Let's Penetrate the Defense: A machine learning model for prediction and valuation of penetrative passes

Pegah Rahimian[1], Dayana Grayce da Silva Guerra Gomes[1], Fanni Berkovics[1], and Laszlo Toka[1,2]

[1] Budapest University of Technology and Economics, Budapest, Hungary
[2] MTA-BME Information Systems Research Group
{pegah.rahimian,gomes.dayanagrayce,fanni.berkovics}@edu.bme.hu
{toka.laszlo@vik.bme.hu}

Abstract. Moving forward and penetrating the defensive zones is crucial for goal scoring in soccer games, yet it involves risky tactics. We propose a novel metric called Expected Value of Potential Penetrative Pass, which measures the likelihood of a potential penetrative pass creating scoring/conceding situations. We show how such a pass value accounting for the effects of crossing defense lines can be decomposed into elementary components. Using the UEFA EURO 2020 spatiotemporal dataset, we train several conventional machine learning and deep learning models to estimate these expected values for all potential penetrative pass situations in the dataset. For the best five and worst five teams in the dataset, we provide a trade-off between the ability to perform penetrative passes, and the expected value of goals those create. Finally, we show the impact of different field sections as the starting location of the penetrative pass to be performed and to create goal scoring situations.

Keywords: Soccer Analytics · Deep Learning · Penetrative Pass.

1 Introduction

Soccer is one of the most popular sports in the world. Similar to any other team sport, the interaction between players, passes in particular, are crucial. Passes are the most frequent actions of the soccer games usually reaching around 800 per game [5]. Thus, measuring the pass impact and capability of the players and teams in terms of successful and valuable passes has gained paramount attention among researchers and sports analyzers. Although media focus on metrics such as the number of passes and their success ratio, and the industry cares more about metrics related to scoring opportunities such as key passes, sports data scientists have combined long- and short-term objectives and introduced novel metrics that consider the effect of short term rewards (e.g., a short successful pass in midfield) in long term objectives (e.g., open valuable space somewhere else on the pitch and create scoring chances) [11], [5],[8],[4],[3],[9]. However, most of these works consider a generic pass or differentiate passes according to their length (e.g., short passes or long balls). On the other hand, Michalczyk in a Stats Perform blog post discusses the effect of line-breaking passes in creating chances[3]. In order to measure the teams' and players' capability of performing penetrative passes, Sotudeh introduces a metric called Potential Penetrative Passes (P3) that measures how they completed the penetrative passes compared to the number of times they had the potential to do so [13]. In this work, we also take into consideration that players need to move forward to create chances and score goals. Furthermore, we concentrate on the penetrative feature of the passes and introduce a decoupled novel metric called Expected Value of Potential Penetrative Pass (xPPP) to measure the likelihood of a potential penetrative pass creating scoring/conceding situations. By analyzing the StatsBomb360 datasets for UEFA EURO 2020, we rank teams according to their performance in penetrative pass completions and the expected value they can create by doing so. We also show the impact of different field sections on their performance.

2 Related Work

Penetrative pass refers to a pass that breaks through the opposition's defensive zone(s). However, it is not straightforward to detect dynamic defensive zone(s) in each situation due to the complex nature of a soccer game. To this end, several studies propose different clustering algorithms for opponent players. Spectral clustering is used by Fernandez et al. [5] to the mean of opponent positions to detect dynamic lines of their formations, and Rahimian et al. [10] to the opponents' locations and velocities to detect pressure on the ball holder. Michalczyk [1] in his Stats Perform blog post proposes Jenks natural breaks optimization with three clusters on outfield players. A role-based approach has been suggested in [9], [2] which finds the formation structure and captures if a pass is attempting to break the line of defenders. Our definition of penetrative passes in this work is the most similar to the definition by Sotudeh [13] as when a player is in a passing situation and there is at least a teammate to receive the ball inside a defensive zone (i.e., a polygon created by the opponent players in front of him (c.f. Figure 1)). However, his work does not provide any pass impact model and simply measures the players' and teams' ability to complete penetrative passes, given all the potential penetrative pass situations. With regards to the pass impact, several action valuation models have been proposed (e.g., [6],[9],[3],[5],[12]). In this work, we investigate the teams' abilities accounting for their performance in completing penetrative passes, and the expected value they can create by doing so.

3 Penetrative Pass Prediction and Valuation

Penetrating the defense is a key tactic in invasion games. When a penetrative pass gets through an opponent's defensive zone, it removes the players it has

[3] https://bit.ly/39uQe6Q

passed by and gets the player in possession of the ball closer to the goal. However, such passes have their own risks and rewards. A successful penetrative pass can lead the player in possession of the ball closer to scoring a goal, and an unsuccessful one might lead to a possession loss. The main objective of the present work is to develop a decoupled metric that measures the expected penetrative value of a pass (i.e., the likelihood of a potential penetrative pass to create goal opportunities). For this aim, in this section, we first describe the dataset and elaborate on our designed machine learning setups to predict the penetrative passes and measure their impact on goal scoring/conceding. We then introduce our proposed metrics to measure players' and teams' performance in terms of penetrative pass completion and valuation.

