PREDICTING SUCCESSFUL PASSES IN FOOTBALL: BENCHMARK DATASET AND ABLATION STUDY

DATASET RECONSTRUCTION SCRIPT

Vid Stropnik (vs2658@student.uni-lj.si), **Lovro Šubelj** (lovro.subelj@fri.uni-lj.si), University of Ljubljana, Faculty of Computer & Information Science, Slovenia

INTRODUCTION

In modern association football, data and the knowledge derived from it play a crucial role in forming tactical plans and analyzing games. We model 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 Hudl Statsbomb [1]. The dataset contains all viable ground passes from the 2022 Men's FIFA World Cup, and the 2020 and 2024 UEFA Men's European Championships.

PLAYER-FORMATION NETWORKS

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 $\mathcal{A}$. We evaluate three instantiations of $\mathcal{A}$ (hub-and-spokes, fully connected, with opponents) to determine the most successful configuration for the downstream prediction tasks (Fig. 1). Furthermore, the networks are enriched with different node and edge features. These include basic player information, different match-moment attributes, and derived player interactions such as heuristically calculated movement trajectories and the fraction of controlled space between players [2] (Fig. 2). The movement trajectories are estimated from the most recent preceding positions of anonymous player nodes, which we identify by solving the graph alignment problem. We validate the contribution of these features to our learning objectives in a thorough ablation study.

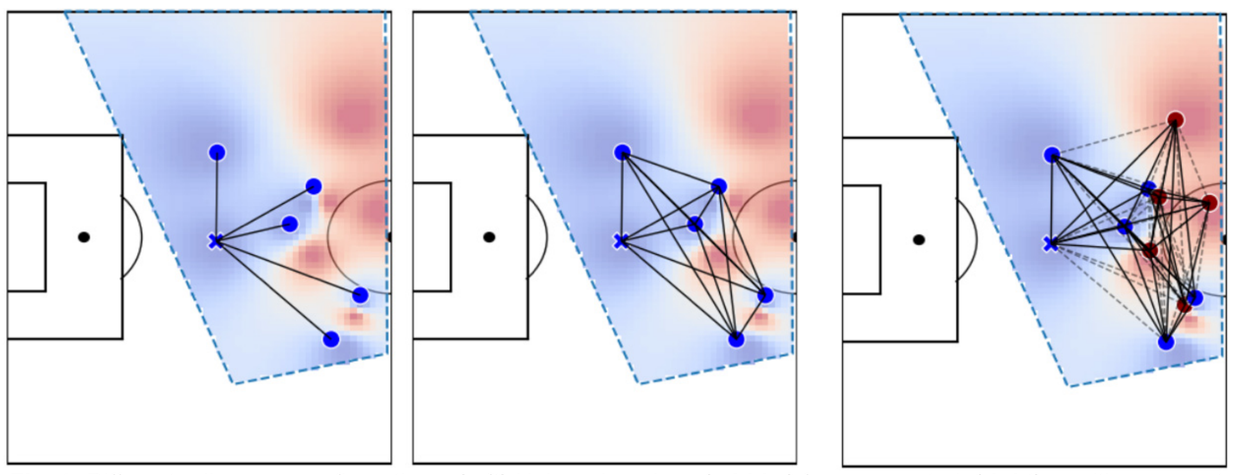

Fig. 1: Different adjacency configurations. **(left)** Hub-and-spokes, **(middle)** fully connected, **(right)** ... with opponents.

