GoalNet: Advancing Counterattack Prediction in Soccer through Gender-Specific Graph Neural Networks

Abstract

Traditional soccer analysis tools emphasize metrics such as chances created and expected goals, leading to an over-representation of attacking players' contributions and overlooking the pivotal roles of players who facilitate ball control and link attacks. Identifying these players could help coaches develop specific tactics and club recruiting. Examples include Rodri from Manchester City and Palhinha who just transferred to Bayern Munich. To address this bias, we developed a model utilizing graph neural networks (GNN) to analyze match events comprehensively. Our research aims to identify players with pivotal roles in a soccer team using GNNs, incorporating both spatial and temporal features.

In our approach, each event in a soccer match is represented as a graph where nodes correspond to players and edges denote interactions. Each node encompasses various attributes, including the player's name, and historical performance metrics such as average pass completion rate. Edges capture interactions between players, such as passes and tackles, with features including pass frequency and distance.We incorporate the last k events to maintain temporal context, accounting for recent interactions. Our model is trained to predict the expected threat (xT) changes for each event, effectively attributing these changes to the contributing players based on their interactions in the previous events. We combine metrics such as degree centrality with the output of the trained GNN model to assign xT changes as credits to players more accurately. To validate the effectiveness of this method, we examined player evaluation outputs, demonstrating that this innovative evaluation method accurately reflects player contributions.

Our findings highlight the significance of these pivotal players in the team dynamics, providing a more nuanced understanding of their impact on the game. This comprehensive analysis using GNNs allows for a balanced evaluation of player contributions, showcasing the indispensable roles of facilitators and initiators in soccer matches.

1. Introduction

Football player performance analysis has traditionally been conducted using prevalent metrics, such as goals, assists, chances created, and expected goals (xG). These are all good metrics for understanding the attacking contributions of players. Still, they often emphasize the contributions of forwards and playmakers while sometimes underestimating the extremely central roles of players who typically work deeper in the field. It is at this point where players like defensive midfielders and central defenders bear the brunt of responsibility in ball retention, transitions, and link-up play that is a sine qua non of how a team performs. Despite their importance, the contributions of such players are rarely reflected in conventional performance evaluations; thus, a gap in our understanding of their impact is created.

Take, for instance, Rodri, the holding midfielder at Manchester City, who facilitates the retention of possession very well and initiates quick transitions; he plays a pivotal role in linking defense with attack. Other midfielders include João Palhinha, who recently joined Bayern Munich from Fulham in a breakthrough deal and shines in the art of breaking opponent's attacks and starting forward play. These players thus contribute a lot to the team dynamics. Still, their contributions are hardly appreciated through traditional metrics, which mainly focus on the number of goal-scoring or assisting actions. The need to devise more comprehensive methodologies for such players' influence is growing.

Using this gap in mind, we find that recent advances in machine learning and graph-based methods give promising results in holistic evaluations of player performance. Graph neural networks are seen as a particularly powerful framework for analyzing soccer matches since they are systems of player interactions. If the model characterizes each player as a node in a system and interactions, such as passes, tackles, and movements, among other events, as edges, GNNs can capture and model the complexity in a game. Furthermore, since GNNs incorporate spatial and temporal features, they might be able to make a more fine-grained sense of how individual players participate in the flow and eventual outcome of the game.

This paper proposes a novel methodology that identifies

pivotal players within a soccer team based on analysis of match events using GNNs. Therefore, we look into a framework where players are individually credited for changes in expected threats (xT) to reflect further players responsible for driving differences within a game from actions that do not directly lead to scoring goals. The approach followed here allows one to interpret, in a slightly more granular level, players' contributions in the two-dimensional and time dimensions of player interactions, especially for those with a role focused on phases of play within ball retention, defense, and linking.

In this study, we model each soccer match event as graphs. In the graphs, nodes represent players and edges represent certain relations between players: either a pass, a tackle, or a movement. Every node also has attributes specific to that player, such as historical performance measures, positional data, and individual passing statistics. The edges capture the interaction between these players, for example, features like pass frequency, distance covered, and event outcomes. Moreover, the GNN model presented considers data from the past k events to ensure a temporal context, such that recent actions and player behaviors are considered in the evaluation.

We further include centrality metrics into the GNN framework for a more precise player evaluation model. Centrality measures, such as degree and betweenness centrality, allow us to quantify how important a particular player is within a network of interactions: those players who are constantly important facilitators. By combining these centrality scores with the outputs of the trained GNN, we can more precisely attribute changes in xT to players, giving a fairer representation of the contributions in each game.

The key contributions of this paper are threefold:

1. We propose a novel GNN-based framework for soccer player evaluation that incorporates both spatial and temporal features of match events.

2. We introduce a methodology for integrating centrality metrics with GNN outputs to better attribute changes in xT to pivotal players.

