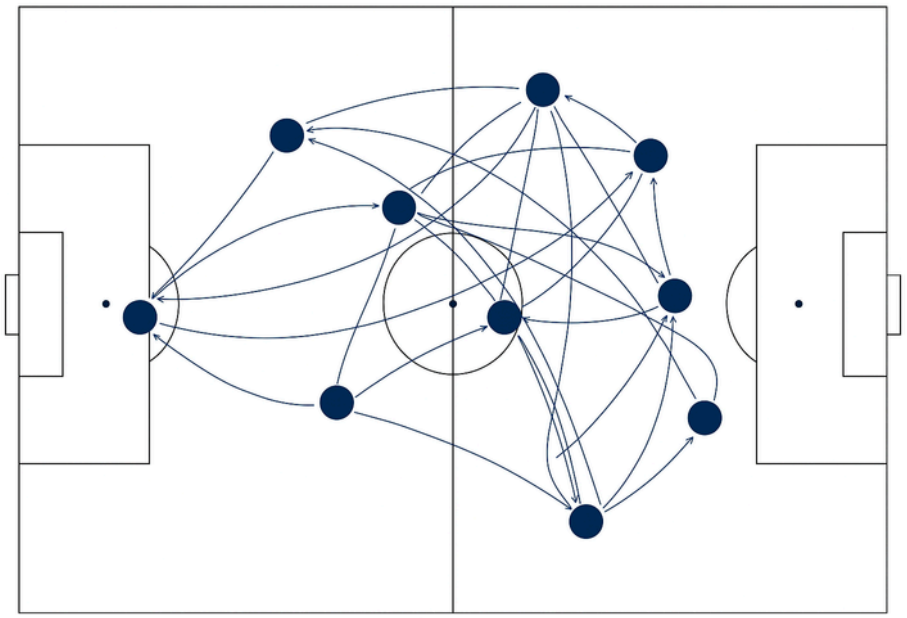
FOOTBALL PASS NETWORKS: STRUCTURE, ANALYSIS AND PREDICTION



GABRIEL CARBINATTO

INTRODUCTION

The objective of this project is to develop a predictive model for football match outcomes using metrics extracted from pass networks between players. The methodology is based on the construction of directed graphs from StatsBomb event data, where nodes represent players and edges represent passes made during matches.
For each team in each match, 28 network topology metrics are computed to capture different aspects of tactical organization: player centrality, connection density, clustering structures, and spectral properties. These metrics are used as features in machine learning algorithms to predict which team will win the match.
The final dataset comprises 2,025 matches across 40 different competitions, totaling 4,050 analyzed pass networks. The final model (an SVM with RBF kernel) achieved an accuracy of 75.33% on the test set, demonstrating that network topology characteristics can effectively predict sports outcomes.
Subsequently, a SHAP value analysis was performed to identify which patterns and features are associated with winning pass networks, revealing the most decisive factors for the tactical success of teams.

DATA COLLECTION

The data used in this project was obtained from StatsBomb, one of the leading providers of detailed professional football data. StatsBomb offers granular event data that captures every action during matches, including passes, shots, ball recoveries, and other events.
The dataset covers a total of 40 combinations of competitions and seasons, representing major football tournaments worldwide:

Domestic Leagues

- **Bundesliga:** 2015/16, 2023/24.
- **La Liga:** 2010/11, 2011/12, 2012/13, 2013/14, 2014/15, 2015/16, 2016/17, 2017/18, 2018/19, 2019/20, 2020/21.
- **Premier League:** 2015/16.
- **Serie A:** 2015/16.
- **Ligue 1:** 2015/16, 2021/22, 2022/23.
- **MLS:** 2023.
- **Indian Super League:** 2021/22.

International Competitions

- **UEFA Champions League:** 2003/04, 2004/05, 2006/07, 2008/09, 2009/10, 2010/11, 2011/12, 2012/13, 2013/14, 2014/15, 2015/16, 2016/17, 2017/18, 2018/19.
- **UEFA Euro:** 2020, 2024.
- **Copa América:** 2024.
- **FIFA World Cup:** 2018, 2022.
- **African Cup of Nations:** 2023.

In total, 2,707 matches were collected, containing 7,254,305 events, resulting in a robust and diverse dataset.

CONSTRUCTION OF PASS NETWORKS

Network Structure

The networks are constructed as directed and weighted graphs, where:

- **Nodes: Represent the players of the team.**
- **Edges: Represent passes between players (directed).**
- **Weights: Relative frequency of passes between each pair of players.**

For each match, two independent networks are created (one for each team), resulting in 4,050 networks in the final dataset.