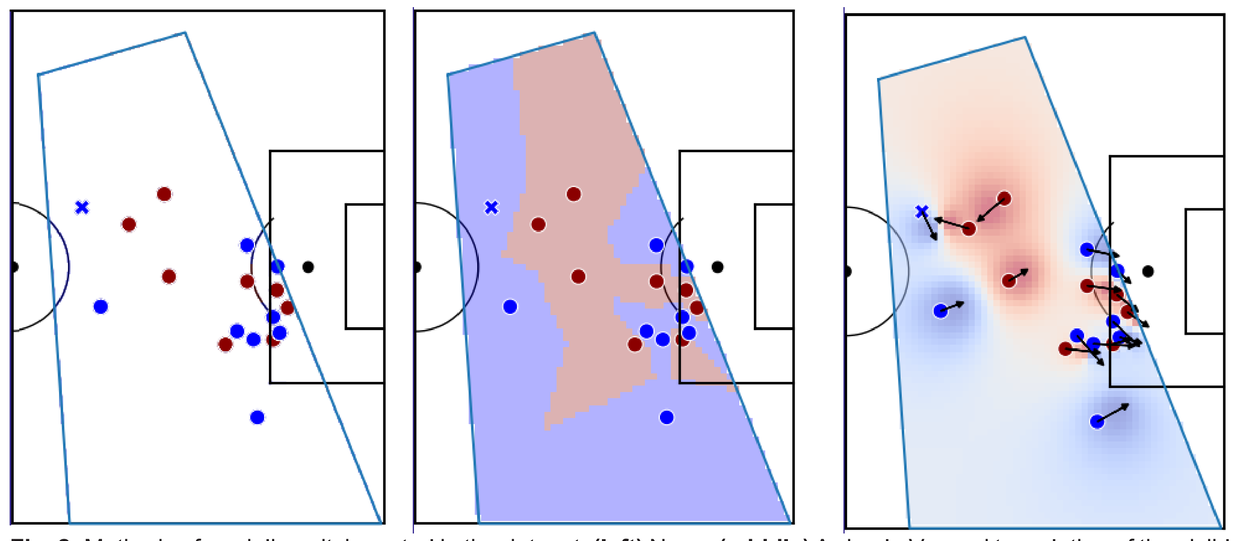

Fig. 2: Methods of modeling pitch control in the dataset. **(left)** None. **(middle)** A simple Voronoi tesselation of the visible space. **(right)** Spearman's model of control using heuristic movement trajectories.

DATASET

The dataset contains 80,332 network representations from 166 matches (Fig. 3, 4). The player locations are sampled from video-stream data and are thus limited by the camera's field of view. We make the dataset reconstruction script publicly available [3]. The script generates networks and their features as **Pandas** dataframes, compatible with the **NetworkX** library and the **PyTorch Geometric** framework, and equipped with informative network card summaries [4].

