Sports Analytics:
2vs1 Situation Analysis in Football

Bhautik Lukhi
(Matriculation No: 4291276)
Faik Erkam Minsin
(Matriculation No: 4125891)

Client: Fabio Casalnuovo, Data Analyst, FC Schalke 04

Course: Statistical Consulting - 314000

Examiner: Rouven Michels

Date of Submission: 29th September 2024

1 Introduction

This report presents the findings of a comprehensive consulting project conducted for Fabio Casalnuovo, a Data Analyst at FC Schalke 04. The objective of the project was to generate actionable insights from player data during 2vs1 situations, a crucial scenario in football where two attacking players face one defender. The analysis was undertaken as part of the Statistical Consulting course under the supervision of the Rouven Michels.

The primary focus was on analyzing double passes in 2vs1 situations using Bundesliga match data. By studying these interactions, we aimed to provide FC Schalke 04 with insights to refine tactical approaches, improve decision-making, and enhance player performance during matches. Specifically, we examined how players used double passes to bypass defenders and create goal-scoring opportunities, with data-driven insights guiding the recommendations.

2 Literature Review

Recent advancements in sports analytics have fundamentally transformed the way football matches are analyzed. Tracking data, which captures the real-time positions of players and the ball, has enabled teams to gain deep insights into game dynamics and tactical decisions. Studies have demonstrated the significant role tracking data plays in providing a competitive edge by identifying patterns in player movement and spatial control. [2].

In football, 2vs1 situations are particularly important in areas where quick decision-making is crucial Peters and Schumacher, in their book ZWEI GEGEN EINS: Starke Entscheider auf dem Platz, [5] present practical strategies for creating and training 2vs1 situations. They emphasize that consistent tactical positioning and rapid decision-making are essential for taking full advantage of such situations on the pitch.

Further expanding on this, these situations involve two attacking players facing one defender, where the attackers must swiftly exchange passes and position themselves to outmaneuver the defender. [8] A successful exploitation of 2vs1 situations can significantly increase a team's chances of breaking through defensive lines. It highlights that spatial awareness and the ability to read the defender's movements are critical skills required for effective decision-making in these scenarios.

A central aspect of breaking down defensive structures in football is the use of intricate passing networks. Grund T(2012) [6] discuss how passing networks allow teams to maintain possession while advancing the ball strategically into the attacking third. Double passes—where the ball is quickly exchanged between two players—play a key role in these networks.

3 Data Presentation

The dataset utilized in this project was sourced from a Bundesliga match between RB Leipzig and Eintracht Frankfurt, played on January 13th, 2024. These datasets are part of the Official Match Data collected by the Deutsche Fußball Liga (DFL). Positional data, which records the location of the players and the ball every 40 milliseconds, is a critical component of our analysis, as it enables a granular examination of player movements and decisions in real time. Each match generates approximately 3.6 million positional data points, while around 1,600 match-related events, including goals, passes, and fouls, are captured in the event data. [1]

3.1 Preprocessing

The initial step involved cleaning and aligning the data. The timestamps in both the event and position data were converted to a common timezone (UTC) to ensure synchronization.

In the **position data**, the player coordinates followed a Cartesian coordinate system, where the X and Y values represent relative positions on the pitch. The summary statistics indicate that the positions were in a minimum and maximum range of $X \in [-54.50, 55.50]$ and $Y \in [-36.78, 37.76]$. These values are arbitrary and do not directly correspond to the real dimensions of a football field. This representation allows for relative positioning but lacks physical context in relation to the pitch's actual dimensions.

In contrast, the **event data** followed the real dimensions of a standard football pitch. The X-axis, representing the length of the pitch, ranged from 0 to 105 meters, while the Y-axis, representing the width, ranged from 0 to 68 meters. The mean positions were $X = 55.07$ and $Y = 37.03$, aligning closely with the middle of the field. The minimum and maximum values for the X and Y coordinates were consistent with the pitch boundaries, i.e., $X_{\text{min}} = 0$, $X_{\text{max}} = 105$, $Y_{\text{min}} = 0$, and $Y_{\text{max}} = 68$.

To enable a seamless integration of both datasets for analysis, the **event data** was transformed to match the Cartesian coordinates used in the **position data**. The X and Y coordinates of the event data were normalized and rescaled to fit within the same Cartesian coordinate system as the position data. This transformation ensured that the spatial context of player movements could be accurately evaluated in a consistent manner across both datasets.