3. We validate the effectiveness of our approach by analyzing player evaluation outputs, demonstrating its ability to provide a more balanced and comprehensive assessment of player contributions, particularly for non-attacking roles.

2. Literature Review

Recent work on machine learning, in general, and Graph Neural Networks (GNNs), has brought a vast improvement in soccer analytics. Hence, it becomes possible for researchers to analyze complex interactions between players

and derive a greater understanding of the player's contributions than was possible using traditional metrics like goals or assists.

2.1. GNNs for Player and Team Evaluation

In the study by Liu et al. [3], GNNs were used to model and evaluate player actions in soccer. The authors introduced the concept of Value of Actions by Estimating Possession (VAEP), a framework that assigns values to every player action based on its likelihood of improving or diminishing a team's chances of scoring or conceding. This approach allowed for more comprehensive evaluations of both attacking and defensive actions, particularly those that go unnoticed by traditional statistics such as goals and assists [3].

The Playerank model by Pappalardo et al. [4] employs GNNs to evaluate soccer players by taking into account not only individual actions but also the broader tactical context of the match. By representing players as nodes and their interactions (such as passes) as edges in a graph, the model identifies influential players who contribute indirectly to their team's success, often through build-up play or defensive work [4].

2.2. Advanced GNN Techniques for Soccer Analytics

Advanced techniques such as Graph Attention Networks (GATs) have been applied to soccer analytics, allowing for a more granular analysis of player interactions. Velickovic et al. [6] demonstrated that using attention mechanisms within GNNs enables the model to assign varying importance to different player interactions, reflecting their influence on the game. This is particularly useful in identifying playmakers who may not score or assist directly but are critical in creating opportunities [6].

Moreover, Anzer et al. [1] demonstrated the potential of GNNs in detecting tactical patterns in soccer. Their semi-supervised model successfully identified frequent subgraph patterns representing tactical plays such as pressing formations or counterattacks. The ability to model team-level interactions and detect tactical setups gives GNNs a unique edge over traditional sports analytics models [1].

2.3. Valuing Player Actions Beyond Goals

A key development in soccer analytics has been to extend the valuation of player actions beyond goals and assists. In work by Decroos et al. [2], the Actions best capture this Speak Louder than Goals model, which assigns value to every player action by predicting the expected increase or decrease in the probability of scoring or conceding as a result of that action. This methodology allows for a better understanding of the contributions of players who are good at the build-up of play or defensively oriented[2].

In summary, GNNs in soccer analytics strongly advance the capability for understanding player contributions. With the spatial and temporal dynamics embedded, GNNs offer an all-inclusive view of how players interact and influence, which clearly gives more precise evaluations of player performances for all tactical roles.

3. Detailed Dataset Description

This study utilizes primarily two datasets: event data and player-specific data. The event data is derived from Statsbomb open data, specifically from the Premier League 2015/2016 season, which contains detailed event-level information such as body part, event type, and event outcome. Player-specific data is collected from Sofascore and includes features such as accurate pass percentage, key passes, and goal conversion percentage.

3.1. Data Preprocessing

We transform the event data into SPDAL[2] format, a structured format developed to represent soccer matches as a sequence of on-the-ball actions. Each action in a game (pass, shot, dribble) is represented by 12 attributes, including the player's ID, team, action type, start and end positions, possible receiver, and the result. Player-specific data is processed every 90 minutes and normalized to all features.

3.2. Response Variable Definition

This study uses the change Expected Threat(xT)[5] as the response variable since it is designed to evaluate better how players and teams contribute to creating goal-scoring opportunities. Specifically, we have two ways of computing the xT changes.

$$\Delta xT = \begin{cases} xT_{\text{current}} - xT_{\text{previous}}, & \text{same team,} \\ xT_{\text{current}} + xT_{\text{previous}}, & \text{different teams.} \end{cases}$$

If two events are from the same team, we want to know how much more threat the current event adds; if they are from different teams, we need to account for the current event erasing the opponent's xT and contributing to their own.

4. Methodology

4.1. Graph Construction

This section elaborates on how graph representations for soccer match events are constructed. The fundamental basis of our model lies in the graph's structure, which amalgamates attributes of individual players and interactions in the context of an event. We construct the graph dynamically for each event from the available information so that spatial and temporal aspects within the match are well captured. Every event is presented as a graph $G = (V, E)$ in which V

denotes the set of players (nodes) and E is the set of interactions (edges) between players. In this section, we describe how we constructed node and edge attributes for the player and event of each available data.

4.1.1 Node Attributes

Every node of the graph represents one player who participated in the current or past k events. The attribute set of a node $v \in V$ is taken from the player-specific information in the dataset. More particularly, every node possesses the following attributes:

- **Player Name:** The player's name is a unique identifier for referencing event-specific and historical player data.