Network Metrics Calculated

For each network, a set of metrics is computed to capture different aspects of its topology:

Basic Metrics

- **Edge Count: Total number of connections.**
- **Network Density: Proportion of realized versus possible connections.**
- **Average Degree: Average connectivity of the players.**

Centrality Metrics

- **Betweenness Centrality: Identifies players who act as "bridges" within the network.**
- **PageRank: Measures the importance of players within the network.**
- **Eigenvector Centrality: Captures influence based on connections to important nodes.**
- **Katz Centrality: A variant of eigenvector centrality that includes a damping parameter.**

Clustering Metrics

- **Clustering Coefficient: Tendency of triangle formation within the network.**
- **Transitivity: Global measure of clustering across the network.**
- **Triangle Count: Number of triangles present in the network.**

Advanced Metrics

- **Reciprocity: Proportion of bidirectional connections.**
- **Assortativity: Tendency of nodes to connect with similar nodes.**
- **Modularity: Community detection within the network.**
- **Weight Entropy: Distribution of pass intensity across edges.**
- **Spectral Metrics: Spectral radius and Fiedler value.**

NETWORK VISUALIZATIONS

Pass Network

Pass networks represent the fundamental structure of ball exchanges between players, with each athlete positioned at their average on-field coordinates and the edges indicating the connections between them.

BORUSSIA DORTMUND (W)

Borussia Dortmund vs FSV Mainz 05

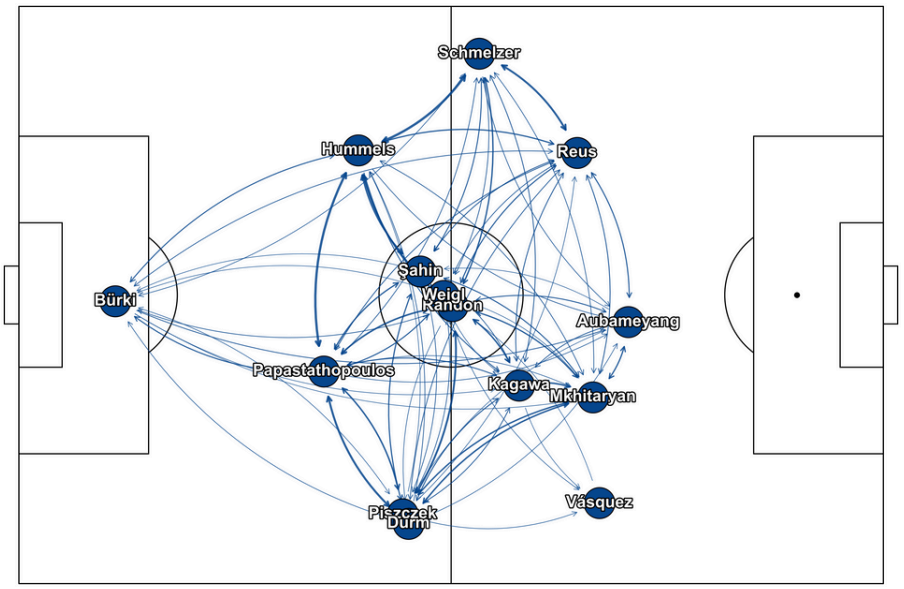


Figure 1: Example of a Pass Network – Winning Team.

WEST BROMWICH ALBION (L)