The following transformations were applied:

- The X-Position was normalized to the range [-52.5, 52.5], and the Y-Position to the range [-34, 34].
- The transformation formula used for the X-Position was:

$$X_{\text{Position}} = \left(\frac{X_{\text{Position}} - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \right) \times 105 - 52.5$$

- Similarly, for the Y-Position:

$$Y_{\text{Position}} = \left(\frac{Y_{\text{Position}} - Y_{\text{min}}}{Y_{\text{max}} - Y_{\text{min}}} \right) \times 68 - 34$$

Finally, to account for the change in direction of play after halftime, the coordinates were mirrored for events occurring after halftime. The halftime timestamp was used as conditional, and all events after this time were mirrored by flipping the sign of both X-Position and Y-Position. Additionally, the play origin (own half or opposition half) was updated accordingly, ensuring consistent analysis across both halves of the match.

This preprocessing step allowed us to create a unified and normalized dataset, ready for further analysis of double passes in 2vs1 situations.

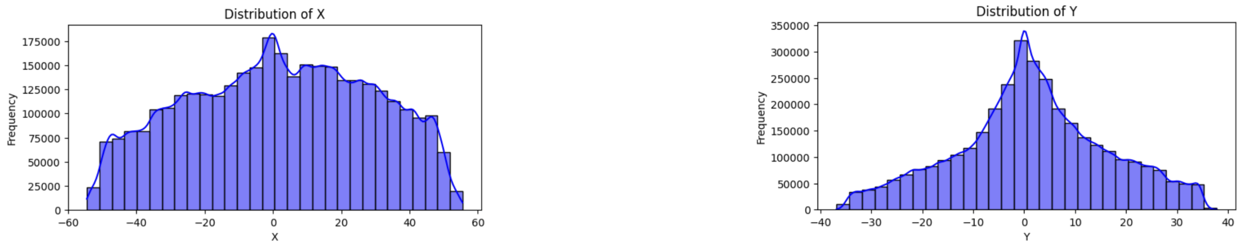
4 Exploratory Data Analysis (EDA)

As part of our analysis of 2vs1 situations, we conducted exploratory data analysis to gain insights into the spatial and temporal distribution of key variables. Below, we provide a detailed interpretation of the distributions of several variables in the dataset, including the player positions and movements.

4.1 Distribution of Variables

We explored several important variables from the dataset, such as X and Y co-ordinates of the players, movement metrics such as S, D, A and Z being the ball height. These variables capture the spatial positioning and dynamic features of players and the ball during matches.The following general interpretations can be made:

- **X and Y (Player Positions)**: Represent player coordinates on the pitch, with higher density near midfield (X = 0) and balanced presence along the vertical axis (Y).
- **S (Speed)**: Shows that players often move slowly, with occasional sprints.
- **D (Distance)**: Indicates shorter movements between events, likely representing quick repositioning during passes.
- **A (Acceleration)**: Suggests gradual directional changes with fewer sharp turns.
- **Z (Ball Height)**: The ball mostly remains on the ground during the passes.



(a) Distribution of X.

(b) Distribution of Y.

Figure 1: Distribution of X and Y (Player Coordinates on the Pitch).

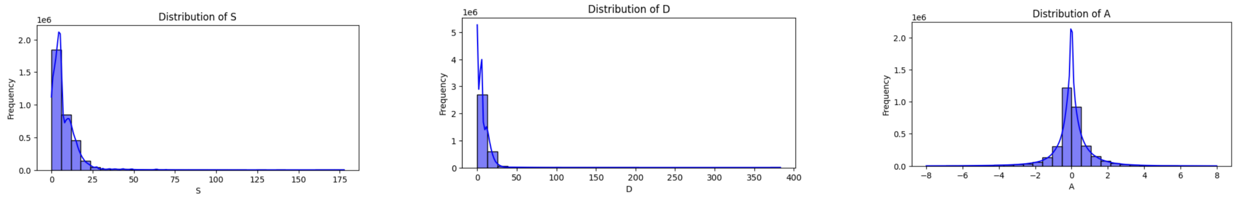


Figure 2: Distributions of Speed, Distance and Acceleration.