- **Player Statistics:** For each player, we include the following performance metrics, which are drawn from the dataset:

    - *Goals*: The total number of goals scored by the player.
    - *Successful Dribbles*: The number of successful dribbles the player completes.
    - *Tackles*: The number of successful player tackles.
    - *Accurate Passes Percentage*: The percentage of passes completed successfully by the player.
    - *Rating*: The overall match rating assigned to the player.
    - *Goal Conversion Percentage*: The player's efficiency in converting shots into goals.
    - *Interceptions*: The number of successful interceptions by the player.
    - *Clearances*: The number of defensive clearances the player makes.
    - *Accurate Passes*: The total number of accurate passes the player makes.
    - *Key Passes*: The number of passes leading directly to a goal-scoring opportunity.

Each node is thus characterized by a feature vector $\mathbf{x}_v$ containing these player statistics:

$$\mathbf{x}_v = \begin{bmatrix} \text{Player Name, Goals, Successful Dribbles,} \\ \text{Tackles, Accurate Passes Percentage} \\ \text{Rating, Goal Conversion Percentage} \\ \text{Interceptions, Clearances} \\ \text{Accurate Passes, Key Passes} \end{bmatrix}.$$

This vector comprehensively describes the player's performance and contributions during the match.

4.1.2 Edge Attributes

Edges $e \in E$ represent player interactions in the current event context. An edge is established between two players if a direct interaction occurs, such as a pass, tackle, or other forms of player-to-player interaction. The edge attributes capture the nature of these interactions. The following attributes are considered for each edge:

- **Player Interactions:** The interactions between players are drawn from the event-specific data. We include the following:

    - *Pass Information*: If the event involves a pass, we capture the pass's start and end locations, as well as the recipient of the pass (identified by the attribute `player_name_recipient`).

    - *Event Type and Result*: The `type_id` and `result_id` fields describe the type of event (e.g., pass, tackle) and its result (e.g., successful or unsuccessful).

    - *Spatial Coordinates*: The starting and ending coordinates of the ball, $(x_{\text{start}}, y_{\text{start}})$ and $(x_{\text{end}}, y_{\text{end}})$, which reflect the spatial dynamics of the event.

    - *xT Value and Change*: The expected threat (xT) value of the event, including the change in xT (`xT_change`) as a result of the interaction.

Each edge e_{uv} connecting player u and player v is described by a feature vector $\mathbf{e}_{uv}$:

$$
\mathbf{e}_{uv} = \begin{bmatrix} \text{Pass Information, Event Type and Result} \\ (x_{\text{start}}, y_{\text{start}}), (x_{\text{end}}, y_{\text{end}}) \\ \text{xT Value, xT Change} \end{bmatrix}.
$$

This vector provides detailed information about the interactions between players during the match event.

4.1.3 Graph Representation

Each event in the soccer match is represented as a graph $G = (V, E)$, where the nodes V correspond to the players, and the edges E represent the interactions. The node feature matrix $\mathbf{X} \in \mathbb{R}^{|V| \times d}$ and the edge feature matrix $\mathbf{E} \in \mathbb{R}^{|E| \times d'}$ are the inputs to our graph neural network (GNN), where d and d' represent the dimensionality of the node and edge feature vectors, respectively.

This graph representation, with both player-specific attributes and interaction-based features, enables our model to capture complex dynamics in soccer matches, allowing for the analysis of every event.

4.2. Model Architecture

This section describes the architecture of the graph neural network, which predicts changes in expected threats (xT) for every soccer event. In doing so, the GNN learns both node- and edge-level features from dynamically formed graphs for every event. The architecture embeds spatial, temporal, and interaction-based features, further enriched with centrality metrics that better define the contributions of pivotal players.

4.2.1 GNN Model

The core of our model is based on a Graph Convolutional Network (GCN) that processes both node and edge features, enabling the model to learn from the structure of player interactions during match events. The GNN model consists of the following components:

- **Graph Convolution Layers:** Our model employs two layers of graph convolutions, implemented using the `GCNConv` operator from the PyTorch Geometric library. Each layer aggregates information from a player's neighboring nodes and edges, effectively learning representations that capture individual player attributes and interactions with others. Let $\mathbf{H}^{(l)}$ represent the hidden node feature matrix at layer l, and $\mathbf{A}$ be the adjacency matrix of the graph. The forward pass at each layer is computed as:

$$
\mathbf{H}^{(l+1)} = \sigma(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}),
$$

where σ is the activation function (ReLU in our case) and $\mathbf{W}^{(l)}$ is the trainable weight matrix for layer l.

- **Edge Feature Processing:** In addition to node features, edge features are processed through a multi-layer perceptron (MLP) to learn representations for player interactions. The MLP consists of two fully connected layers with ReLU activations. For each edge, let $\mathbf{e}_{uv}$ represent the edge feature vector between nodes u and v. The MLP transforms $\mathbf{e}_{uv}$ as follows:

$$
\mathbf{e}'_{uv} = \text{ReLU}(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{e}_{uv})),
$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weight matrices for the MLP layers. The processed edge features are then incorporated into the GNN to enrich the representation of player interactions.