West Bromwich Albion vs Chelsea

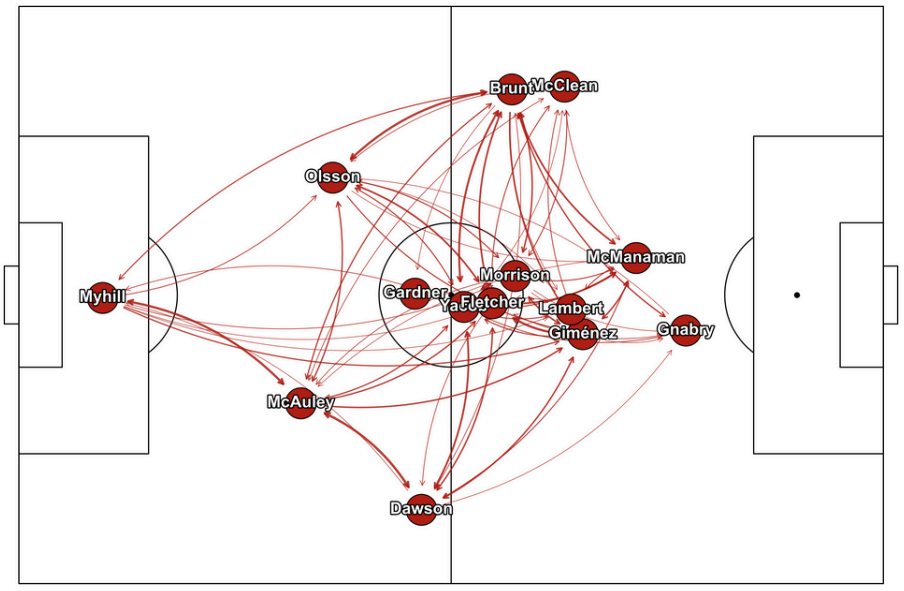


Figure 2: Example of a Pass Network – Losing Team.

NETWORK VISUALIZATIONS

Density Map

The density maps are constructed using a sophisticated methodology that combines the pass network with spatial analysis:

- **Point Sampling: For each player, multiple points are generated at their average on-field positions.**
- **Pass-Based Interpolation: Additional points are interpolated along the pass lines (edges), with the number of points proportional to the connection weight.**
- **Density Estimation: Kernel Density Estimation (KDE) is applied with a bandwidth of 0.3 to smooth the distribution.**
- **Visualization: Colored contours indicate zones of higher passing activity.**

This approach enables the identification of zones with higher concentrations of activity on the pitch, revealing spatial patterns that are not apparent in traditional pass network visualizations.

REAL MADRID (W)

Real Madrid vs Sporting Gijón

PASSING NETWORK **DENSITY MAP**

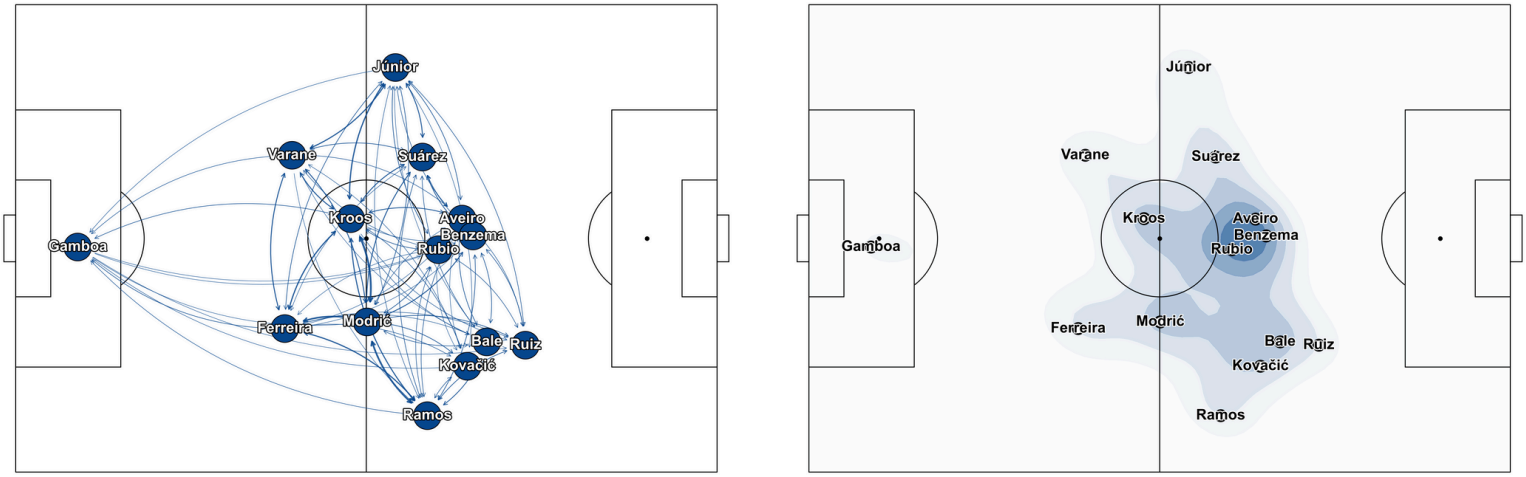


Figure 3: Example of a Density Map – Winning Team.

