Predicting successful passes in football: Benchmark dataset and GNN ablation study

V. Stropnik and L. Šubelj

University of Ljubljana, Faculty of Computer and Information Science, Slovenia

In modern association football, data and the knowledge derived from it play a crucial role in forming tactical plans and analyzing games. We explore the capability of modeling a passer's decision-making when selecting a target during a football match using graph deep learning on player-formation networks. We present a methodology for constructing a benchmark dataset of networks from open data provided by the sports data company Hudl Statsbomb [1]. The dataset contains all viable ground passes from the 2022 Men's FIFA World Cup, and the 2020 and 2024 iterations of the UEFA Men's European Championship.

The constructed networks $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ are parametrized with sets of player nodes $\mathcal{V}$ and different configurations of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ summarized by the adjacency matrix $\mathbf{A}$. We evaluate three instantiations of $\mathbf{A}$ (hub-and-spokes, fully connected, with opponents) to determine the most successful configuration for the downstream machine learning tasks. Furthermore, the networks are enriched with node features $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times n}$ and edge features $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}| \times m}$. These correspond to different match-moment attributes $\mathbf{X}$ and their derived player interactions $\mathbf{E}$, such as the heuristically calculated movement trajectories of (otherwise anonymous) players or the fraction of controlled space between two players [2] (Voronoi vs Spearman). We validate the contribution of these features to our learning objectives in a thorough ablation study.

The final dataset contains 80,332 network representations from 166 matches. We evaluate the usefulness of the dataset on three tasks: regression of the coordinates of a successful pass, prediction of the target zone of the pass, based on the positional play pitch division, and classification of the role of the pass recipient. The proposed model based on the Graph Transformer architecture [3] achieves the best results in all tasks when compared to other graph convolutional architectures (state-of-the-art for similar tasks [4]) as well as baselines omitting the network structure altogether. The best-performing regression model achieves a median Euclidean distance error of 6.78 ± 0.21 yards (6.20 ± 0.19 meters). The best-performing classification models achieve an accuracy of 55 ± 2 % on the positional play zone task (20 zones) and 42 ± 1 % on the player role prediction task (23 roles). We additionally show that the model error increases with the ground-truth pass distance and that uncertainty is higher when predicting progressive passes in the attacking third.

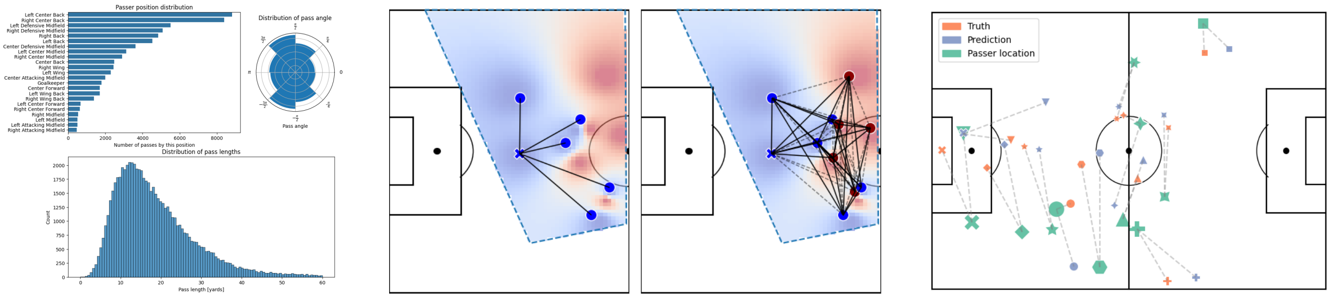

Figure 1: *(left)* Ground-truth statistics of successful passes in the dataset. *(middle)* Hub-and-spokes and fully connected networks with Spearman pitch control model. *(right)* Twelve randomly selected examples of predicted passes.

The results demonstrate that the proposed approach can effectively model the passer's decision-making and that networks enriched with appropriate features can significantly contribute to advanced football analytics. We make the dataset reconstruction script publicly available [5] (since directly sharing the dataset is outside the terms and conditions). The script generates networks compatible with the *NetworkX* library and *PyTorch Geometric* framework (`.dill` format) equipped with informative network card summaries [6].

[1] `https://github.com/statsbomb/open-data`

[2] Spearman *et al.* (2017) Physics-based modeling of pass probabilities in soccer, *MIT Sloan Sports Analytics Conference*, pp. 14.

[3] Shi *et al.* (2021) Masked label prediction: Unified message passing model for semi-supervised classification, *International Joint Conference on Artificial Intelligence*, p. 1548.

[4] Wang *et al.* (2024) TacticAI: An AI assistant for football tactics, *Nature Communications* **15**(1), 1906.

[5] Stropnik (2024) Analysis of player-formation graphs for predicting football passes. `https://github.com/wwwidonja/GraphFC`

[6] Bagrow & Ahn (2022) Network cards: Concise, readable summaries of network data, *Applied Network Science* **7**, 84.

This work has been supported in part by Slovenian Research Agency ARRS under the program P5-0168.