- **Global Mean Pooling:** After the graph convolutions, a global mean pooling layer is applied to aggregate node embeddings into a single graph-level embedding, representing the overall event. This step ensures that the contributions of all players in the event are considered
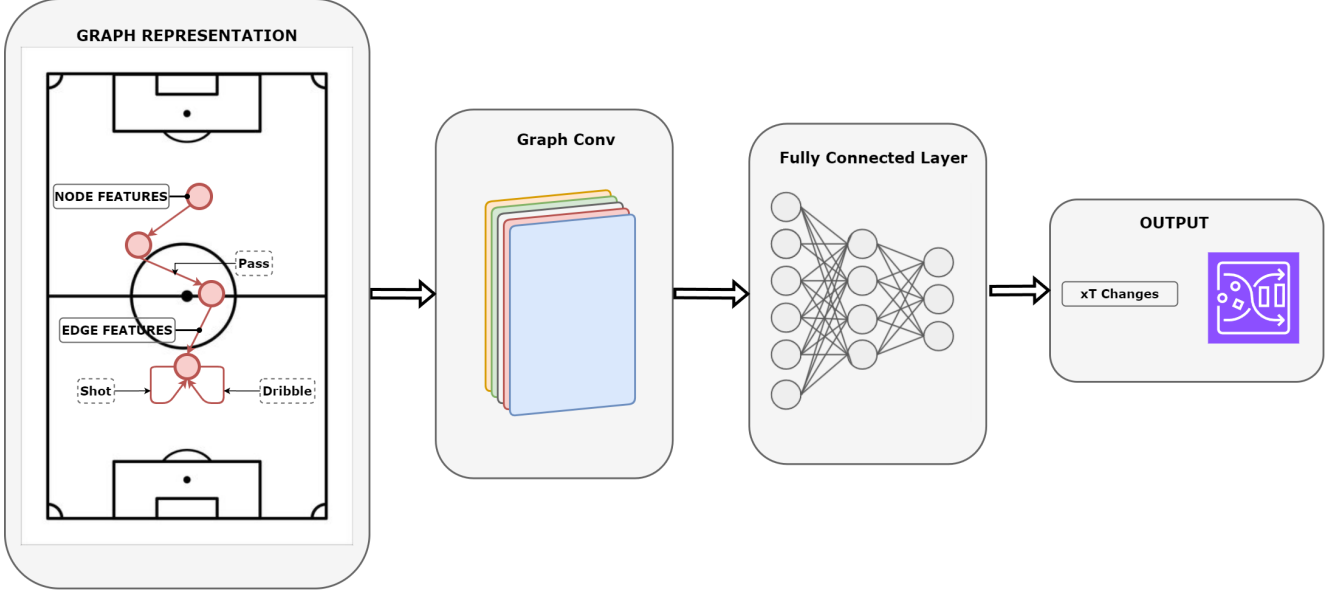
Figure 1: Model Visualization: Graph representing interactions and players will be fed to Graph convolution blocks and fully connected layer for prediction

when predicting the xT change. The global pooling operation is defined as:

$$\mathbf{z} = \frac{1}{|V|} \sum_{v \in V} \mathbf{h}_v,$$

where $\mathbf{h}_v$ is the hidden representation of node v after the graph convolutions, and $|V|$ is the number of nodes (players) in the graph.

- **Fully Connected Layers:** The pooled graph embedding $\mathbf{z}$ is passed through two fully connected layers with ReLU activations to predict the final output, the xT change for the event. The output of the model is a scalar value $\hat{y}$, representing the predicted xT change:

$$\hat{y} = \mathbf{W}_4 \cdot \text{ReLU}(\mathbf{W}_3 \cdot \mathbf{z}),$$

where $\mathbf{W}_3$ and $\mathbf{W}_4$ are the weights of the fully connected layers.

The entire model is trained using the Mean Squared Error (MSE) loss function to minimize the difference between the predicted xT changes and the actual xT changes:

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

where N is the number of training samples, y_i is the true xT change, and $\hat{y}_i$ is the predicted xT change for the i-th event.

4.2.2 Integration of Node Embeddings for xT Distribution

We employ GNNs in our model to generate meaningful node embeddings that capture the influence and importance of players involved in each event. Classic centrality metrics from graph theory, such as degree centrality, betweenness centrality, and closeness centrality, are helpful to measure a player's role in static networks. Still, they fail to fully capture all the complexity and dynamics of soccer matches. In other words, we do not use these static measures but make use of the learned node embeddings from the GNN to dynamically distribute the values of expected threat (xT) to the players based on their involvement in events taking place in the match.