RENNES (L)

Bastia vs Rennes

PASSING NETWORK **DENSITY MAP**

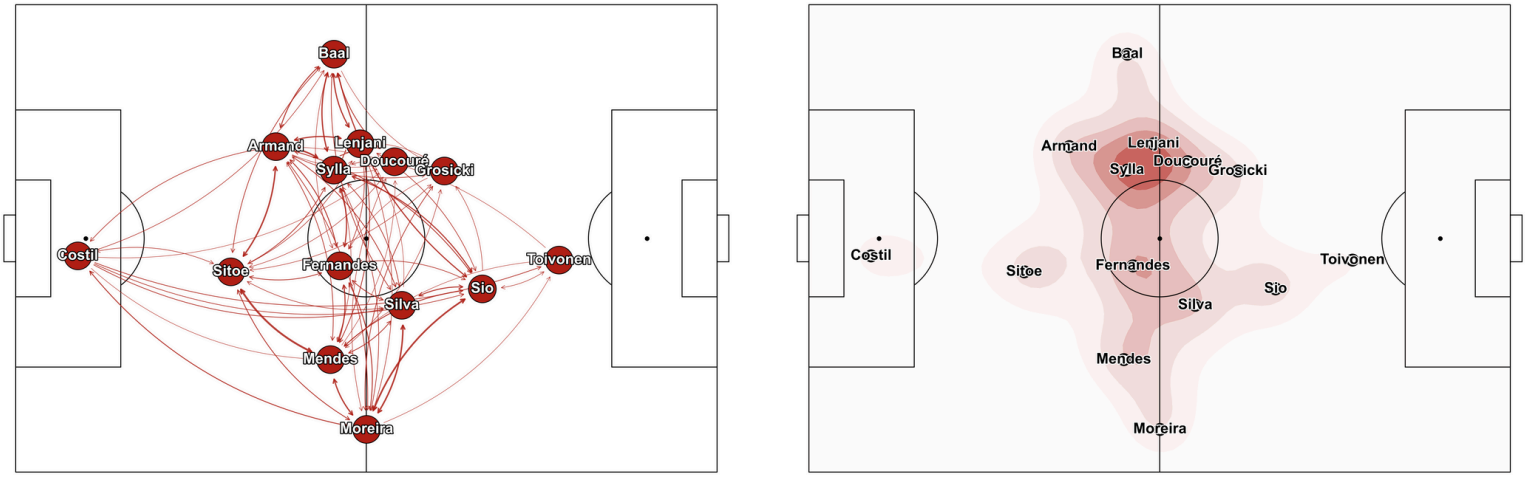


Figure 4: Example of a Density Map – Losing Team.

NETWORK VISUALIZATIONS

Key Players and Main Connections

The identification of key players is based on the weighted degree of the nodes, calculated using the following methodology:

- **Degree Calculation: Sum of the weights of all edges connected to the player (both incoming and outgoing).**
- **Normalization: Values are normalized between 0 and 1, considering all players on the team.**
- **Selection Threshold: Players with a normalized degree greater than 0.7 are classified as key players.**
- **Visualization: Highlighted with yellow borders.**

Main connections are determined by selecting each player's two strongest connections, avoiding duplicates to maintain visual clarity. This filtering allows the focus to remain on the most important relationships without overloading the visualization.

BAYERN MUNICH (W)

