probability of pass being selected can be used for determining the value of possession given its current state, as presented by authors [14].

4.3.1 Pass success probability

We implemented the SoccerMap architecture to estimate the pass success probability as described in Section 3.3.1. However, since the loss is computed from the single pixel representing the end location of the pass, it was required that we decide if we use the not perfectly synchronized end of the pass from the event data or the start of the tracking data coordinates of the beginning of next contact achieved from the synchronization algorithm. We decided to use the latter since most passes were intercepted or passed successfully. However, the problem of incorrect representation of the end pixel occurred for blocked passes. ORTEC did not recognize the player who blocked a pass as its receiver. Therefore, using the synchronization algorithm, we got the location when the next player controlled the pass after the block. On the other hand, this approach allowed us to consider the players' location thanks to the correct synchronization of the end location of the pass.

We replicated the design choices used by J. Fernandez et al. as described in Section 3.3.2. We used similar input tensors sized 104 x 68 x 13, presented in Appendix B. While not stated explicitly by the authors, we decided to standardize the layers with the distance to the goal and the distance to the ball, accordingly, the 7th and the 9th layer. Moreover, we optimized the same loss function - the negative log-likelihood, used the same batch size - 32, utilized the Adam optimization algorithm with the learning rate $1e-4$ and the default setting of betas, made use of the early stopping technique with patience set to $1e-3$.

We split the data in the following way: passes from 80% of games were used for training, 10% of games were used for validation and 10% of games were used for testing. Once we trained the model, similarly to the authors, we applied the temperature scaling procedure with a value of 0.5 [21]. Moreover, after the training, we evaluated the model on unseen data using the loss and expected calibration error with quantile binning strategy and 10 bins, presented in Equation 4.5.

$$ECE = \sum_{k=1}^{10} \frac{|B_k|}{N} |(\frac{1}{|B_k|} \sum_{i \in B_k} y_i) - (\frac{1}{|B_k|} \sum_{i \in B_k} \hat{y}_i) \tag{4.5}$$

B_k represents the set of examples in the kth bin, N is the number of observations, y_i stands for true labels and $\hat{y}_i$ for predicted labels.

We trained the model in the *PyTorch v1.13* environment using *NVIDIA TITAN Xp* GPU.

4.3.2 Pass value

Also, for the pass value models, we decided to follow J. Fernandez et al. approach and implement the SoccerMap architecture as described in Section 3.3.2. Therefore, we compiled two different models, one on successful and one on unsuccessful passes. Furthermore, we used the tensors consisting of the same layers as the authors. Their structure is presented in Appendix B. To estimate the team lines, we used Jenks Natural Breaks algorithm [22]. Moreover, we transformed the output of the neural network from [0, 1] interval to [-1, 1], we optimized the same loss function - the mean squared error, used the same batch size - 16, utilized the Adam optimization algorithm with the learning rate $1e-5$ and the default setting of betas and made use of the early stopping technique with patience set to $1e-6$.

We split the data in the following way: passes from 244 games were used for training, passes from 30 games were used for validation and passes from 30 games were used for training. Similarly to the pass probability model, we evaluated the models using the loss and expected calibration error with quantile binning strategy, represented in Equation 4.5.

We trained the model in the *PyTorch v1.13* environment using *NVIDIA TITAN Xp* GPU.

4.3.3 Finding potential passes

Once we computed the value of the pass, we calculated its expected value. We approximated it using the expected value of a pass to a specific zone, as described in Section 3.3.2 and Equation 3.8

Contrary to the physics-based approach, we computed the value for every frame when the player who made the pass controlled the ball. We selected the start of the potential pass as the ball location in the frame when the value was the highest and the end of the pass as the pixel with the maximal value. As the receiver of the potential pass, we estimated the player from the same team, who was the closest to the end location of the pass at the moment it should have been made.

4.3.4 Creating potential pass network

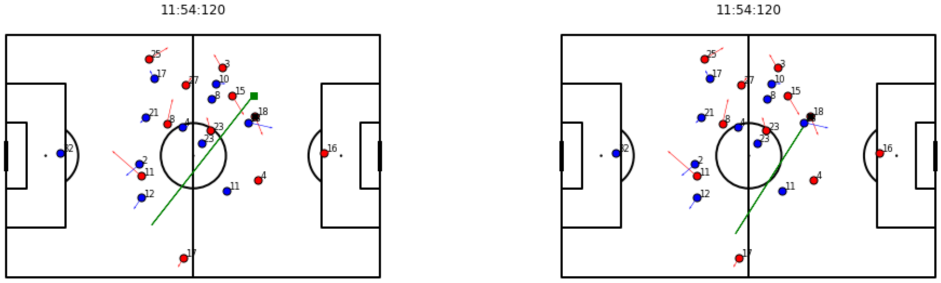
As for the physics-based approach, we computed the potential pass network similarly. Namely, we found the potential pass for every situation when players made a pass. Then, we made a network. We calculated the location of nodes as the average location of potential passes and their receptions. We estimated the size of the node as the number of passes normalized per 90 minutes played. Then, we assigned the node's colour as the average value of passes. The vertices between two nodes represented the presence of more than three passes from one player to another. We calculated their width as an average number of passes from one player to another and assigned their colour as the average value added by passes between 2 players.

5 Results

5.1 Data

In the project, we decided to use the data from the 2021/22 Eredivisie season, the most recent fully completed season of the highest division in Netherlands. We preprocessed and synchronized the data as described in Section 4.1. Using the synchronization algorithm provided by AFC Ajax resulted in 90% successful data synchronization, i.e. correct estimation of the frame when a player started the contact, made a pass, and the next player started their contact.

Moreover, this approach allowed us to unify the process and use only tracking data coordinates in this project. The difference between the event data coordinates of the pass *(a)* and tracking data coordinates of a pass *(b)*, calculated as ball location in the frame when the pass was made and when the next player received the pass, is presented in Figure 5.1 for the situation from the 11*th* minute and 52*nd* second of PSV - AFC Ajax game, which happened on 21.01.2022. The situation is available to investigate under the link [Video link].



(a) Event data coordinates of a pass

(b) Tracking data coordinates of a pass

Figure 5.1: The difference between event data and tracking data coordinates of a pass

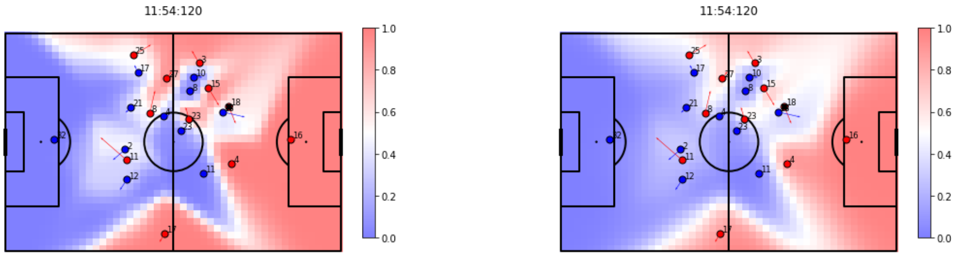
In the figure, blue dots represent the position of Ajax players, and red dots represent the location of PSV players. Arrows represent their velocities, and green line represents the pass.

5.2 Physics-based approach

5.2.1 Pass success probability

Potential Pitch Control Field

We implemented the Potential Pitch Control Field as described in Section 4.2.1. The difference between the parameters which L. Shaw decided to use *(a)* and the parameters that we utilized *(b)* is presented in Figure 5.2 for the same situation as in the previous section.