In this sense, the ultimate goal of training the GNN should not be considered computing a more accurate xT. Rather, it is to attach this xT to individual players based on what they have learned from the representations (embeddings) in the graph. These embeddings encode much information about how the player interacts, contributes to ball progression, and affects game dynamics. The GNN model learns these representations by performing multi-layer convolutional operations capturing spatial and temporal dependencies in the match events. When the model trains, node embeddings can be extracted to distribute the xT value for each event proportionally. The steps involved include the following:

1. **Extract Node Embeddings:** After passing the data through the GNN layers, the node embeddings for

5

each player involved in the event are obtained. These embeddings represent the player's contribution to the event in a multidimensional feature space.

2. **Normalize Node Embeddings:** The magnitude (norm) of each node embedding is calculated. This magnitude reflects the player's relative importance within the event context. To ensure the xT value is distributed proportionally, the node embeddings are normalized such that the sum of all node magnitudes for an event equals the total xT.

3. **xT Distribution:** The xT value for each event is then distributed among the players based on the normalized node embeddings. Players with higher embedding magnitudes receive a more significant share of the xT, as their learned representation suggests a stronger contribution to the event's outcome.

This approach provides a dynamic and context-aware mechanism for assigning xT to players, leveraging the power of GNNs to go beyond traditional centrality metrics. By using the node embeddings learned from the model, we can reflect the nuanced roles players play in ball progression and game facilitation, particularly for those players whose contributions are often overlooked in conventional statistical analysis.

Integrating node embeddings as a basis for xT distribution allows for a more accurate and fair assessment of player contributions. Players who are involved in critical transitions or build-up play but may not be directly associated with goal-scoring opportunities are appropriately credited for their impact on the match. This leads to a more holistic evaluation of player performance, which is particularly beneficial for defensive midfielders and central defenders, who play pivotal roles in facilitating team dynamics.

4.3. Temporal Context

The temporal aspect of soccer matches is critical in understanding player performance and interactions. Soccer is a continuous, dynamic sport where the actions of players and teams evolve throughout the game, making it essential to account for the time-dependent nature of events. To accurately model player contributions, it is important to capture the spatial relationships between players and the sequence of events leading up to each match situation. Our approach incorporates temporal context by integrating information from preceding events into the graph representation used by the graph neural network (GNN) model.

In our methodology, we represent each event not as an isolated occurrence but within the broader temporal context of the previous k events. This enables the model to consider how the current event is influenced by prior actions and decisions made by players. Specifically, we embed the

temporal sequence into the graph structure by ensuring that both nodes (players) and edges (interactions) reflect information from recent events, thus allowing the model to learn temporal dependencies that impact player contributions.

The integration of temporal context is achieved through the following mechanisms:

1. **Sliding Window of Events:** For each event, we include the previous k events in the graph construction. This sliding window ensures that the graph captures recent interactions and player involvement, allowing the model to understand how prior plays influence the current game state. The choice of k is determined based on empirical performance, balancing between capturing relevant temporal information and maintaining computational efficiency.

2. **Temporal Edge Features:** Each edge in the graph is associated with not only the spatial features (such as pass frequency and distances) but also temporal features, including the time elapsed since the start of the match and the time difference between successive events. These features provide the model with a richer representation of the temporal flow of the game, helping to capture how quickly actions unfold and how the timing of events affects player interactions and outcomes.

3. **Event Sequencing and Causality:** By incorporating temporal context, we allow the model to learn causality in player interactions. For instance, a defensive interception may trigger a counterattack, where a sequence of quick passes leads to a scoring opportunity. The temporal sequencing of these actions is vital for understanding the causal relationship between defensive and offensive contributions. Our approach ensures that such sequences are represented in the graph, thereby enhancing the model's ability to assign xT values based on the spatial positioning of players and their roles in key moments leading up to goal-scoring opportunities.

4. **Temporal Weighting:** In addition to considering the previous k events, we introduce a temporal weighting mechanism that emphasizes the most recent actions while gradually reducing the influence of earlier events. This weighting scheme allows the model to focus on more relevant recent interactions, reflecting that events occurring closer to the current event are often more influential in determining outcomes. Temporal weighting is applied to both node and edge features, ensuring that the model captures the time-varying importance of player actions.

By incorporating temporal context, our GNN model can capture the immediate spatial relationships between players and the broader temporal dynamics of the game. This approach enhances the model's ability to assign credits to players based on their involvement in sequences of events rather than evaluating actions in isolation. Including temporal features is particularly valuable in assessing players who contribute to build-up play or transitions, as their influence often spans multiple events and cannot be adequately captured by static event-based metrics.

Overall, the integration of temporal context ensures that our model reflects the dynamic, evolving nature of soccer matches, providing a more accurate and nuanced understanding of player contributions over time.

5. Experiments and Results

5.1. Experimental Setup

This section outlines the methodology used to train and validate the SoccerGNN model, including hardware specifications, training process, and evaluation metrics.