5 Summary Statistics

Game ID	Game	Total Double Passes
DFL-MAT-J03YHV	Eintracht Frankfurt vs RB Leipzig	81
DFL-MAT-J03YIV	1. FC Union Berlin vs VfL Wolfsburg	45
DFL-MAT-J03YIW	Bayer 04 Leverkusen vs FC Bayern München	110
DFL-MAT-J03YK5	Sport-Club Freiburg vs Bayer 04 Leverkusen	128
DFL-MAT-J03YK7	1. FSV Mainz 05 vs VfL Bochum 1848	33
DFL-MAT-J03YKL	Borussia Dortmund vs VfB Stuttgart	107

Table 1: Double Passes Identified in Six Bundesliga Games

5.1 Double Pass Identification

To identify double passes, we developed a custom method to analyze consecutive passes between the same players from the same team within a short time frame. We synchronized the event and position data and applied logical conditions to detect the sequence of passes. The process involved pre-processing the event data to filter for passes, aligning timestamps, and mapping passes to the corresponding position data.

The process began by synchronizing both the event and position datasets. Event data provides information about game actions like passes, kicks, shots etc., while position data records the precise locations of players and each moment. Synchronizing these datasets allowed us to accurately map each pass event to the players' positions on the field at that specific time.

Once the data was aligned, we used conditional to detect double passes. We focused on sequences where two consecutive passes occurred between the same two players, from the same team. The criteria ensured that Player A passed to Player B, and shortly after, Player B returned the ball to Player A, forming a rapid double pass exchange. This method allowed us to identify only those passes that fit the characteristics of double passes, excluding regular passing sequences.

5.2 Distribution of Time Spent for Each Double Pass

We analyzed the temporal dynamics of double passes from the DFL-MAT-J03YHV (Eintracht Frankfurt vs RB Leipzig) match. For each double pass, we computed the time difference between the first and second pass. Most passes were completed within 1-2 seconds, while a secondary peak was observed around 3 seconds.

Key Findings:

- Most double passes are completed within 1 to 2 seconds, with durations ranging from 0 to over 7 seconds.
- There is a gradual decrease in frequency for durations beyond 2 seconds, with events lasting over 6 seconds considered outliers.
- A secondary peak was observed around 3 seconds, suggesting strategic or contextual reasons for longer double pass durations.

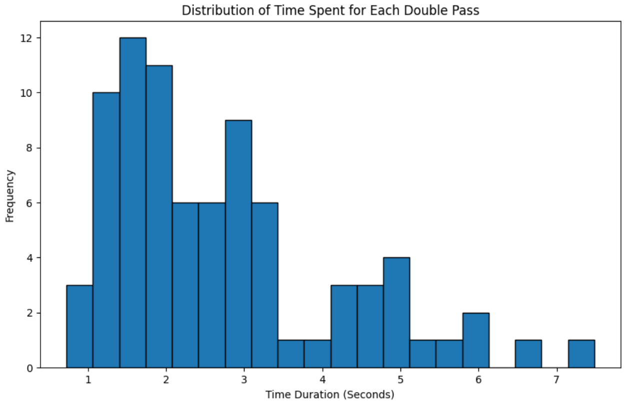


Figure 3: Distributions of Time spent for each Doublepass

5.3 Double Pass Visualizations

In Figures 4 and 5, we visualize the double passes extracted from match data. The arrows represent the direction of the passes, with the color coding indicating the half from which the pass originated. Yellow arrows depict passes that originated in the team's own half, while red arrows depict passes that originated in the opposition half. The visualizations help highlight the spatial patterns of double passes across different areas of the pitch. Notably, there are clusters of activity near the center of the pitch and closer to the attacking third, demonstrating how players exploit space in key areas of the field during 2vs1 situations.

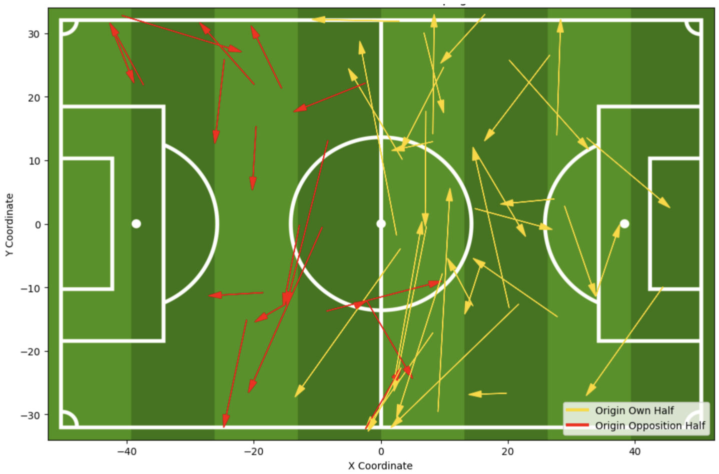


Figure 4: Double passes for RB Leipzig.

