A Graph-Based Approach to Football Analytics: Leveraging StatsBomb Event Data for Advanced Machine Learning Projects

1. Introduction: From Event Streams to Relational Graphs

1.1 The Limitations of Traditional Football Analytics

The quantitative analysis of football has traditionally relied on aggregated performance metrics and linear statistical models. These approaches, while providing valuable summary insights such as total passes, distance covered, or shot counts, often fall short of capturing the sport's intrinsic complexity. Football is a multi-agent system of continuous, non-linear interactions where the effectiveness of a single action is deeply dependent on the dynamic context of all players, the ball, and the pitch. The predictability of match outcomes is notoriously difficult to model due to the myriad of variables, including tactical decisions, player psychology, and unpredictable occurrences, which create a highly volatile environment for quantitative analysis.[1]

Furthermore, traditional methods that treat events as isolated data points fail to account for the crucial interdependencies that define a match. While some studies have successfully used "complex networks" like passing networks to predict match outcomes, a more holistic approach is required to capture the full, dynamic game state.[2] A core challenge in modeling this domain is the multi-agent nature of the game and the fact that models must be invariant to the order of players, as the event itself is the primary focus, not who performs it.[3]

1.2 The Graph-Native Nature of StatsBomb Event Data

The StatsBomb Open Data, though presented as a series of chronological events, possesses an inherent graph structure that makes it uniquely suitable for advanced relational analysis. Each event—whether a pass, a tackle, a shot, or a duel—can be conceptualized as a node in a network. The relationships between these events, such as their temporal sequence, spatial proximity, and causal links, form the edges of a complex, dynamic graph.

The dataset's rich attributes serve as the raw material for this graph representation. Fields like related_events provide explicit links between actions, such as a pass leading to a shot.[4] The

freeze_frame data captures a snapshot of player locations at a critical moment, providing a geometric context that is often missing from event-based data.[4] The

possession and play_pattern attributes naturally segment the game into meaningful sub-graphs, enabling the analysis of distinct offensive or defensive sequences. This structure directly addresses the challenges of handling multi-agent systems and player-order invariance by modeling the relationships between entities rather than just their individual actions.[3]

1.3 The Transformative Potential of Graph-Based Machine Learning

Graph Neural Networks (GNNs) are a class of machine learning models specifically designed to operate on data structured as graphs. Their ability to model interactions within such structured data makes them exceptionally well-suited for analyzing the intricate dynamics of football.[5] GNNs move beyond an analysis of individual actions to understand the collective "game state" as a cohesive entity.

By representing players as nodes and their interactions as edges, GNNs can model context-dependent behaviors that traditional models cannot.[6] For example, the effectiveness of a pass is not only a function of the passer and receiver but also of the positioning of surrounding players who may block passing lanes or create space.[6] GNNs can account for these contextual factors, providing a more nuanced understanding of how player interactions contribute to the success of a formation or strategy.[7] The integration of GNNs with other architectures, such as Recurrent Neural Networks (RNNs) or LSTMs, is a particularly powerful approach, as it allows for the simultaneous modeling of both spatial and temporal dependencies.[3]

2. Conceptual Framework: Modeling StatsBomb Data as a Graph

2.1 Defining the Graph Schema: Nodes, Edges, and Attributes

The proposed projects will employ a heterogeneous graph schema, which allows for multiple types of nodes and edges, thereby providing a more accurate representation of the complexity within the StatsBomb dataset.

- **Nodes:**
  - **Player Nodes:** Represent individual athletes. Node features would include id, name, team_id, and position_id.[4]
  - **Event Nodes:** Represent every action in the dataset. Node features would include the event id, type, location, timestamp, duration, and outcome-specific attributes (e.g., pass.outcome, shot.statsbomb_xg).[4]
  - **Team Nodes:** Represent the two competing teams, with aggregated player or formation data as features.
  - **Possession Nodes:** Represent a unique possession chain, as defined in the dataset.[4]
- **Edges:**
  - **Temporal-Sequence Edges:** Directed edges linking one Event node to the next within the same possession, capturing the flow of play over time.
  - **Player-Action Edges:** Directed edges from a Player node to an Event node, signifying that the player performed the action.
  - **Possession-Event Edges:** Edges linking Event nodes to their parent Possession node.
  - **Related-Event Edges:** A critical, explicitly defined directed edge linking events that are causally or associatively related, such as a Pass to a Shot.[4]
  - **Spatial-Proximity Edges:** Edges between Player nodes based on a distance threshold at a specific timestamp. These edges are dynamic and are particularly valuable when reconstructed from the freeze_frame data.[4]
  - **Under-Pressure Edges:** A special, contextual edge from a Pressure event to the Pass or Carry event it affects, as defined by the StatsBomb documentation.[4]

2.2 Second and Third-Order Insights: The Power of Context and Causality

The StatsBomb dataset contains attributes that, when used to construct a graph, enable a deeper level of analysis than a simple event log. For example, the related_events field is not merely an identifier; it is a direct representation of a causal or associative relationship within the game. A Shot is not simply followed by a Goalkeeper event; it is the direct cause of it. Similarly, a successful Interception is directly related to a failed Pass. This allows for the construction of a semantically meaningful graph that models the flow of causality, which is a far more powerful representation for a predictive model than a simple temporal graph.