(a) Potential Pitch Control Field with L. Shaw parameters

(b) Potential Pitch Control Field used in the project

Figure 5.2: The difference between Potential Pitch Control Field with L. Shaw parameters and parameters used in this project

In the figure, red zones represent zones controlled by PSV, and blue ones represent zones controlled by Ajax. White areas stand for zones with equal control by each team.

Closest defender influence

We implemented the closest defender influence on the pass success probability model (for more details, see Section 4.2.1). The evaluation of the model is presented in Table B.1.

	Brier score	**AUC**	**Precision**	**Recall**
Model	0.135	0.774	0.818	0.969
Average pass completion	0.170	0.5	0.799	1

Table 5.1: The evaluation of the closest defender model

Then, we created the layer with the pass success probability for each of the 1600 (52 x 30) zones on the pitch. Figure 5.3 presents the layer for the previously mentioned situation. In the figure, the darker the zone, the higher the probability of making a successful pass to this area.

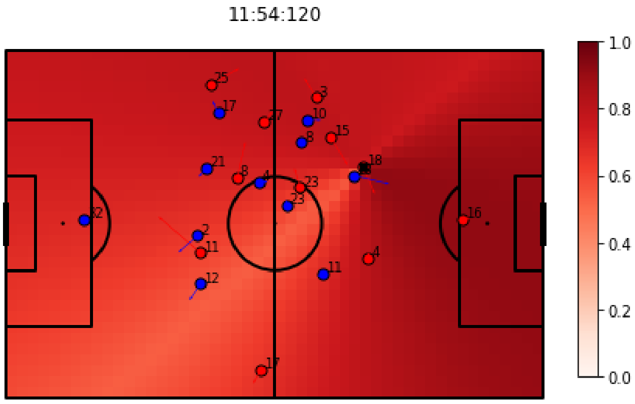


Figure 5.3: Layer with the pass success probability given closest defender influence

Probability of ball staying in the field

To calculate the probability of the ball staying in the field for each of 1600 (52 x 30) equally sized zones, we applied the methodology described in Section 4.2.1. The layer for the situation from the 11*th* minute and 52*nd* second of the PSV - AFC Ajax game is presented in Figure 5.4.

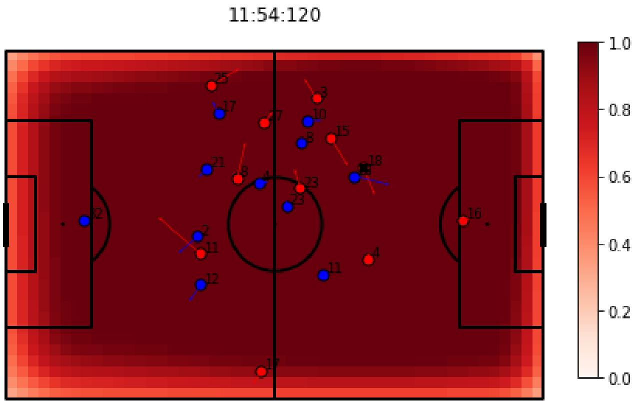


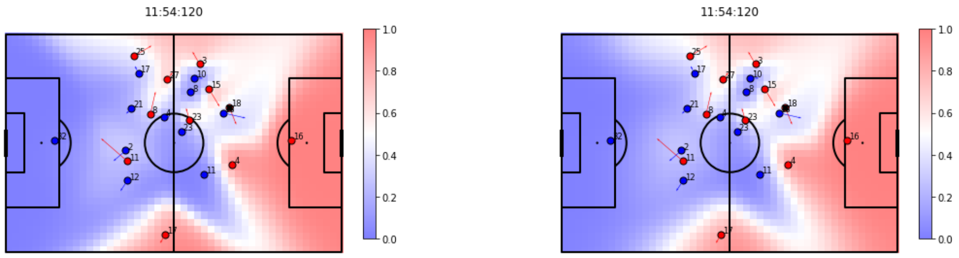
Figure 5.4: Layer with the probability of ball staying in the field

Similarly to the closest defender influence layer, the darker the colour, the higher the probability that if the player aims at a specific zone, the ball stays in the field.

Pass error and Potential Pitch Control Field

We implemented the procedure described in Section 4.2.1. The comparison between the Potential Pitch Control Field (*a*) and smoothed Potential Pitch Control Field (*b*) is shown in Figure 5.5.

In a similar fashion as in the previous comparison between Potential Pitch Control layers, red zones represent zones controlled by PSV, blue ones represent ones controlled by Ajax and white zones represent areas controlled to the same degree by teams.



(a) Potential Pitch Control Field

(b) Smoothed Potential Pitch Control Field

Figure 5.5: The difference between Potential Pitch Control Field and smoothed Potential Pitch Control Field

Combining layers into pass probability

As described in Section 4.2.1, we multiplied smoothed Potential Pass Control layer, closest defender influence layer and the layer with the probability that the ball would stay in the field. We visualised the procedure together the result for the analysed situation, which is provided in Figure 5.6.

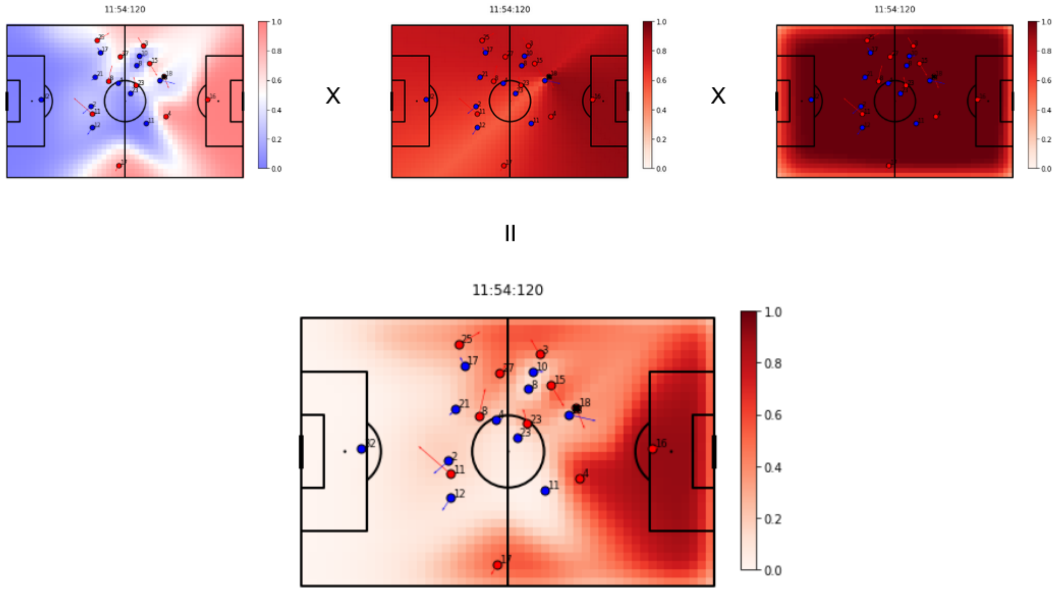
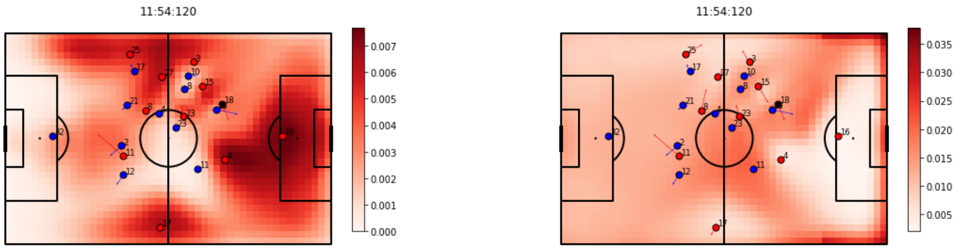


Figure 5.6: Physics-based pass success probability

As in the previous sections, the darker the colour of the zone, the higher the pass success probability.