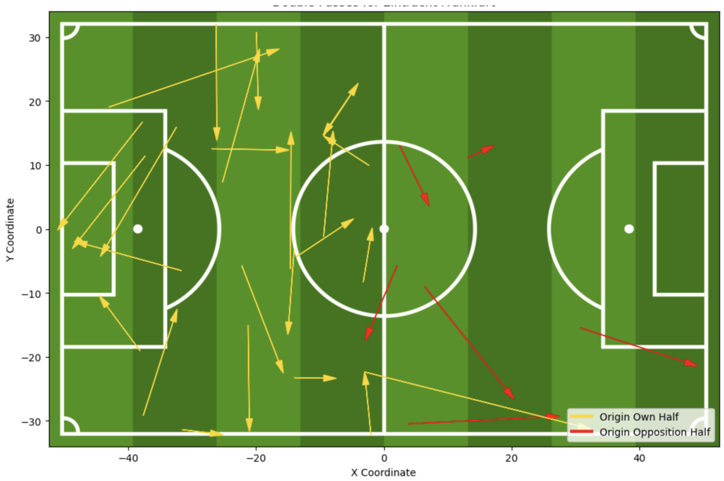


Figure 5: Double passes for Eintracht Frankfurt.

5.4 Distribution of Distance for Each Double Pass

Using the X and Y coordinates of players, we calculated the distances for each double pass. Most passes were within 10 meters, but we observed a secondary peak at 17 meters, potentially indicating specific spacing strategies in 2vs1 scenarios.

Key Findings:

- Most double passes occurred at distances around 10 meters, with very short passes (0-5 meters) being rare.
- A few outliers were observed for longer passes over 30 meters.
- After 10 meters, pass frequency gradually decreased, though an unexpected peak around 17 meters suggests strategic or contextual factors for mid-range passes.

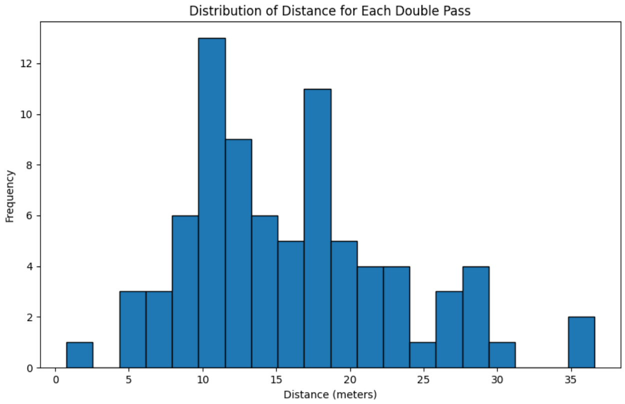


Figure 6: Distributions of Distance for each Doublepass

5.5 Field Movements of Players During Double Passes

We extend the previous analysis of double pass distribution by incorporating the analysis of field movements, including player acceleration and ball movement during each double pass. Specifically, we investigate how players utilize their positioning, movements, and acceleration dynamics to create space and facilitate successful double passes, using the second double pass visualization as a representative example. Our objective is to provide a comprehensive view of how players' spatial dynamics, both in terms of positioning and movement, interact with the tactical utility of double passes across different areas of the field.