The freeze_frame data associated with Shot events transforms a sparse event log into a series of rich, contextual snapshots. A major limitation of event data is the lack of continuous player tracking information. However, the freeze_frame array, documented in Appendix 3, provides a snapshot of all relevant players and their locations at the precise moment of a shot.[4] This allows for the reconstruction of the spatial geometry of the game at critical junctures, enabling a model to learn about defensive formations, passing lanes, and player spacing. This elevates spatio-temporal analysis from a simple sequence of events to a series of rich, graph-based game states.

2.3 Key Table: Mapping StatsBomb Fields to Graph Components

The following table provides a structured blueprint for the initial data parsing and graph construction phases, explicitly mapping raw StatsBomb fields to their designated roles within the proposed graph models.

StatsBomb Field	Graph Component	Description/Role
event.id	Node	Unique identifier for each event node.
player.name, player.id	Node Attribute	A unique identifier and name for a Player node.
type.name, type.id	Node Attribute	Defines the type of Event node (e.g., Pass, Shot).
location, end_location	Node Attribute	Spatial coordinates of the event or its end point.

related_events	Edge	Creates a directed, causal edge between two event nodes.
timestamp	Node Attribute	Used to order event nodes and create temporal-sequence edges.
possession	Node Attribute	Links events to a specific possession chain, creating a sub-graph.
under_pressure	Edge	Creates a contextual edge from a Pressure event to a subsequent event.
freeze_frame	Node Attribute	Provides a snapshot of surrounding players, used to create dynamic spatial-proximity edges between Player nodes.
pass.outcome	Node Attribute	Outcome attribute for a Pass event node.
shot.statsbomb_xg	Node Attribute	Expected goals value attribute for a Shot event node.

3. Project Idea 1: Predictive Offensive Momentum and Threat Assessment

3.1 Problem Statement: Quantifying Offensive Flow Beyond xG

While Expected Goals (xG) is a powerful metric for evaluating the quality of a shot, it provides a static assessment of a single action. It fails to capture the dynamic build-up of play that precedes a shot. This project aims to predict the probability of a possession chain resulting in a goal or a high-quality scoring opportunity by modeling the entire flow of offensive play, moving beyond the traditional shot-centric view.

3.2 Graph Construction for Offensive Sequences

For each unique possession defined in the dataset, a distinct sub-graph is created. The nodes would represent offensive Players and the Events they perform, such as Pass, Carry, Dribble, and Shot. Defensive Player nodes can also be included from freeze_frame data to provide contextual information.

The graph would be connected by Temporal-Sequence edges based on the index field, Player-Action edges linking players to their events, and Related-Event edges, such as from a Pass to a Shot.[4] Node features would include

location, type.name, pass.length, carry.duration, and the under_pressure boolean.[4]

3.3 Proposed GNN Architecture and Workflow

A hybrid GNN-RNN model is proposed for this task. First, a GNN, such as a Graph Attention Network (GAT) or GraphSAGE, would process the graph at each time step. The GNN's role is to learn a rich vector representation (an embedding) for each Event and Player node, incorporating information about their immediate neighborhood, such as nearby players or preceding actions.

Second, the sequence of these node embeddings, ordered by timestamp, would be fed into a Recurrent Neural Network (RNN) like an LSTM or GRU. This RNN would then model the temporal evolution of the possession. The model would be trained to predict the final outcome of the possession, whether it results in a goal, shot, or turnover. The output would be a continuous score representing offensive threat, providing a dynamic metric for the value of each action within the possession. The model's performance could be evaluated using standard classification metrics and a comparative analysis against a baseline xG model to demonstrate the added value of the graph-based approach.

A graph-based approach can model offensive threat as a dynamic, accumulating value rather than a static metric. A traditional xG model provides a score based on the context at a single moment. A GNN-based model, by modeling the entire possession chain of passes, carries, and dribbles, can learn that a sequence of seemingly low-value actions can lead to a high-value outcome. This allows a model to recognize specific patterns of play, such as a rapid counter-attack versus a slow build-up, and assign a real-time "threat score" to each pass or carry. This moves the analysis from "what happened" to "what is likely to happen next," enabling more proactive tactical decision-making.