5.2.2 Pass value

We estimated the reward and risk of each pass using the formulas presented in Section 4.2.2. The reward (a) and risk (b) for each possible pass during the situation mentioned in previous section is presented in Figure 5.7.



(a) Reward of the pass

(b) Risk of the pass

Figure 5.7: Physics-based risk and reward of the pass

Then, we subtracted the risk from the reward, achieving the pass value, shown in Figure 5.8.

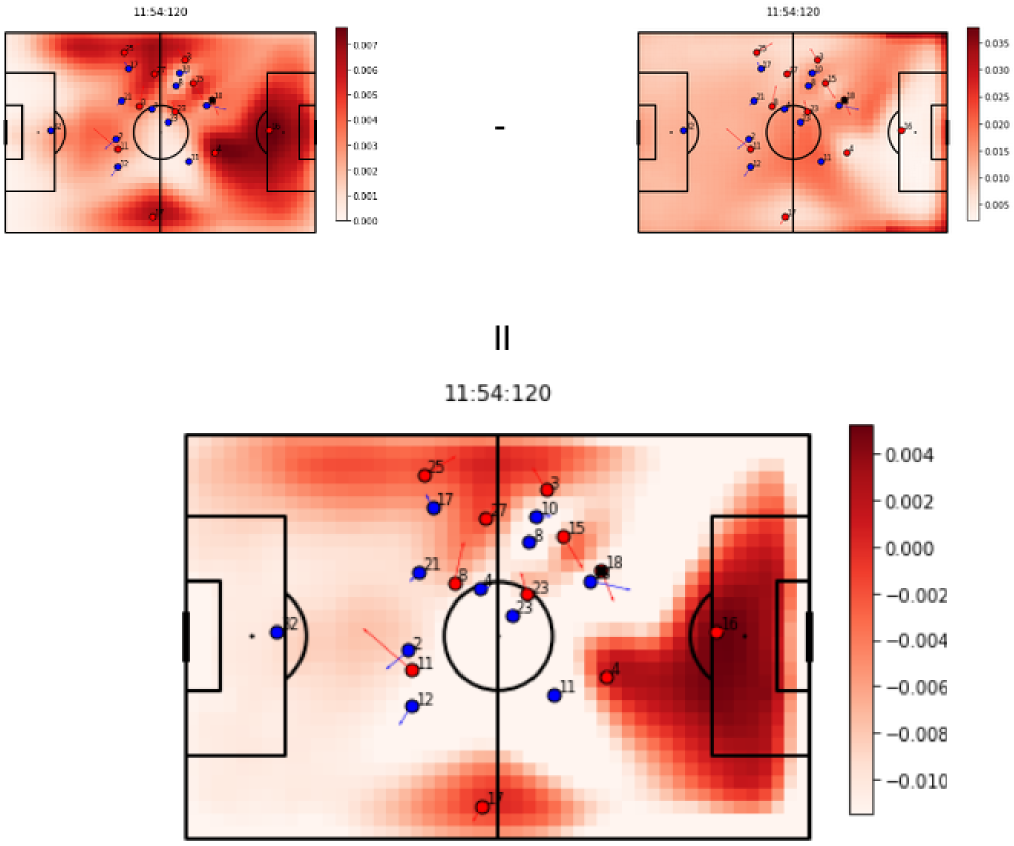
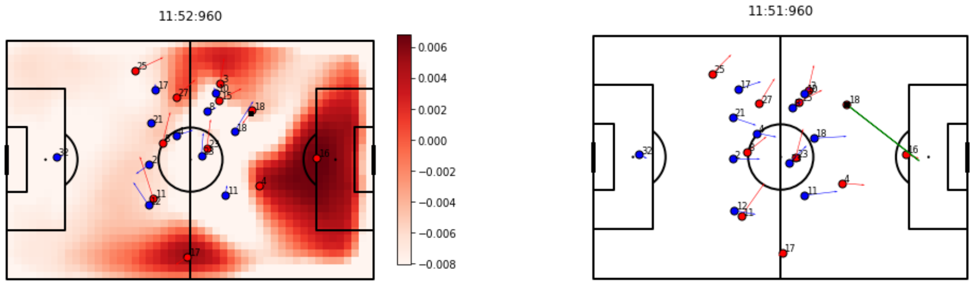


Figure 5.8: Physics-based pass value

5.2.3 Finding potential passes

As previously described in Section 4.2.3, as the end location of the potential pass we assigned the zone where the value is maximized for the frame when the value was the highest. In this case, our approach suggested making a pass 0.4 seconds earlier than it happened during the game. For the analyzed situation, the end coordinates of a potential pass are $x = 91.35, y = 32.94$. The potential pass is visualized in Figure 5.9 (b) together with its value (a).



(a) Physics-based value of a potential pass (b) Potential pass

Figure 5.9: Potential pass estimated with physics-based approach

We assigned the receiver of the potential pass the player with number 16 playing for PSV, J. Drommel, since his individual Potential Pitch Control Field was the highest, as shown in Figure 5.10.

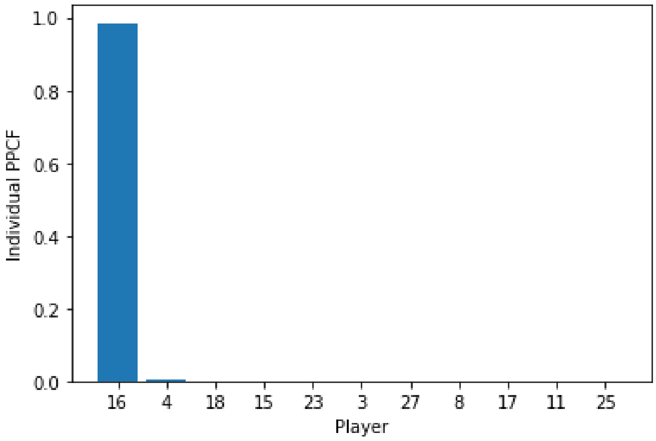


Figure 5.10: Potential Pitch Control for individual players of the attacking team

5.2.4 Creating pass network

We found the potential pass for each situation when a pass was made during the PSV - AFC Ajax, which happened on 23.01.2022. Then, we created a potential pass network from the passes as described in Section 4.2.4. The potential pass network for AFC Ajax from the analyzed game is presented in Figure 5.11.