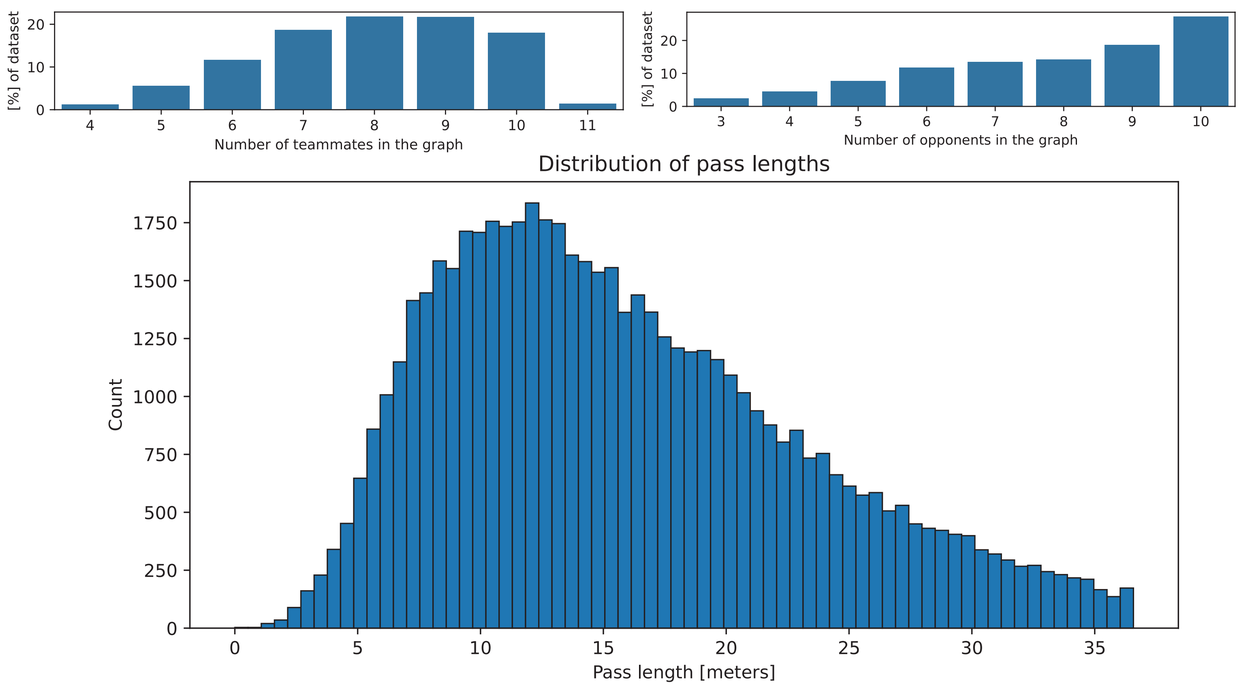

Fig. 3: The distributions of **(top)** node (player) counts in the dataset, and **(bottom)** successful pass lengths, to which the graphs in the ablation study correspond. The median pass measures ~ 12 meters.

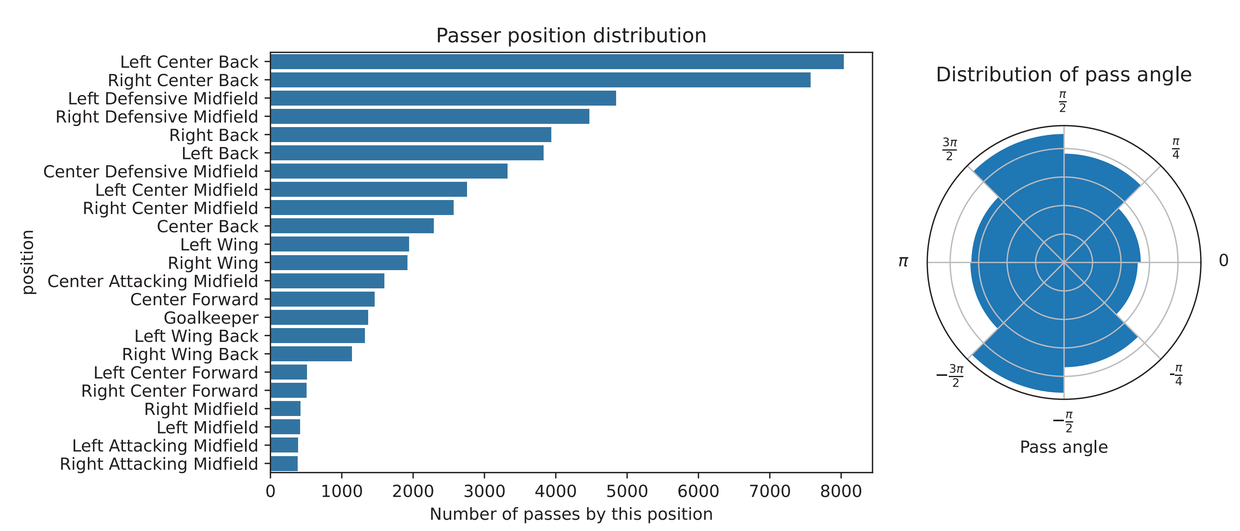

Fig. 4: Ground-truth statistics of successful passes. The majority of passes are made by backline players and midfielders, and aren't necessarily progressive.

EXPERIMENTAL RESULTS

We evaluate the applicability of our dataset on three downstream prediction tasks:
1. **Regression of the target coordinates** of a successful pass (Tbl. 1 & Fig. 5-6),
2. Prediction of 20 target **zones of the positional play pitch division** (Fig. 7, left),
3. Classification of **23 playing roles** of the pass recipient (Fig. 7, right).

The proposed model based on the Graph Transformer architecture [5] can effectively model the passer's decision-making and achieves the best results in all tasks when compared to the state-of-the-art [6]. We show that utilising the graph structure, utilizing graph edge features, and estimating the players' movement trajectory all contribute to the predictive performance of all tasks. The detailed ablation is shown for the pass coordinate regression task (Tbl. 1).

Backbone	Pitch Control	Trajectories?	Connectivity	Opponents?	Median error distance [m]
Transformer	Spearman	Yes	Fully Connected	Yes	6.46 ± .05
GATv2	Spearman	Yes	Fully Connected	Yes	7.13 ± .09
Transformer	**Voronoi**	**Yes**	**Fully Connected**	**Yes**	**6.19 ± .18**
Transformer	Voronoi	No	Fully Connected	Yes	8.36 ± .26
Transformer	Voronoi	Yes	Fully Connected	No	7.21 ± .07
Transformer	Spearman	Yes	Hub-and-Spokes	No	7.25 ± .07
GCN	Voronoi	Yes	Fully Connected	Yes	14.32 ±.05
n/a (MLP)	Spearman	Yes	n/a	No	11.01 ± .04

Tbl. 1: Results of the target regression ablation study. The Transformer architecture outperforms alternative GNN backbones, while the inclusion of pitch control features and heuristic player trajectories improves model performance.