Bayern Munich vs Schalke 04

PASSING NETWORK | **KEY PLAYERS & MAIN CONNECTIONS**

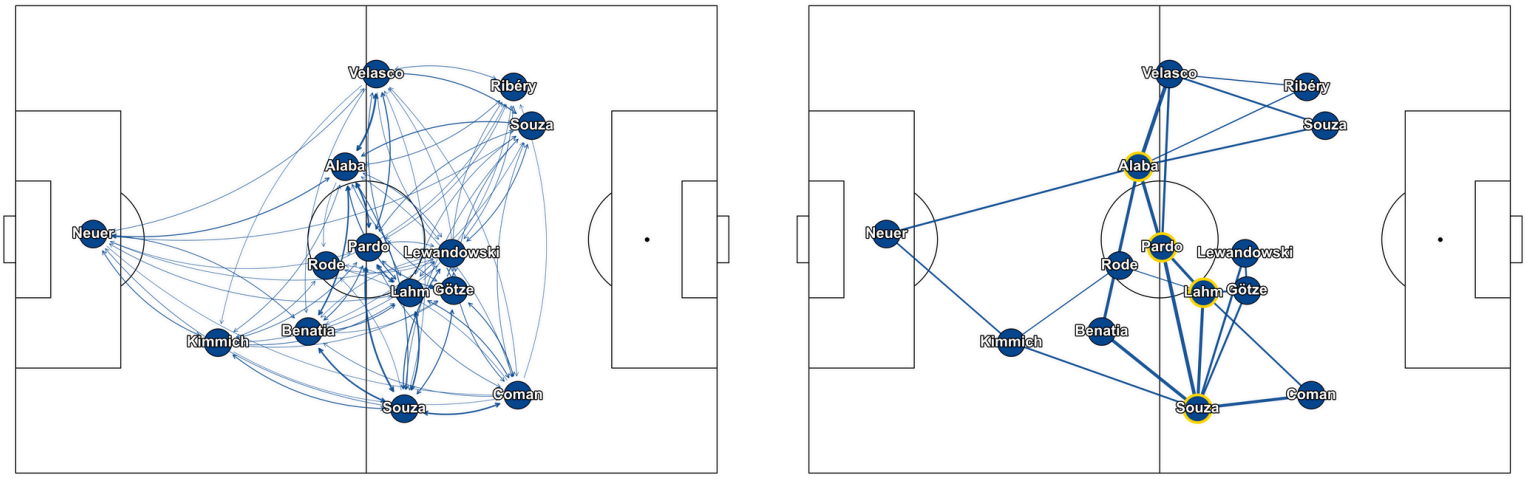


Figure 5: Example of a Key Players and Main Connections Visualization – Winning Team.

SERBIA (L)

Brazil vs Serbia

PASSING NETWORK | **KEY PLAYERS & MAIN CONNECTIONS**

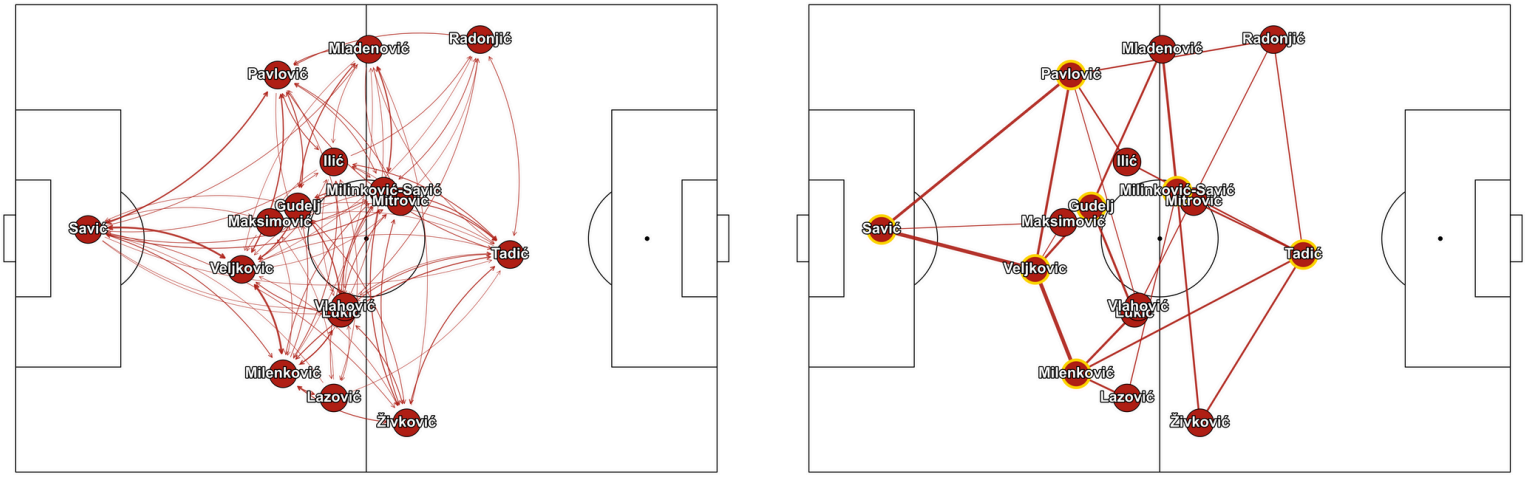


Figure 6: Example of a Key Players and Main Connections Visualization – Losing Team.

NETWORK VISUALIZATIONS

PLAYMAKERS AND ZONE CONNECTIONS

Playmakers are identified through an analysis based on betweenness centrality:

- **Calculation: Weighted betweenness centrality applied to the graph converted to an undirected form.**
- **Normalization: Values normalized between 0 and 1 within each team.**
- **Identification threshold: Players with normalized centrality greater than 0.7.**
- **Visualization: Highlighted with orange borders.**

Zone connections implement a tactical division of the pitch into three main sectors:

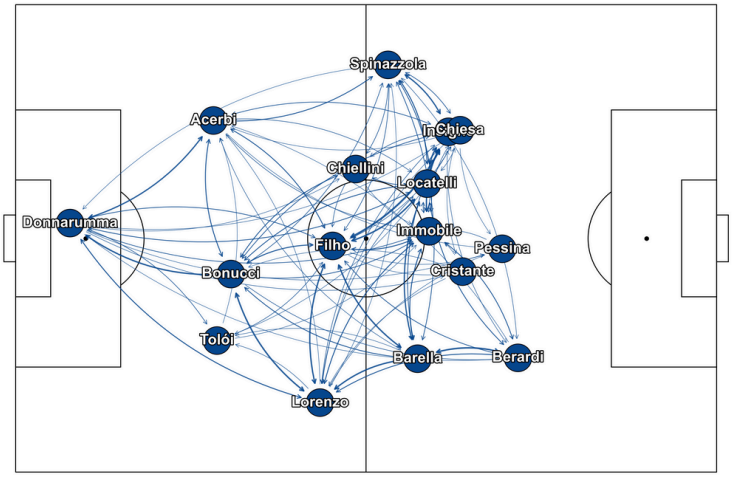
- **Defensive Zone: x = [0; 40] (first third).**
- **Midfield Zone: x = [40; 80] (middle third).**
- **Attacking Zone: x = [80; 120] (final third).**

In the visualization, connections between different zones are prioritized, highlighting the team's ability to progress play and transition between sectors.

ITALY (W)

Italy vs Switzerland

PASSING NETWORK



PLAYMAKERS AND ZONE CONNECTIONS

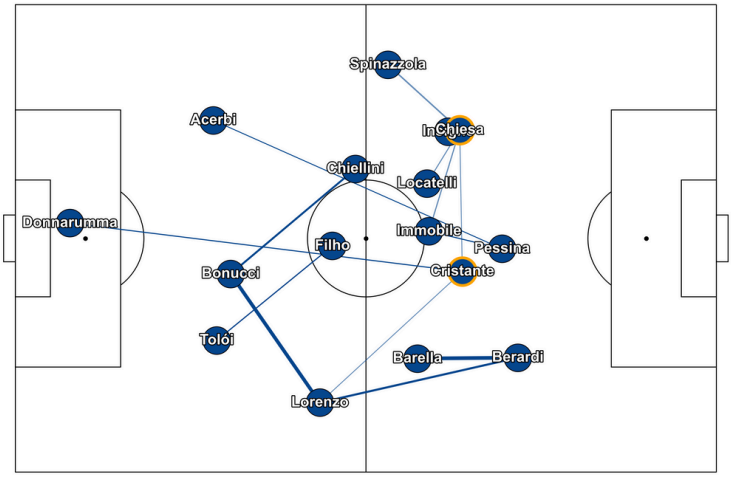
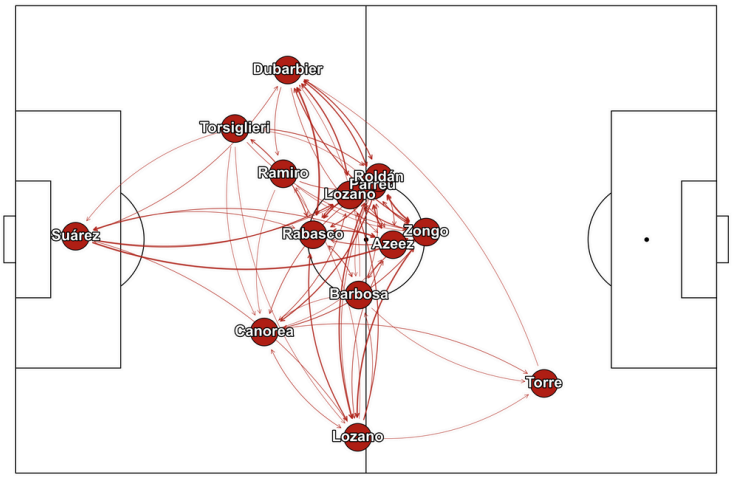


Figure 7: Example of a Playmakers and Zone Connections Visualization – Winning Team.

ALMERÍA (L)

Barcelona vs Almería

PASSING NETWORK



PLAYMAKERS AND ZONE CONNECTIONS

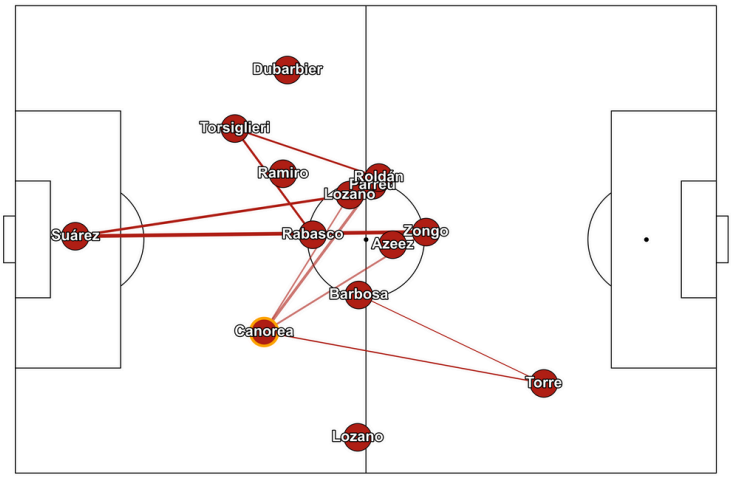


Figure 8: Example of a Playmakers and Zone Connections Visualization – Losing Team.