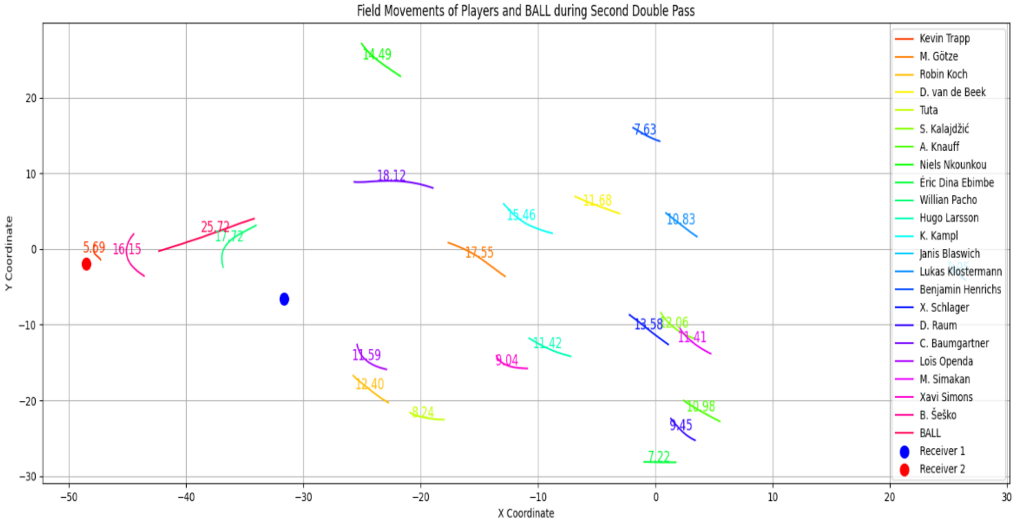


Figure 7: Player movements and ball trajectories during double passes

6 Modeling the Double passes

To model double passes we created a dummy variable for DoublePass. We combined the relevant position data and then labeled the original dataset by adding a new column to indicate whether each event involved a double pass.

The features used for the analysis were X and Y coordinates, speed(S), distance(D), and acceleration(A) of the players involved. To ensure accurate model training, we used standard 70-30 split for training and tasting sets. We then standardized the features to ensure consistency and proper scaling across the model training process. For modeling the double pass events, we applied three machine learning models: Logistic Regression, Random Forest, and Weighted Random Forest. The following features were used: X and Y coordinates, speed (S), distance (D), and acceleration (A). The target variable was a dummy indicator (IsDoublePass) for whether a double pass had occurred or not.

The class distribution of the target variable was highly imbalanced:

- Class 0 (no double pass): 3,254,707 instances.
- Class 1 (double pass): 126,201 instances (around 3.75%).

7 Results

We began with Logistic Regression [3] as a baseline model. Logistic regression is a simple and widely used classification algorithm that models the probability of an event using a linear combination of input features. While it offers easy interpretability, it is limited by its assumption of linear relationships between the features and the outcome. In dynamic scenarios like football, where player movements and interactions are non-linear, logistic regression struggles to capture the complexity of the data. As a result, it failed to detect any double passes (Class 1), yielding a recall of 0%, primarily due to the imbalanced nature of our dataset, where only 4% of events were double passes.

Next, we employed Random Forest [4], an ensemble learning method that constructs multiple decision trees from random subsets of the data. Each tree independently predicts whether an event is a double pass, and the final prediction is based on majority voting among the trees. Random Forest is well-suited for handling non-linear relationships and complex data, making it a better fit for our football tracking data. This model improved recall for Class 1 to 53%, but because it treated all classes equally, the model still favored the majority class (non-double passes).

To better handle the imbalanced dataset, we applied Weighted Random Forest [7], a variation of the Random Forest algorithm that assigns higher importance to misclassifying the minority class (double passes) by adjusting class weights. In this case, we used a 10:1 weighting ratio to increase the model's sensitivity to double pass events. Additionally, we increased the number of decision trees to 200, further boosting the model's ability to generalize. This approach resulted in the best overall performance, with a recall of 61% for Class 1 and a perfect precision of 1.0, making it the most suitable model for predicting double passes. The incremental improvements highlight how effectively adjusting algorithms and considering data imbalance can significantly enhance performance, especially when the minority class is of critical importance.

7.1 Model Performance

The performance of three different models: Logistic Regression, Random Forest (Unweighted), and Weighted Random Forest. Each model was evaluated using precision, recall, and F1-score metrics. One of the key challenges was to improve recall for Class 1 without sacrificing overall model performance.

The following table summarizes the performance metrics for each model:

Model	Accuracy (%)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Logistic Regression	96.26	0.00	0.00	0.00
Random Forest	98.21	0.74	0.53	0.62
Weighted Random Forest	**98.53**	**1.00**	**0.61**	**0.76**

Table 2: Comparison of model performance

7.1.1 Hyperparameter Tuning

In an effort to further improve the model's performance, we applied hyperparameter tuning using two approaches: GridSearchCV and BayesSearchCV. GridSearchCV is an exhaustive search method that evaluates every possible combination of hyperparameters in a predefined grid. While comprehensive, it is highly time-consuming and inefficient, particularly when dealing with large datasets or complex models like Random Forest. In contrast, BayesSearchCV employs Bayesian optimization, which intelligently narrows down the search by focusing on the most promising regions of the hyperparameter space based on previous evaluations. This approach proved to be faster and more efficient, as it learns from each iteration and avoids regions with poor hyperparameters.