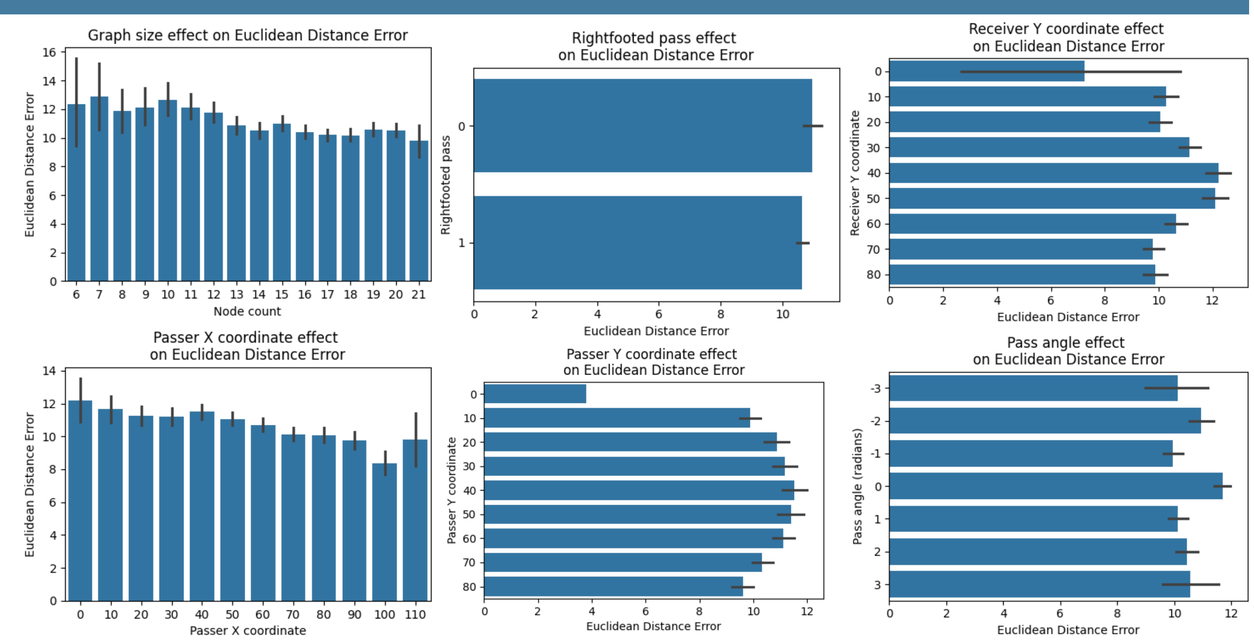

Fig. 6: Diagnostic plots for the pass coordinate regression task. Passes from the sidelines have a smaller error.