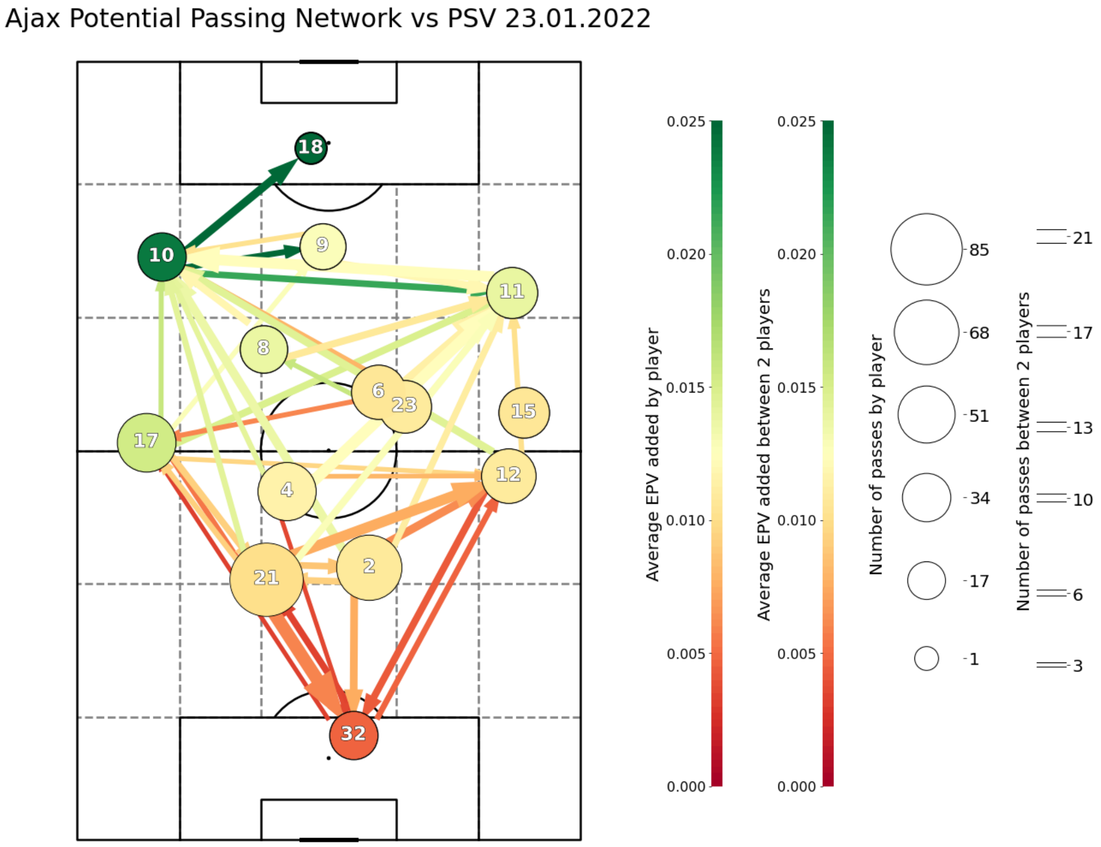


Figure 5.11: Potential Pass Network for AFC Ajax created with the physics-based approach

5.3 Machine Learning approach

5.3.1 Pass success probability

To estimate the pass success probability given the current situation on the pitch, we implemented the neural network with the SoccerMap architecture as described in Section 4.3.1 to estimate the pass success probability. The results compared to results from J. Fernandez et al. are presented in Table 5.2 [14].

	Loss	ECE
our implementation	0.261	0.0070
Fernandez et al.	0.190	0.0047

Table 5.2: The average loss and calibration value for the pass success probability model

The outcome of the model for the analyzed situation is presented in Figure 5.12.

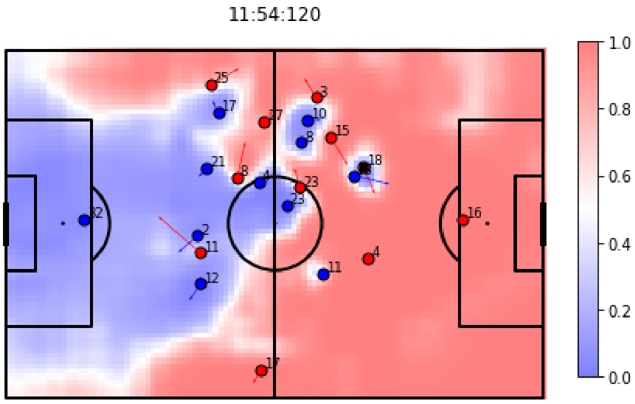


Figure 5.12: Pass success probability estimated with machine learning method

5.3.2 Pass value

We implemented two models, one for successful and the other one for unsuccessful passes, as described in Section 4.3.2. The results compared to results from J. Fernandez et al. are presented in Table 5.3 for successful passes and Table 5.4 for unsuccessful ones [14].

	Loss	**ECE**
our implementation	0.0092	0.0021
Fernandez et al.	0.0075	0.0011

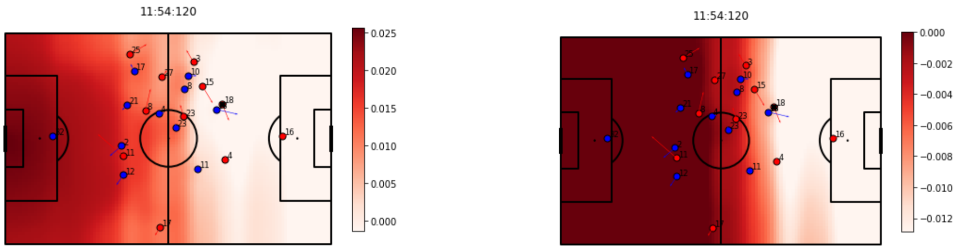
Table 5.3: The average loss and calibration value for the pass value model for successful passes

	Loss	**ECE**
our implementation	0.0087	0.0019
Fernandez et al.	0.0085	0.0015

Table 5.4: The average loss and calibration value for the pass value model for unsuccessful passes

Then, we calculated the pass value given the current situation for successful (*a*) and unsuccessful (*b*) passes, which is provided in Figure 5.13.

The darker the colour, the higher the value.



(a) The value of successful pass

(b) The value of unsuccessful pass

Figure 5.13: The value of successful and unsuccessful pass estimated with machine learning approach

As the next step, we calculated the expected value of the pass given the current situation, which is shown in Figure 5.14.

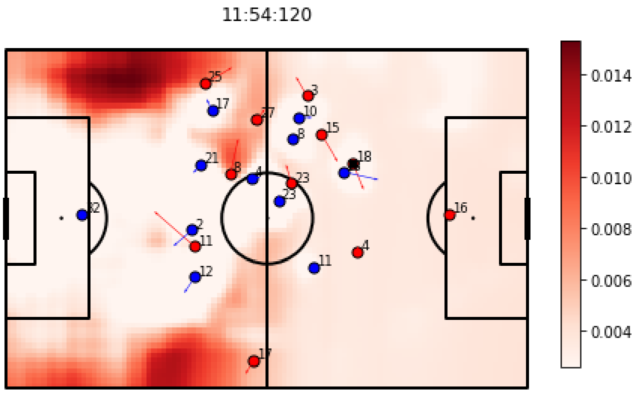


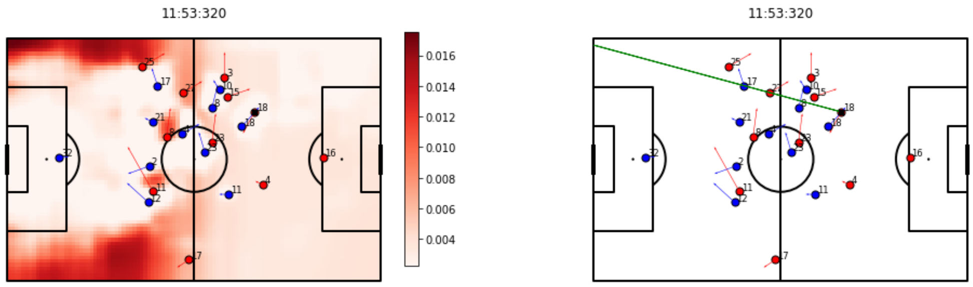
Figure 5.14: The value of a pass approximated with machine learning methods

Similarly to the previous figures, areas with darker colour represent zones with higher values.