4. Project Idea 2: Quantifying and Predicting Defensive Pressing Success

4.1 Problem Statement: Analyzing the Effectiveness of the "Press"

Defensive pressure is a crucial tactical concept, but its effectiveness is difficult to quantify. A Pressure event, as a data point, does not inherently guarantee a successful outcome. This project aims to model the effectiveness of defensive pressing actions and predict the likelihood of a turnover, foul, or other loss of possession resulting from a specific defensive sequence.

4.2 Graph Construction for Defensive Transitions

The analysis would focus on creating sub-graphs centered around Pressure or Counterpress events. The nodes would include the pressing Player (defender), the pressured Player (attacker), and all surrounding Players sourced from freeze_frame data.[4] Events related to the defensive action, such as a

Pass, Carry, Dribbled Past, or Foul Committed, would also be included as nodes.

Edges would consist of Spatial-Proximity edges between players and a special Under-Pressure edge linking the defender's Pressure event to the attacker's corresponding Pass or Carry event.[4] Node features would include

player.position, team.id, location, and the under_pressure and counterpress booleans.[4] Edge features could include the distance between players.

4.3 Proposed GNN Architecture and Workflow

A GNN model like GraphSAGE is well-suited for this node classification task. The GNN's role is to learn embeddings for each player within the defensive sub-graph, incorporating their spatial relationships and the nature of their interactions. The model would be trained to predict the outcome of the defensive action (e.g., a binary classification of successful turnover vs. unsuccessful pressure). The model would learn which player configurations—such as a defender in a specific location relative to the attacker and their teammates—are most likely to lead to a successful outcome. The performance of the model would be evaluated using metrics like accuracy, precision, and recall.

A model trained on counterpress events can learn a team's tactical "defensive identity." The counterpress attribute specifically tags pressing actions that occur within 5 seconds of an open-play turnover.[4] By focusing on these events, a GNN can learn to distinguish teams that execute a high-intensity, immediate press from those that do not. The model's learned embeddings for teams or players could then serve as a quantitative measure of their pressing effectiveness, providing a data-driven complement to qualitative tactical analysis. This moves the analysis from "did they press?" to "how effective is their press, and under what conditions?"

5. Project Idea 3: Player Synergy and Tactical Role Discovery

5.1 Problem Statement: Uncovering Hidden Player Relationships and Roles

Team success in football is not solely a function of individual talent but of how players work together to execute a strategy.[6] This project uses GNNs to discover and quantify player partnerships and nuanced tactical roles that go beyond simple pass counts or goal contributions.

5.2 Graph Construction for Player Interaction Networks

This project uses a more static, longer-term graph. A single graph can be built for a team over an entire match or even a season. The nodes of the graph would represent the Players on a given team. An edge would be created between two Player nodes for every Pass event between them, with the edge being directed from the passer to the recipient.[4]

The edges would be rich with features, including pass.length, pass.angle, under_pressure, pass.outcome, and whether the pass was a cross, cut-back, shot-assist, or goal-assist.[4] Aggregating these features over time would provide a comprehensive representation of the nature of the partnership between two players.

5.3 Proposed GNN Architecture and Workflow

A GNN model, such as a Graph Convolutional Network (GCN), would be trained as an unsupervised learning task. The GCN would learn a low-dimensional embedding for each Player node. The goal is to learn a representation where players who perform similar roles or have strong synergies are closer to each other in the embedding space.

The workflow would involve aggregating all Pass events for a chosen team over a period to build the player interaction graph. The GCN would then be trained to learn the player embeddings. Post-training, the embeddings can be clustered to identify tactical roles, such as a "deep-lying playmaker" or a "target man," based on the network of passes. The distance between player embeddings can also be used as a quantitative measure of their synergy. This information can be used for lineup selection, scouting, and opponent analysis.[6]

GNN-based player embeddings can identify players who contribute significantly to offensive success without being directly involved in goals or assists. While traditional analysis might focus on players with high goal or assist numbers, a GNN, by processing the entire passing network with contextual features, can learn that a player who consistently makes short, low-risk passes to a teammate who then makes a high-value pass has a crucial, yet hidden, tactical role. Their embedding would reflect this indirect contribution, allowing coaches to identify undervalued players or strategic roles that are not obvious from a box score. This moves the analysis from simple metrics to a holistic understanding of player contributions.

6. Project Idea 4: Spatio-Temporal Event Detection (Advanced Research)

6.1 Problem Statement: Predicting Unlabeled Events from Game State

The most ambitious and valuable application of machine learning in sports is the ability to predict what *will* happen next based on the current game state. This project, which directly links to existing research on modeling 2D trajectory data, aims to predict the occurrence of specific, high-value events such as a Foul Committed, Miscontrol, or Interception by modeling the dynamic spatio-temporal relationships of the game.[3]

6.2 Graph Construction with freeze_frame Snapshots

This project requires a highly dynamic graph that evolves with each timestamp. The nodes of this graph would represent all Players on the pitch, the Ball, and the Pitch itself (represented by grid points). The freeze_frame data provides the necessary snapshots to create the core structure of this graph.[4]

Edges would be created to represent Spatial-Proximity between all players, a Ball-Player edge for the player in possession, and Temporal edges between a player's nodes at consecutive timestamps. Node features would include player.id, team.id, position.id, and location. Edge attributes could include distance and relative velocity, derived from player trajectories over time.