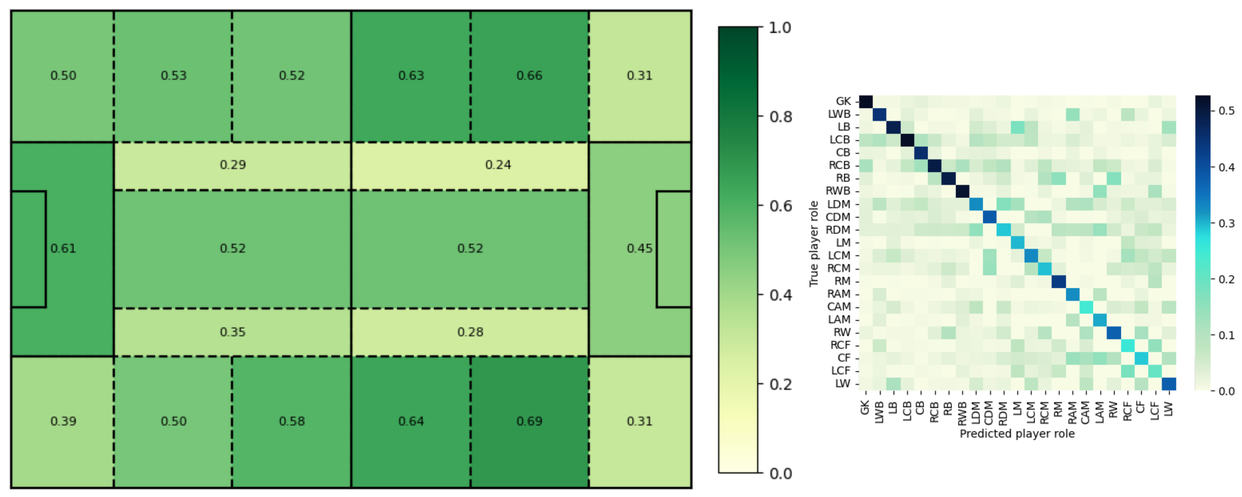

Fig. 7: The best-performing models achieve an accuracy of **(left) 55 ± 2%** on the positional play zone task and **(right) 42 ± 1%** on the player role prediction task.

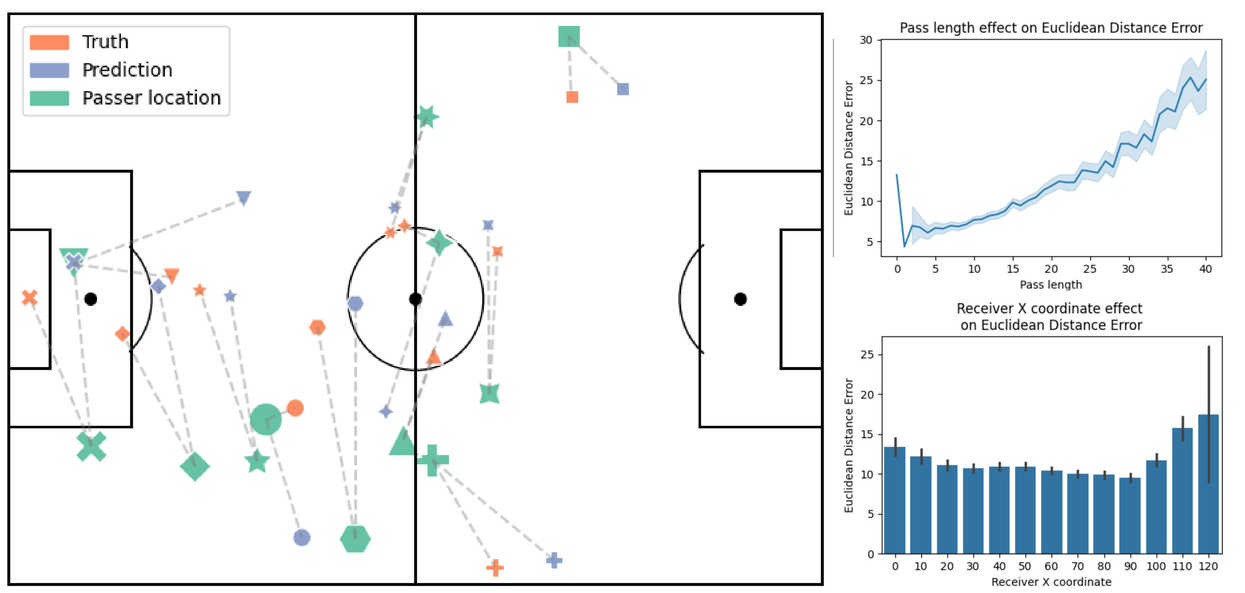

Fig. 5: (left) Twelve examples of the predicted coordinates of passes. **(right)** The model performance deteriorates with the pass distance and progress up the pitch.

[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] Stropnik (2024) Analysis of player-formation graphs for predicting football passes. https://github.com/wwwidonja/GraphFC
[4] J. Bagrow, Y.-Y. Ahn, Network cards: concise, readable summaries of network data (2022).
[5] Shi et al. (2021) Masked label prediction: Unified message passing model for semi-supervised classification, International Joint Conference on Artificial Intelligence, p. 1548.
[6] Wang et al. (2024) TacticAI: An AI assistant for football tactics, Nature Communications 15(1), 1906.