5.3.3 Finding potential passes

To approximate the potential pass, we investigated all frames when the player had control over the ball and chose the one with the highest pass value (for more details, see Section 4.3.3). In this situation, we selected the pass which happened 0.8 seconds before the pass was made since this was the one with the highest value. The pass and its value are shown in Figure 5.15.

We assigned the player with number 25, R. Doan, as the receiver of the potential pass since they were the closest to the end location of the pass.



(a) Value of potential pass (b) Potential pass

Figure 5.15: Potential pass estimated using machine learning approach

5.3.4 Creating pass network

Similarly to the physics-based approach, we found potential passes for each passing situation during the analysed game. Then, we aggregated them and visualised them in the form of a network (for more details, see Section 4.3.4). The potential pass network for AFC Ajax from the analysed game is presented in Figure 5.16.

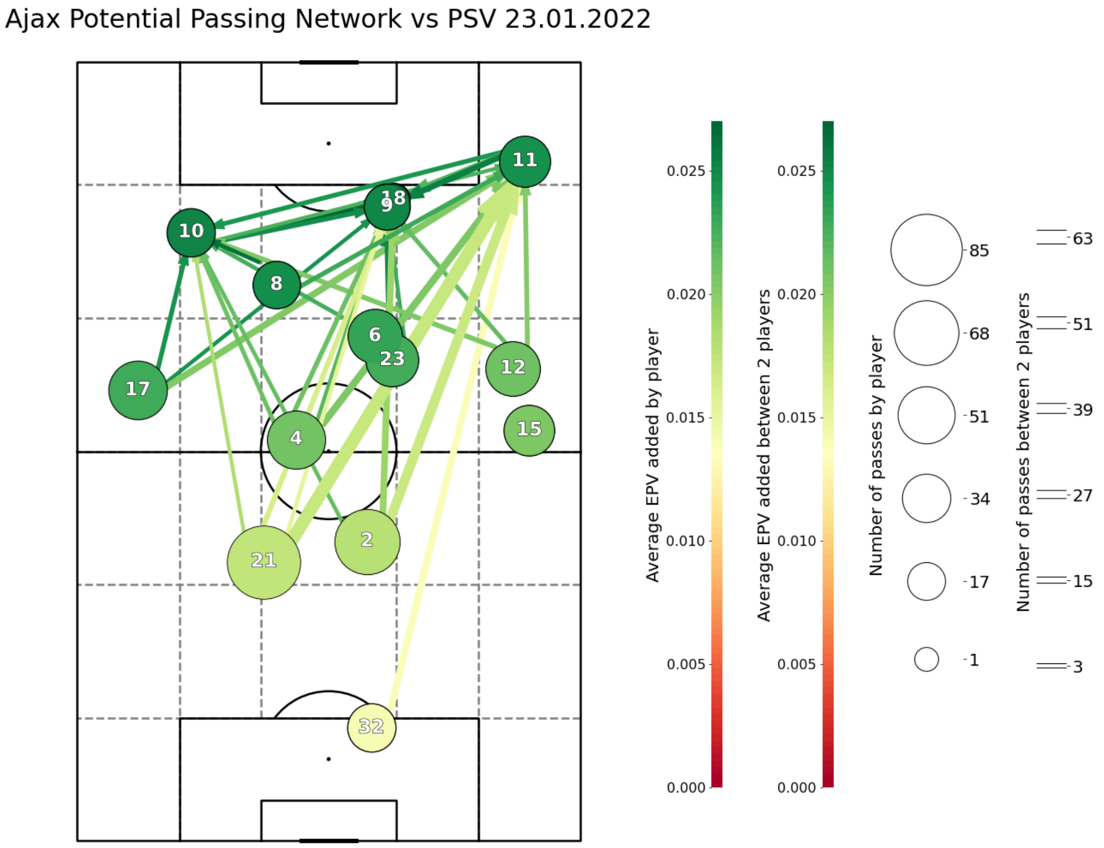


Figure 5.16: Potential Pass Network for AFC Ajax created with the machine learning approach

6 Discussion & Analysis

6.1 Data

The synchronization procedure yielded superior results than the technique suggested by J. Michalak, who proposed adding the offset between the event and tracking data to event timestamps [17]. Their approach yielded perfect synchronization for about 75 % of the investigated frames. Additionally, it did not provide information concerning the start of contact and the subsequent player′s start of contact.

Furthermore, using tracking data coordinates for pass location allowed us to better calculate features for both the closest defender model and SoccerMap. This improved accuracy in feature calculation enhances the reliability of the models used in this study.

6.2 Physics based approach

6.2.1 Pass success probability

Potential Pitch Control Field

One of the problems with the parameters used by L. Shaw is that they tend to overvalue the control probabilities on the wings for the attacking team, far away from the attacking player. This occurs because the defenders are often slow to cover these areas effectively if their maximal velocity was set to $5m/s$. Using a different maximum speed that the player can achieve, the defending team has a higher control probability for the zones closer to the sideline, far away from the attacking player. Therefore, the control probability presented in Section 5.2.1 may provide a value closer to the true one.

However, it is worth mentioning that, firstly, W. Spearman estimated the parameters using a set maximum speed of $5m/s$ and maximal acceleration of $7m/s^2$ [6]. Therefore, the parameters that maximize the log-likelihood may differ for different physical parameters. Secondly, when a pass is made, not all players will accelerate to their highest possible speed in the ball's direction. Nonetheless, the maximal speed may differ between players.

Closest defender model

Although the evaluation metrics of the model, presented in Section 5.2.1, are lower than in the pass success probability models in the literature [10], we argue that this is not a significant concern. The features utilized in the model focus only on the relationship between the ball and one attacking and defending player. Moreover, the model is better than the baseline - average pass success. Furthermore, as expected, the distance between the attacking and defending player and the angle between the direction the ball was played and the defending player were the most important features. The layer provides a valuable approximation of the angles that were covered by the defending players when the pass was made.

Moreover, the model provides a possibility to provide a deeper analysis than just a probability layer. Figure 6.1 presents how B. Brobbey, the closest defender in the analyzed situation, influenced the probability of a successful pass while O. Boscagli had control over the ball.

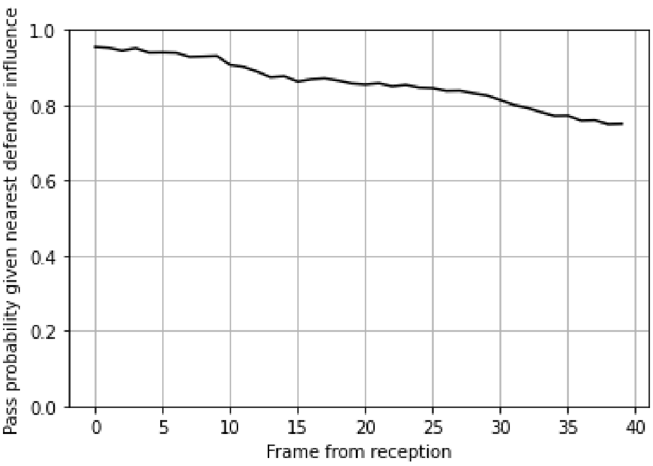


Figure 6.1: Pass probability over the time O. Boscagli controlled the ball

Furthermore, this model enables the analysis of how the defender's position could influence the pass success probability to a specific zone. For the analyzed situation, we can examine the impact of B. Brobbey's position on the probability of a successful pass to the exact end location. This analysis is illustrated in Figure 6.2. In the figure, it is assumed that the defender runs with the same speed as during the game ($3.33m/s$) directly at the player with the ball.