6.3 Proposed GNN Architecture and Workflow

A hybrid GNN-RNN model that processes a sequence of freeze_frame-like snapshots is the appropriate architecture. The GNN layer would process the graph at each timestamp to create a rich, context-aware embedding for the entire game state. The sequence of these game state embeddings would then be fed into an RNN or LSTM layer to model the temporal flow of the match. The model would be trained on sequences of game states to predict the next event. For example, it could learn the spatial and temporal patterns that precede a Foul Committed by a specific player.

The workflow would involve first creating a high-frequency, interpolated dataset of player locations, potentially using linear interpolation between known location and end_location points to approximate a "trajectory." The series of spatio-temporal graphs would then be constructed from this data. The model would be trained on a sparse set of labeled events (e.g., Foul Committed, Interception). The model's predictive performance would be evaluated on unseen data.

This project addresses a fundamental challenge in sports machine learning: predicting events from non-event data. The StatsBomb dataset, while event-based, provides sufficient data points and rich attributes—particularly the location and freeze_frame fields—to create a proxy for continuous tracking data. By using these snapshots to train a GNN-RNN, a system can be built that learns to recognize the spatial and temporal patterns that precede key events. This moves the analysis from a post-game summary to a potential real-time predictive system for in-game decision support, representing a new frontier in data-driven sports strategy.

7. Conclusion: The New Frontier of Football Analytics

7.1 Summary of GNNs in Football Analytics

The projects outlined in this report demonstrate how a graph-based approach can fundamentally transform football analytics. By modeling the intricate web of relationships within the StatsBomb dataset, GNNs enable a move from simple statistical analysis to a deeper, context-aware understanding of the game. GNNs are uniquely positioned to model player relationships, tactical formations, and the dynamic flow of play in a way that traditional methods, which struggle with the inherent complexity of the sport, cannot.[1] Each project—from quantifying offensive threat to discovering player synergy—leverages the graph-native structure of the data to provide more nuanced and actionable insights.

7.2 Challenges and Future Work

Implementing these projects is not without its challenges. The graphs for a full match or an entire season could be massive, presenting significant computational hurdles. Furthermore, while the StatsBomb data is rich, it is still an event-based dataset and not a continuous tracking dataset, which introduces a level of data sparsity that must be addressed, for example, through interpolation. A long-term research direction would involve combining this event data with high-frequency player tracking data to further enhance the models' predictive capabilities.[5]

7.3 The Future of Data-Driven Football

These projects lay the foundation for a new paradigm in football analytics. By moving beyond traditional metrics, a graph-based approach can revolutionize tactical decision-making, player scouting, and fan engagement. GNNs provide a means to gain a deeper understanding of the complex relationships and interactions that define successful formations, allowing teams to make more informed and precise tactical decisions.[7] The ability to simulate potential game scenarios and identify hidden contributions marks a significant evolution in how teams can leverage data to optimize performance and gain a competitive edge.

Works cited

1. Predicting Soccer Match Outcomes:A Data-Driven Approach | by Fion Ouyang | Medium, accessed on September 7, 2025, https://medium.com/@fion.ouyang/predicting-soccer-match-outcomes-a-data-driven-approach-df6bea2429ec
2. Predicting soccer matches with complex networks and machine learning - arXiv, accessed on September 7, 2025, https://arxiv.org/html/2409.13098v1
3. Graph Neural Networks for Events Detection in Football - DiVA, accessed on September 7, 2025, https://kth.diva-portal.org/smash/get/diva2:1845172/FULLTEXT01.pdf
4. Open Data Events v4.0.0.pdf
5. Game State and Spatio-temporal Action Detection in Soccer using Graph Neural Networks and 3D Convolutional Networks - arXiv, accessed on September 7, 2025, https://arxiv.org/html/2502.15462v1
6. (PDF) Smart Football Formations: The Power of Graph Neural Networks in Recommendation Systems - ResearchGate, accessed on September 7, 2025, https://www.researchgate.net/publication/383025412_Smart_Football_Formations_The_Power_of_Graph_Neural_Networks_in_Recommendation_Systems
7. (PDF) Graph Neural Networks for Personalized Football Formation Strategies in Sports Analytics - ResearchGate, accessed on September 7, 2025, https://www.researchgate.net/publication/383025658_Graph_Neural_Networks_for_Personalized_Football_Formation_Strategies_in_Sports_Analytics