5.1.1 Training and Validation

The SoccerGNN model was trained using the Adam optimizer with a learning rate of 1×10^{-4}, incorporating a weight decay of 1×10^{-4} to prevent overfitting. The training process was conducted over 25 epochs, with the dataset split into 80% for training and 20% for validation. The model was trained using backpropagation and stochastic gradient descent, with a batch size adjusted based on system memory.

The model was trained on a high-performance L4 GPU with high-RAM configuration. The training process, including all 25 epochs, took approximately 3 hours to complete. This hardware configuration ensured faster computation during the graph neural network operations and allowed for efficient processing of the high-dimensional data.

The SoccerGNN architecture consists of two graph convolutional layers followed by fully connected layers for node embeddings and graph-level predictions. The model outputs the expected goals (xG) for each event. Weight initialization was performed using the Xavier method for linear layers and Kaiming initialization for graph convolution layers, helping to stabilize gradients and prevent issues like gradient vanishing or explosion during training.

A learning rate scheduler was employed to reduce the learning rate by a factor of 0.5 every 10 epochs, promoting convergence and avoiding local minima. Early stopping was also applied to avoid overfitting, with the model stopping if validation loss did not improve for a pre-defined number of epochs.

5.1.2 Evaluation Metrics

The performance of the SoccerGNN model was evaluated using two regression metrics: **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)**. These metrics provide a quantitative assessment of how closely the predicted xG values align with the actual values.

Mean Squared Error (MSE): MSE measures the average squared difference between the predicted and actual xG values. It is more sensitive to larger errors, making it useful for identifying significant deviations:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Mean Absolute Error (MAE): MAE measures the average magnitude of errors between the predictions and actual values. Unlike MSE, it is less sensitive to outliers, making it useful for evaluating the overall fit of the model:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Both MSE and MAE were logged at the end of each epoch to monitor both training and validation performance. These metrics provided insight into how well the model generalizes to unseen data and helped ensure that the model's predictions became progressively more accurate over time.

5.2. Results

In this section, we present the performance of the SoccerGNN model during training and validation. The results include visualizations of the training and validation loss, as well as evaluation metrics (MSE and MAE) tracked across epochs.

5.2.1 Model Performance

The model's performance was evaluated using both training and validation data across 25 epochs. The training loss steadily decreased, while the validation loss showed a similar trend, suggesting effective learning and generalization.

Validation Loss: The figures below illustrate the progression of the validation losses (MSE and MAE) over the epochs.

Training and Validation MSE/MAE: The figures below present the progression of MSE and MAE across both the training and validation datasets, providing additional insight
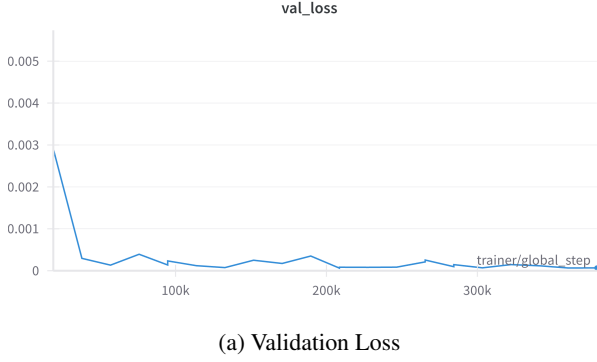
(a) Validation Loss

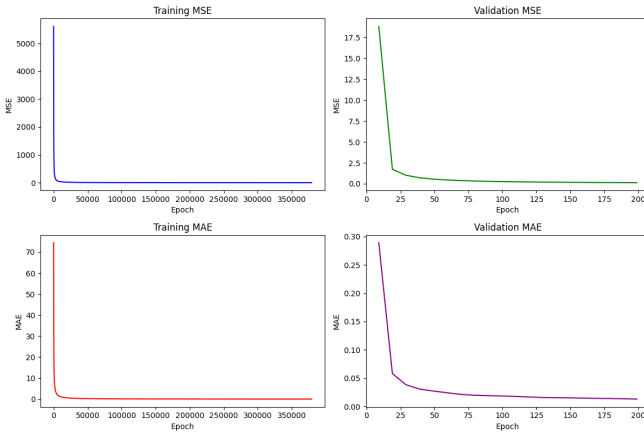Figure 2: Validation Loss over Epochs

Figure 3: Combined Learning Curves: Training and Validation Loss, MSE, and MAE

into how well the model minimizes error throughout the epochs.

The final validation MSE and MAE values reached [insert value] and [insert value], respectively, indicating the model's ability to generalize to unseen data.

5.2.2 Player Evaluation

One of the core objectives of this study is to evaluate individual player contributions using the learned node embeddings from the GNN. These embeddings provide insight into each player's role in ball progression and team coordination.

xT Contribution Assignment: The node embeddings were used to assign xT contributions to each player, with higher node embedding magnitudes corresponding to greater xT contributions. This approach highlights players who significantly impact expected goals through indirect actions such as ball progression or defensive maneuvers.