PREDICTIVE MODELING

Problem Formulation

The predictive modeling relies exclusively on network property metrics derived from pass networks to determine, for each match, which team won and which team lost. The problem is formulated as a binary classification task (0-1), where:

- **Input: Differences between the 28 network metrics of the two opposing teams.**
- **Output: Binary classification.**

For each match, two records are generated: one representing the winning team (target = 1) and the other representing the losing team (target = 0), both using the differences between their respective network metrics as features.

Validation Strategy

The validation process followed rigorous practices to ensure reliable performance estimates:

- **Hold-out Test Set: 15% of the matches (304 matches, 608 records) were set aside before any training.**
- **GroupKFold Cross-Validation: Prevents data leakage between records from the same match.**
- **Nested Cross-Validation: Unbiased hyperparameter tuning using 5 outer folds and 3 inner folds.**
- **Evaluation Metric: G-Mean (geometric mean of sensitivity and specificity), suitable for balanced problems.**

Model Results

Four machine learning algorithms were tested, all using only the differences between network metrics as input features:

MODEL	Accuracy (CV)	Accuracy (Test)
SVM (RBF)	0.7838 ± 0.0130	0.7533
Logistic Regression	0.7780 ± 0.0127	-
XGBoost	0.7618 ± 0.0112	-
Random Forest	0.7571 ± 0.0101	-

Table 1: Results from Nested Cross-Validation and Final Test Evaluation.

The SVM model demonstrated excellent generalization, with a small gap between cross-validation and final test performance, indicating low overfitting.

Interpretability: SHAP Analysis

SHAP Values

To understand which network features have the greatest influence on the prediction, we applied SHAP (SHapley Additive exPlanations) to the Logistic Regression model.