On the other hand, as mentioned in Section 4.2.1, we are aware that multiplying layers, which may not be perfectly independent, is not mathematically correct. However, this approach produces a visually interpretable layer with angles covered by the defending players. Also, the data provided by ORTEC did not include information on whether a pass was blocked or intercepted. Incorporating this information into the calculations may have led to a more precise model without the bias from passes that were kicked out.

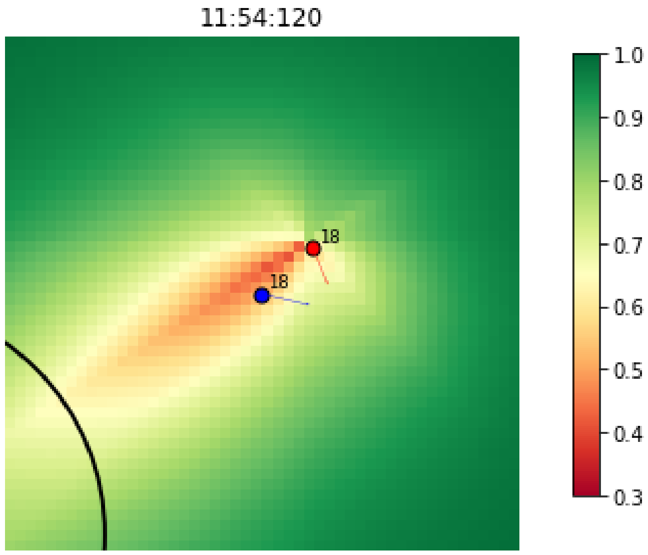


Figure 6.2: Pass probability given defender's position

Probability of ball staying in the field

The layer for the probability of the ball staying in the field, presented in Section 5.2.1, resulted in an intuitive outcome. If a player aims the pass close to the corners or the sidelines, the probability that the ball would stay in the field is the lowest. Additionally, longer passes were more likely to be kicked out of bounds.

However, no data is available regarding the zone the player aimed at. Therefore, this layer is just an approximation derived from experts' assumptions. If there were a possibility to collect the necessary data, i.e. where precisely the player wanted to pass to, we would be able to quantify the parameters of the distribution by maximizing the log-likelihood. Moreover, we could approximate the values with respect to the pass length and the angle at which it was played.

Pass error and Potential Pitch Control Field

Our approach yields a similar outcome to the Potential Pitch Control Field. However, as presented in Section 5.2.1, the defending team is even more likely to control long balls close to the sideline. Additionally, the control probability of an individual player isolated on the wing is lower since the ball can likely be played to an area they do not control.

Conversely, similarly to the layer with the probability of the ball staying in the field, the distribution of the passes is just an approximation since there is no real-world data regarding the exact location the player aimed at.

6.2.2 Pass value

Our model overcomes the problem of overvaluing long passes near the sidelines. Moreover, after analyzing multiple situations with the football experts from AFC Ajax, we observed a correlation between the approximated optimal end location of a pass and their choices. However, there is no objective measure of the optimal pass in every situation. In some situations, the team wants to play riskously and maximize the probability of scoring the goal, while in other instances, the main objective is not losing the ball.
In the situation analyzed in Section 5.2.2, a pass to the goalkeeper is identified as providing the highest value. Moreover, a pass to the player with the number 4 is also considered a pass which provides high value. Passes to both right and left wing generate lower value since they are less likely due to lower pitch control on the right side and Brobbey's pressure, which reduces the possibility of a successful pass to the left side. Importantly, structuring the pass probability and pass value problems in an understandable way helps explain why certain passes should have or should not have been selected.
There are also disadvantages to this approach. Football is a dynamic game. Based on the situation, the value of the pass and interception in the same zone may lead to different goal scoring probabilities. A pass between two lines to the same zone provides a higher value than an average one. Intercepting the ball with only one defender in front may lead to a better chance within the next few seconds than intercepting a ball with the entire team behind the ball. Moreover, a cross was suggested using positional expected threat in situations when plenty of players were in the penalty box due to pitch control in this area. Due to their style of play, some teams prefer not to cross the ball as their strikers can't win areal duels. A situation when a pass to a striker was recommended due to the number of attacking players in the penalty box is illustrated in Figure 6.3

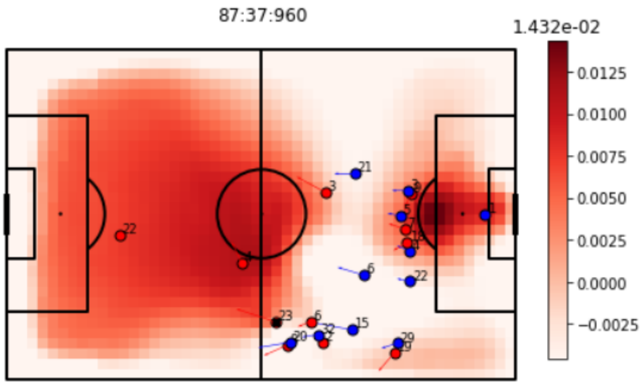


Figure 6.3: Situation when cross to the box was the most valuable pass

6.2.3 Finding potential passes

The approach that we adopted enables us to find potential passes throughout the entire duration of player control over the ball. It provides more information about the players they could have played the ball compared to solely focusing on the frame when the pass was made. Therefore, our approach helps identify missed passing situations because of a higher number of analyzed frames.

However, despite being included in the closest defender model, the time to pass after controlling the ball is less relevant than expected. Usually, football players need some time to fully control the ball and make a pass. Therefore, we expected that the potential pass would not be suggested in the first possible frames. This is presented in Section 5.2.3, as the same pass is suggested 1.5 seconds earlier. Nevertheless, since time was not a very relevant feature, the approach recommended a pass in early frames when the player had control over the ball, and the pressure had not started yet.

6.2.4 Creating pass network

Aggregating all potential passes in the form of the network provides valuable insight regarding which players could have passed to each other, which players were available to receive passes and the value each player could have created. This approach proves to be a useful tool when compared to the pass network. After the game, coaches and analysts can analyze which players could have received more passes, which players missed the most dangerous passes or passing directions the players could have investigated more frequently. Then, the tool would help to find and investigate the most crucial situations during the game.

The network presented in Section 5.2.4 provides an interesting result. As expected, the passes from the goalkeepers and defenders are less valuable than ones from attacking players. Moreover, the network is balanced; there is no overwhelming majority of potential passes forward to the wingers or strikers. Also, there is not one direction or combination of passes from one player to the other that stands out.

The hypothesis which can be derived from this potential pass network is the availability of strikers to receive passes. They could have received passes of high value. However, this network alone does not provide enough context for the coaches. Thus, based on the hypotheses stated from the network, one can quickly query specific situations when the strikers could have received the ball or Ajax's number 10, D. Tadić, could have made a valuable pass. Then, situation analysis takes less time since the timestamps of crucial situations are available.

One of the downsides of this method is the time complexity. The game between AFC Ajax and PSV, which consisted of 870 passes, required multiple hours on a personal device to find all potential passes, even though every fifth frame was analyzed.