Top Player for Each Team: The table below highlights the top player from each team based on xT contributions. This ranking is determined by the node embeddings learned from the GNN, which quantify each player's influence in facilitating team actions and goal-scoring opportunities.

Rank	Player	xT	Team
1	Simon Francis	2.289	AFC Bournemouth
14	Aaron Ramsey	1.176	Arsenal
63	Ashley Westwood	0.427	Aston Villa
70	Nemanja Matić	0.395	Chelsea
59	Yohan Cabaye	0.460	Crystal Palace
5	Gareth Barry	1.548	Everton
29	Danny Drinkwater	0.823	Leicester City
48	Jon Flanagan	0.529	Liverpool
77	Aleksandar Kolarov	0.364	Manchester City
11	Daley Blind	1.251	Manchester United
23	Jonjo Shelvey	0.958	Newcastle United
131	Lewis Grabban	0.177	Norwich City
44	Victor Wanyama	0.580	Southampton
3	Glenn Whelan	1.648	Stoke City
182	Jan Kirchhoff	0.080	Sunderland
21	Ashley Williams	0.965	Swansea City
4	Toby Alderweireld	1.621	Tottenham Hotspur
141	Alessandro Diamanti	0.159	Watford
192	Joleon Lescott	0.063	West Bromwich Albion
40	Manuel Lanzini	0.614	West Ham United

Table 1: Top Player from each Team by xT

This table provides an overview of the top-performing player from each team, highlighting their key contributions to expected threat (xT).

Overall Top 20 Players: The following table displays the top 20 players ranked by their xT contributions in a specific match. These rankings are derived from the GNN model, showcasing the players with the most significant influence in facilitating goal-scoring opportunities.

This ranking emphasizes players whose contributions may not be reflected in traditional statistics like goals or assists but play pivotal roles in linking play or disrupting the opposition. For example, Andrew Surman, a midfielder, topped the ranking due to their ability to distribute the ball and maintain team cohesion despite not registering any direct goal involvement.

5.3. Analysis of Findings

The results of this study demonstrate the ability of the GNN model to identify key players whose contributions may not always be reflected in traditional metrics like goals or assists. By utilizing Expected Threat (xT) values, the model sheds light on players who play crucial roles in team

Rank	Player	xT	Team
1	Simon Francis	2.289	AFC Bournemouth
2	Andrew Surman	2.193	AFC Bournemouth
3	Glenn Whelan	1.648	Stoke City
4	Toby Alderweireld	1.621	Tottenham Hotspur
5	Gareth Barry	1.548	Everton
6	Eric Dier	1.470	Tottenham Hotspur
7	Charlie Daniels	1.438	AFC Bournemouth
8	Ross Barkley	1.433	Everton
9	Matt Ritchie	1.405	AFC Bournemouth
10	John Stones	1.273	Everton
11	Daley Blind	1.251	Manchester United
12	Jan Vertonghen	1.223	Tottenham Hotspur
13	Steve Cook	1.220	AFC Bournemouth
14	Aaron Ramsey	1.176	Arsenal
15	Harry Arter	1.176	AFC Bournemouth
16	Dan Gosling	1.110	AFC Bournemouth
17	Ibrahim Afellay	1.096	Stoke City
18	Michael Carrick	1.069	Manchester United
19	Morgan Schneiderlin	1.067	Manchester United
20	Erik Pieters	1.002	Stoke City

Table 2: Top 20 Players by xT

dynamics, particularly those involved in creating scoring opportunities through indirect actions like ball progression, positioning, and defensive contributions.

AFC Bournemouth's Tactical Effectiveness: AFC Bournemouth features prominently in the overall top 20 xT rankings, with six players making the list. This points to the club's highly effective tactical setup, which may have been centered around dynamic midfield play and forward momentum. Players like Simon Francis and Andrew Surman excelled in their ability to generate offensive opportunities, showcasing their pivotal roles in controlling the game. The high ranking of Bournemouth players reflects a system that emphasizes coordinated team movement and the facilitation of scoring threats across multiple positions, not just the forwards.

Evolving Defensive Roles: The inclusion of traditional defenders like Toby Alderweireld and Eric Dier from Tottenham Hotspur in the top 20 rankings underscores the modern evolution of defensive roles in soccer. Rather than focusing solely on stopping attacks, these players contributed significantly to their team's offensive plays. Their ability to distribute the ball from deep positions, engage in build-up play, and even participate in forward movements illustrates how defenders are now required to facilitate attacks, making them integral to both defensive solidity and offensive creativity.