However, despite the optimized search, the tuned model performed worse overall, with lower recall and precision for Class 1. This highlights that hyperparameter tuning, while powerful, is not guaranteed to improve model performance universally. The results depend heavily on the quality of the data, its underlying structure, and the problem at hand. In cases of imbalanced datasets like ours, even intelligent search methods like BayesSearchCV may struggle if the data quality or feature relevance is lacking. This subjectivity in performance improvement suggests that hyperparameter tuning can enhance certain models, but it may not always result in better outcomes, especially when recall for minority classes is the primary goal. Therefore, data quality and feature engineering can be just as critical as tuning itself.

8 Conclusion and Reflection

In conclusion, this report provides a comprehensive analysis of 2vs1 situations in football, focusing on the role of double passes as a key strategy in breaking defensive structures. Throughout the analysis, we have combined advanced sports analytics techniques, data-driven insights, and machine learning models. The integration of positional and event data allowed us to explore not only how double passes are executed but also their effectiveness in different phases of play, particularly in midfield and attacking zones.

One key observation is the importance of context in double passes. These passes are not merely technical maneuvers; their success is influenced by player positioning, decision-making speed, and defensive pressure. Although our analysis focused on double pass events in 2vs1 situations, it became evident that many other contextual factors, such as the positioning of nearby defenders or the tactical setup of the opposing team, could impact outcomes. Incorporating such contextual data could significantly improve the predictive power of future models, and it would be an important direction for future research and analytics.

The use of machine learning models provided valuable insights but also revealed the limitations inherent in working with highly imbalanced datasets. While Weighted Random Forest proved to be the most effective model for identifying double passes, its performance underscores the fact that achieving a high recall for minority events in sports data is a complex task. The challenge lies not just in tuning hyperparameters or selecting the right model but also in ensuring that the data being fed into the model captures the true complexity of real-world football scenarios. This brings us to a broader reflection: no model, however well-tuned, can substitute for high-quality, relevant data. Improving data quality—whether through better feature engineering, incorporating more contextual elements, or capturing richer player interactions—remains a cornerstone of effective sports analytics.

Moreover, the exploratory data analysis highlighted important trends in player movements, pass timing, and spatial distribution, particularly in how players adapt to different areas of the field. These insights could be directly applied by coaches and analysts to refine tactical instructions. For instance, the tendency for double passes to occur in less pressured midfield areas suggests that teams might benefit from deliberately engineering 2vs1 situations in tighter, more defensively congested spaces like the attacking third. Understanding how to manipulate space effectively in these key areas is crucial for teams looking to capitalize on such strategic opportunities.

Looking ahead, the conclusions drawn from this analysis could be expanded to include more prediction factors, such as player fatigue, ball control quality, or even external conditions like weather or pitch quality, which could all influence double pass effectiveness. Additionally, further investigation into the defenders' roles in 2vs1 situations—particularly in terms of how their positioning and actions impact the likelihood of a successful double pass—would provide even more nuanced insights.

Finally, this project emphasizes the growing importance of sports analytics not just as a tool for post-match analysis but as a proactive instrument for tactical planning and in-game decision-making. The insights gained can serve as a foundation for developing more sophisticated models help teams maximize their performance in critical moments of the game. While machine learning and data analytics provide powerful tools, their real value lies in how they are applied to practical, real-world scenarios, where human intuition and strategic thinking must go hand-in-hand with data-driven insights.

This report underscores the potential of double passes in football, provides a solid methodology for analyzing them, and highlights the importance of integrating high-quality data with advanced analytical techniques. There is no doubt that the future of football tactics will be increasingly shaped by insights drawn from such analyses, and the key to success will be the ability to balance data-driven decision-making with the intuition and experience that comes from deep knowledge of the game.

References

[1] Dfl deutsche fußball liga gmbh: Official match data. `https://www.dfl.de/en/topics/match-data/official-match-data/`.

[2] Gennady Andrienko, Natalia Andrienko, Guido Budziak, Jason Dykes, Georg Fuchs, Tatiana Von Landesberger, and Hendrik Weber. Visual analysis of pressure in football. *Data Mining and Knowledge Discovery*, 31:1793–1839, 2017.

[3] J. Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365, 1944.

[4] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[5] Schumacher A Peters B. Zwei gegen eins: Starke entscheider auf dem platz. 2022.

[6] Grund Thomas U. Network structure and team performance: The case of english premier league soccer teams. *Social Networks*, 34(4):682–690, 2012.

[7] Stacey Winham, Robert Freimuth, and Joanna Biernacka. A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining*, 6, 12 2013.

[8] Yuji Yamamoto and Keiko Yokoyama. Common and unique network dynamics in football games. *PloS one*, 6:e29638, 12 2011.