6.3 Machine Learning approach

6.3.1 Pass success probability

Our implementation of the original framework produced worse results than the original implementation by J. Fernandez et al. [14]. However, the authors used data from two seasons of the Premier League, which were provided by the same company. Therefore, their model is more robust to the synchronization errors. Additionally, the company may have used a different definition of a successful pass than ORTEC. However, our implementation still achieved significantly lower loss than each of the baselines presented in J. Fernandez's and L. Bornn's work [8].

In our opinion, this approach tends to overvalue the probability of long balls played down the wing, as presented in Section 5.3. Also, similarly to the physics-based method, it does not account for the body orientation of the player. Moreover, it is not explainable. It is challenging to determine why a pass to a certain player was considered likely or not. Furthermore, since we learn the model from the pixel where a pass ended, the approach suffers from the survivorship bias. If a player aimed to a specific location, but the ball was intercepted by a defender, the model would learn that the pass to the intercepted location was unlikely instead of considering the location where the player aimed the ball at. Also, in our implementation, we did not exclude offside players.

Since the model is black-box, we can not easily state why the values down the wing are high. One possible hypothesis is that since the neural network learns from the end location of the pass, the winger can control the ball in these areas since they are closer to the potential end location of the ball than the defender. However, it is challenging to state if the situation on the ball enables easy delivery of the ball to these areas.

6.3.2 Pass value

Similarly to the implementation of the pass success probability model, our models resulted in higher loss and ECE than implementation by J. Fernandez et al. [14]. However, we claim that the difference is not high, as presented in Section 5.3.2. Moreover, the values on the pitch, presented in Section 5.3.2, reflect our assumptions about football. A successful pass closer to the goal would lead to a higher probability of scoring the goal within the next 15 seconds. Similarly, losing the ball far away from the goal is not as dangerous as losing it close to the own goal.

However, similarly to the pass probability approach, it is hard to explain the reasons behind the value since the method is black-box. Moreover, since we claim the pass success probability on the wings is overestimated, the expected value of the passes is inflated.

6.3.3 Finding potential passes

Similarly to the physics-based approach, our analysis considered the entire time during the player's control over the ball. Different frames resulted in a different approximation of the optimal target of a pass. As in the physics-based approach, the time to pass from controlling the ball to the pass was not an essential factor. Moreover, while using a physics-based approach allowed us to explain why a pass earlier could provide a higher value, these explanations are impossible with the black-box approach.

In the example presented in Section 5.3.3, the pass was suggested earlier than it was actually made. It was played to the player who was offside in that specific moment. Moreover, we argue that it would be impossible for the passing player to pass exactly to the sideline without any margin of error. Even a slight inaccuracy would lead to the ball being kicked out.

6.3.4 Creating pass network

Using the machine learning approach, the passing network presented in Section 5.3.4 indicates that the passes to the strikers or wingers are often overvalued. The model suggests that most passes should have been played to number 11, Antony, number 10, Tadić, number 18, Brobbey, and number 9, Haller, who were attacking players for AFC Ajax. Therefore, it is hard to utilize this network as a useful application to provide value for the football organization.

However, the relatively low time complexity is one advantage of the machine learning approach. Once the model is trained, analyzing every frame between the start of the contact with the ball and the moment of the pass for each of the 870 passes took less than 3 hours.

7 Conclusion

To conclude this paper, we provide answers to the research questions presented in Section 1.

We introduced potential pass networks, a visualisation technique that aids in identifying passes which could have happened during the game, given that the players had always chosen the most valuable option. Importantly, this tool should not be used without context but is helpful to make hypotheses. Based on the hypotheses stated, it allows us to quickly query crucial situations during the game. Moreover, one should be aware that the players will not always choose the most valuable option, since they would like to reorganise and make a more valuable pass within the next couple of seconds. Therefore, it is not straightforward that certain players should have or should not have passed between each other more or less frequently.

With the created closest defender influence model, we identified areas on the pitch which were the most and the least available, given the opposing team's pressure on the ball. Moreover, we found different interesting applications of this model, which can allow the teams to analyse situations in-depth. Additionally, we introduced the concept of the error associated with the pass, which helped us reduce the overestimation of the probability of long passes near the sidelines. However, we claim it would be even more precise if the data regarding the zones, where players aimed the ball at, were available. These two additional layers helped us achieve the more precise pass success probability layer while maintaining explainability.

Moreover, the value of the pass using both risk and reward provides a more comprehensive approximation of the most valuable zones since it considers the potential risk of the rival intercepting the ball. However, the use of the action-based approach, instead of the position-based one, would provide a better estimation of the pass value since football is a dynamic game.

The outcome of the physics-based approach and machine learning approach differs a lot. Our physics-based method does not overvalue long passes close to the sideline and is easily explainable to the non-technical coaching staff. Moreover, it correlates with experts' choices concerning the most valuable pass. However, it is not as computationally efficient as the machine learning method. Neural network with SoccerMap architecture overestimates the probability of the ball being played to the wing and suffers from survivorship bias. Moreover, since it is black-box, it does not answer the questions concerning the factors making the pass more or less probable. However, the value of the action is action-based and more relative to the current situation on the pitch. Both state-of-the-art approaches to evaluating possessions, On-Ball Scoring Opportunity and the framework by J. Fernandez et al., introduced the transition probability model [6] [14]. The frameworks are useful to evaluate the possession value given the current situation on the pitch. Still, they can provide biases towards the most frequently chosen passes as they use the transition probability models. Our approach can help evaluate individual passing options, not only possessions.

8 Future work

To further improve on the approach to evaluate potential passes, we recommend the implementation of an action-based value model for successful and unsuccessful passes instead of a position-based one. This approach would provide better insights into the quality of the pass since football is a dynamic game.
Moreover, we suggest incorporating a probabilistic approach to players' movement when calculating the Potential Pitch Control Field. This could help to account for the fact that when a pass is played, not all players will sprint towards the ball's estimated destination, as the current implementation suggests.
Furthermore, with increased computational power, we propose investigating the interception and control probabilities based on the pass end location. Given these calculations, such a layer of pass success probability can be created, which would be more precise than the Potential Pitch Control Field.
We also suggest investigating the potential passes using a machine learning approach using event and tracking data provided by the same provider since possible synchronization problems are a potential threat to the validity of our implementation. Moreover, it would be interesting to investigate different machine learning techniques, such as graph neural networks.
Additionally, we suggest using reinforcement learning techniques in modelling since a pass may not provide an immediate value but may lead to maximizing it within the next couple of actions or seconds.
Last but not least, we recommend that clubs internally collect the data concerning the error associated with the pass. One of the potential data collection ideas is to conduct multiple experiments with players aiming at the object and estimating the deviation from the target.

Literature

[1] P. Gould and A. Gatrell, “A structural analysis of a game: The Liverpool v Manchester United cup final of 1977,” *Social Networks*, vol. 2, no. 3, pp. 253–273, January 1979.

[2] J. Duch, J. S. Waitzman, and L. A. N. Amaral, “Quantifying the Performance of Individual Players in a Team Activity,” *PLoS ONE*, vol. 5, no. 6, E. Scalas, Ed., e10937, Jun. 2010.

[3] J. M. Buldú, J. Busquets, I. Echegoyen, and F. Seirullo, “Defining a historic football team: Using Network Science to analyze Guardiola’s F.C. Barcelona,” *Scientific Reports*, vol. 9, no. 1, Sep. 2019.