Midfield Mastery: Midfielders such as Glenn Whelan from Stoke City and Ross Barkley from Everton also featured prominently in the rankings. Their high xT values point to their ability to maintain possession, distribute the ball efficiently, and control the tempo of the game. These players excel in linking defense with attack and ensuring the smooth progression of the ball into dangerous areas. Notably, Andrew Surman's high ranking highlights his ability to be a critical playmaker despite not recording goals or assists. This reinforces the importance of midfielders who contribute to team cohesion and ball movement rather than just the final output of goals.

Contributions from Lesser-Known Players: Players like Simon Francis, who topped the ranking, and Steve Cook, another high-ranking Bournemouth defender, are examples of players whose contributions are often overlooked in traditional analyses. The xT metric allows for a more nuanced understanding of their roles, particularly in facilitating scoring opportunities through interceptions, progressive passes, and key defensive actions. These insights could be invaluable for tactical analysis, helping coaches identify hidden contributors who may not receive widespread recognition but are critical to the overall success of their team.

Team Balance and Strategy: The findings also reveal that successful teams are those that have balanced contributions from multiple players across different positions. For example, Everton had three players in the top 20, highlighting the balanced nature of their team's strategy, where both midfielders (Ross Barkley) and defenders (John Stones) play key roles in generating scoring opportunities. This balance ensures that a team is not overly reliant on a single player for creativity and progression, making them more resilient in various match scenarios.

5.4. Implications for Coaches and Clubs

With our model, it is then possible for the coaches to coach players specifically toward developing those particular player roles, strategically plan their game tactics, and make well-informed decisions during a match by singling out the key but often unnoticed performers who hold key interactions together in matches. Clubs can use these insights to recruit players more efficiently and evaluate transfers at a higher level of effectiveness through authentic on-field contributions. Usually, underrated players are relatively cheaper. This allows better insight into the team dynamics itself, leading to the derivation of player development programs and long-term team composition, besides financial planning, all in aid of better competitive strategies and fan engagement as there is a more nuanced appreciation of player value.

5.5. Limitations

Many other underlying factors may affect the model's effectiveness across leagues and at different levels of play because of different styles and strategies. Furthermore, the dynamic nature of football and potential biases in model training could affect the accuracy and relevance of the findings, requiring careful management and continuous improvement of the model to address these issues accurately. Lastly, although clubs could use this to identify key players they want, we still lack a way to determine whether a player can fit in the new environment.

6. Conclusion

6.1. Summary of Contributions

The project has proven that GNNs will improve the analysis of soccer matches by presenting a dynamic interplay of roles across players, going beyond existing quantifications like goals and assists. Formulating each event in a soccer match as a graph with nodes representing players and edges for interactions between them, this model, including spatial and temporal data, was sensitive to the small differences between player contributions. This approach thus gave a wider picture of player impact and underscored the relevance of numerous facilitative roles, most of which get lost in standard analyses. This ability of the GNN model to predict, especially the attribution of changes in xT to individual players based on their interactions, certainly opened up new perspectives that could significantly impact tactical decisions, training programs, and recruitment strategies. Even though it is littered with data quality issues, model complexity, and computational demands, the results surely help in making a point of how applying advanced machine learning techniques can be very transformative for sports analytics. This work should open new research efforts and practical applications that could help to refine coaching strategies or improve our understanding of game dynamics in professional football.

6.2. Future Work

Future work building on this study will include incorporating more detailed player and game data—including real-time tracking and physiological metrics—to enhance the accuracy and applicability of the model for real-time game scenarios. An extension across different leagues and sports would enable insights into different tactical styles and player roles, making the model generalizable. Increasing the interpretability of models will make these advanced analytics more accessible to coaches and sports practitioners who are not deeply versed in machine learning. Further exploration into the predictive capacity of player injuries and fatigue management will greatly benefit players' health and overall team performance. Such advancements could signif-

icantly change the landscape of sports analytics, providing a more dynamic and strategic interpretation across various sporting disciplines.

References

[1] G. Anzer and C. Stöcker. Detection of tactical patterns using semi-supervised graph neural networks in soccer matches. In *Proceedings of the MIT Sloan Sports Analytics Conference*, pages 1–12, 2020.

[2] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, pages 1851–1861, New York, NY, USA, 2019. ACM.

[3] G. Liu, Y. Luo, O. Schulte, and T. Kharrat. Deep soccer analytics: learning an action-value function for evaluating soccer players. *Data Mining and Knowledge Discovery*, 34(5):1531–1559, 2020.

[4] L. Pappalardo, P. Cintia, and A. Rossi. Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. In *Proceedings of the ACM International Conference on Intelligent Systems and Technology*, pages 1–15. ACM, 2021.

[5] K. Singh. Introducing expected threat (xt).

[6] P. Velickovic, M. Simonovsky, and Q. Xu. Graph attention networks for predicting sports outcomes. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):1234–1245, 2021.

A. GitHub Repository

The source code related to this paper can be found at the following URL:
```
https://github.com/JackyJiang0410/
GoalNet.git
```