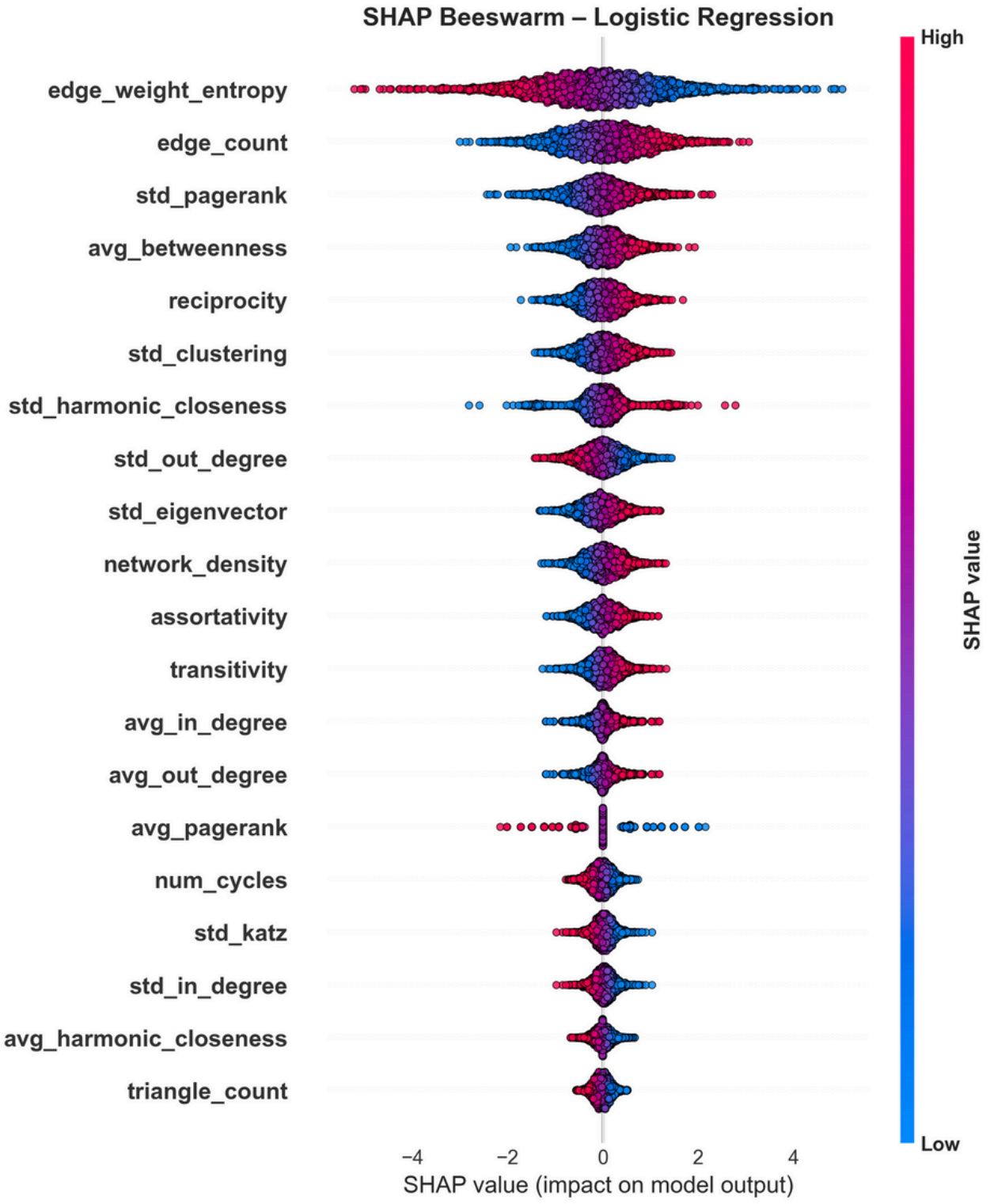


Figure 9: SHAP Beeswarm plot showing the distribution of SHAP values for each feature.

INTERPRETABILITY: SHAP ANALYSIS

Top 5 Most Important Metrics

Edge Weight Entropy (Importance: 0.990)

- **Interpretation:** Distribution of pass intensity.
- **Football Impact:** Teams with a more uniform distribution of passes (low entropy) tend to win.
- **Pattern:** Monotonically decreasing – lower entropy leads to better performance.

Edge Count (Importance: 0.669)

- **Interpretation:** Total number of passing connections between players.
- **Football Impact:** Teams with more diverse connections between players gain an advantage. This reflects tactical variety and multiple passing options.
- **Pattern:** Monotonically increasing – more connections indicate better performance.

Standard Deviation of PageRank (Importance: 0.465)

- **Interpretation:** Variability in the importance of players within the network.
- **Football Impact:** Greater variability suggests a clear hierarchy of player influence, with some players being central to the team's play.
- **Pattern:** Monotonically increasing – a well-defined hierarchy is advantageous.

Average Betweenness Centrality (Importance: 0.370)

- **Interpretation:** Average presence of players acting as "bridges" in the network.
- **Football Impact:** Teams with players that connect different sectors of play have a tactical advantage. These are the connectors who bridge the defensive and attacking phases.
- **Pattern:** Monotonically increasing – more playmakers lead to better performance.

Reciprocity (Importance: 0.339)

- **Interpretation:** Proportion of bidirectional connections between players.
- **Football Impact:** High reciprocity indicates mutual passing between players, reflecting chemistry. Well-connected teams maintain possession through consistent exchanges.
- **Pattern:** Monotonically increasing – higher reciprocity reflects better team cohesion.

CONCLUSIONS

This study demonstrates that network topology metrics derived from passing patterns can predict football match outcomes with significant accuracy (75.33% on the test set). Analysis of the five most important metrics reveals critical tactical insights that distinguish winning teams from losing ones.

Teams achieve better outcomes when passes are distributed more evenly across players. This uniform distribution (low entropy) suggests that tactical systems based on collective participation and team-wide involvement are more effective than those that concentrate high edge weights within the network.

Successful teams establish more diverse passing connections among players, reflecting sophisticated tactical approaches with multiple offensive and defensive options. This connectivity across various players indicates that athletes are able to connect and exchange passes with a larger number of teammates, expanding ball circulation possibilities and making it harder for opponents to defend.

Teams benefit from having clearly defined roles in which certain players take on greater importance within the network structure. This variability in player influence reflects natural differences in positional responsibilities and individual capabilities, where some players naturally become more central to the team's passing structure.

The most effective teams feature players who act as connectors between different areas of the pitch, particularly linking the defensive and attacking phases. These transition players enable fluid exchanges and maintain tactical coherence throughout matches.

High levels of bidirectional passing between players indicate strong cohesion and mutual understanding within the team. This reciprocity reflects practiced combinations and trust among teammates, favoring one-twos and quick passing sequences.