[4] P. Power, H. Ruiz, X. Wei, and P. Lucey, “Not All Passes Are Created Equal,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, August 2017.

[5] W. Spearman, A. Basye, G. Dick, R. Hotovy, and P. Pop, “Physics-Based Modeling of Pass Probabilities in Soccer,” in *Proceedings of MIT Sloan Sports Analytics Conference*, 2017.

[6] W. Spearman, “Beyond Expected Goals,” in *Proceedings of MIT Sloan Sports Analytics Conference*, 2018.

[7] F. P. Alguacil, J. Fernández, P. P. Arce, and D. J. T. Sumpter, “Seeing in to the future: Using self-propelled particle models to aid player decision-making in soccer,” in *Proceedings of MIT Sloan Sports Analytics Conference*, 2020.

[8] J. Fernández and L. Bornn, “SoccerMap: A Deep Learning Architecture for Visually-Interpretable Analysis in Soccer,” in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, Springer International Publishing, 2021, pp. 491–506.

[9] G. Anzer and P. Bauer, “Expected passes,” *Data Mining and Knowledge Discovery*, vol. 36, no. 1, pp. 295–317, January 2022.

[10] P. Robberechts, M. V. Roy, and J. Davis, “Un-xpass: Measuring Soccer Player’s Creativity,” in *Proceedings of Statsbomb Conference 2022*, 2022.

[11] U. Dick, D. Link, and U. Brefeld, “Who can receive the pass? A computational model for quantifying availability in soccer,” *Data Mining and Knowledge Discovery*, vol. 36, no. 3, pp. 987–1014, March 2022.

[12] S. Rudd, “A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains,” in *New England Symposium on Statistics in Sports*, 2011.

[13] K. Singh, *Introducing expected threat (XT)*, https://karun.in/blog/expected-threat.html, Accessed on 23.04.2023 11:01, 2018.

[14] J. Fernández, L. Bornn, and D. Cervone, “A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions,” *Machine Learning*, vol. 110, no. 6, pp. 1389–1427, May 2021.

[15] L. Shaw, *Advanced football analytics: building and applying a pitch control model in python*, https://www.youtube.com/watch?v=5X1cSehLg6s&t=1102s&ab_, Accessed 23.04.2023 12:10, Youtube, 2020.

[16] D. Sumpter, *Soccermatics*, https://soccermatics.readthedocs.io/en/latest/lesson7/OffBallRuns.html, Accessed on 27.04.2023 18:29, 2022.

[17] J. Michalak, “Identifying football players who create and generate space,” M.S. thesis, Uppsala University, 2022.

[18] A. Fujimura and K. Sugihara, “Geometric analysis and quantitative evaluation of sport teamwork,” *Systems and Computers in Japan*, vol. 36, no. 6, pp. 49–58, 2005.

[19] G. Rolland, *Analyzing liverpool attacks*, https://gabs-rol43.medium.com/analysing-liverpool-attack-e6d32c6c9a57, Accessed 18.05.2023 19:01, Medium, 2020.

[20] M. Bruinsma, “Counter press in elite football. A quantification of the if, how and why questions,” M.S. thesis, Universiteit van Amsterdam, 2020.

[21] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, JMLR.org, 2017, pp. 1321–1330.

[22] G. F. Jens, “Optimal data classification for choropleth maps,” *Occasional paper. University of Kansas*, 1977.

Appendices

A SoccerMap architecture

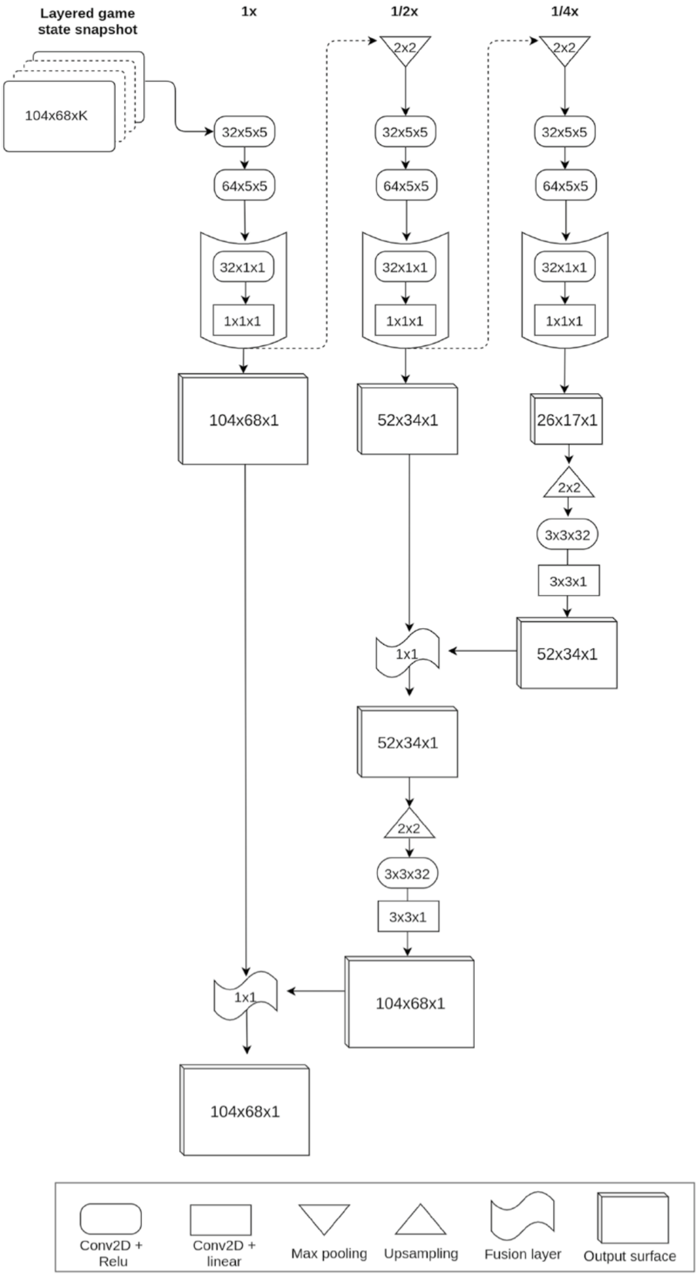


Figure A.1: SoccerMap architecture [**Fernandez et al. 2021**].

B Layers used in the SoccerMap neural network

Feature	PP	PV
1 for possession player location (x, y)	x	x
Possession team players' speed (m/s) (x)	x	x
Possession team players' speed (m/s) (y)	x	x
1 for defending player location (x, y)	x	x
Defending team players' speed (m/s) (x)	x	x
Defending team players' speed (m/s) (y)	x	x
Angle between every location and the goal	x	x
Standardized distance between every location and the goal	x	x
Standardized distance between every location and the ball	x	x
Sine of the angle between every location and the ball location	x	
Cosine of the angle between every location and the ball location	x	
Sine of the angle between carrier velocity vector and the every location	x	
Cosine of the angle between carrier velocity vector and the every location	x	
Index of the closest possession team line to closest location		x
Index of the closest opponent team line to closest location		x
Number of possession team's players between the ball and every other location		x
Number of defending team's players between the ball and every other location		x
Number of possession team's players between the opponent's goal and every other location		x
Number of defending team's players between the opponent's goal and every other location		x
Pass probability surface		x

Table B.1: Layers used in the SoccerMap neural network models for pass probability (PP) and pass value (PV) **[Fernandez et al. 2021]**