3.1 Dataset and Preprocessing

StatsBomb events and StatsBomb360 events datasets[4] for UEFA EURO 2020 were obtained using the statsbombpy API[5]. Both datasets were fetched and then merged. The StatsBomb events dataset contains 110 columns detailing aspects of each event, and the StatsBomb360 events dataset contains 7 columns depicting the position of each player caught in the frame of the action for each event in the visible area. The datasets included all matches played by 24 teams participating in UEFA EURO 2020.

3.2 Potential Penetrative Pass Situation

As the first step, we follow the definition of a potential penetrative pass according to [13], as when a player is in a passing situation and there is at least a teammate to receive the ball inside a defensive zone (i.e., a convex hull created by the opponent players in front of him (c.f. Figure 1)). For constructing these situations from the dataset, we first filtered the synchronized dataset for all "Forward Pass" actions with any outcomes such as being penetrative or not, and being successful or turnover. This is because we are constructing a pre-event situation. Furthermore, we assume that the first one-third and the last one-fourth of the field should be ignored since passes that originated in these areas are not valuable for our analysis. Thus, we consider only the area between 40 and 90 meters on the touchline of the field. Now for all filtered passes, we construct a convex hull from all visible opponent players' locations in front of the ball using ConvexHull class from the spatial library of scipy[6]. Next, we check all visible teammate players' locations in the visible area. If there exists at least one teammate player lying in the created convex hull (as the potential receiver), we mark the frame as the potential penetrative pass situation, and non-potential penetrative pass, otherwise. This is done using Delaunay class from spatial library of scipy[7].

[4] https://statsbomb.com/articles/soccer/statsbomb-announce-the-release-of-free-statsbomb-360-data-euro-2020-available-now/
[5] https://github.com/statsbomb/statsbombpy
[6] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.ConvexHull.html
[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.Delaunay.html

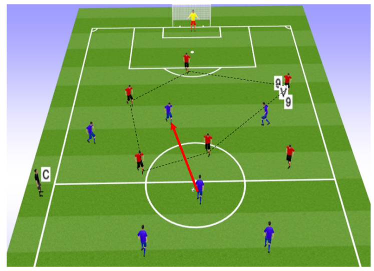


Fig. 1: An example of visible area and a successful penetrative pass

3.3 Penetrative Pass Label Generation

The next step is to check whether or not the potential penetrative pass is completed (i.e., converted into an actual penetrative pass). For this aim, for all potential penetrative pass situations, we check the outcome and receiver of the pass. If the pass was successfully received by one of the teammate players lying in the convex hull, we label the pass as "penetrative". For the rest of the passes including unsuccessful passes, and successful passes received by a teammate player outside of the convex hull, we label them as "non-penetrative".

3.4 Penetrative Pass Decomposed Model

We aim to estimate the likelihood of a potential penetrative pass to create scoring/conceding situations. However, this is a pre-event metric and we still do not know its outcome, i.e., whether it is going to be penetrative or not. Therefore, estimating such a value is not as simple as predicting the probability of goal scoring or conceding within the next preset number of actions. To estimate such a value, we propose a novel decomposed model which takes into consideration different outcomes of a potential penetrative pass, and estimates its likelihood to create scoring/conceding opportunities. To do so, we first introduce the following components:

– **xPP: Expected Value of Penetrative Pass**: Likelihood of an actual penetrative pass to create goal scoring situation;
– **xPC: Expected Penetrative Pass Completion:** Likelihood of a potential penetrative pass completion;
– **xPPP: Expected Value of Potential Penetrative Pass:** Likelihood of a potential penetrative pass to create goal-scoring/conceding situation.

Now we explain how to estimate each component.

xPP: This is a post-event metric, i.e., we know the pass was performed and the outcome was penetrative, and we want to estimate the probability of goal scoring within the next five actions. In order to calculate this metric we have:

$$xPP = \Big(Pr(\text{Penetrative pass} \rightarrow goal) \times Value(Goal)\Big) + \Big(Pr(\text{Penetrative pass} \rightarrow no - goal) \times Value(No - goal)\Big), \tag{1}$$

where $Value(Goal) = 1$, and $Value(No - goal) = 0$. Therefore, we assume that the expected value of a penetrative pass equals to Pr(Penetrative pass $\rightarrow$ goal). We treat estimating this value as a classification model. Since this is a post-event metric, we set the previous 2 actions of the current pass and the pass itself (3 actions in total) as the game state, and predicting goal scoring within the next 5 actions.

xPC: This is a pre-event metric, i.e., we have been given a potential penetrative pass situation, and we are not aware of its outcome. In order to estimate the likelihood of potential penetrative pass completion (i.e., a successful penetrative pass will be completed from a potential situation), we simply train a classification model, by setting the previous three actions of the current potential penetrative pass as the game state, and we set the labels as the outcome of the pass (i.e., penetrative or non-penetrative) like it is elaborated in Section 3.3.

xPPP: This is again a pre-event metric, in which we have been given a potential penetrative pass situation without knowing its outcome, and we would like to measure the likelihood of creating goal scoring/conceding within the next 5 actions of the current situation. Estimating this metric depends on the different outcomes of the pass. To this end, we first show how a pass value accounting for the effects of penetration can be decomposed into other components. Thus, we introduce the decoupled xPPP metric in (2):

$$xPPP = \Big(Pr(Penetrative) \times V(Penetrative)\Big) + \Big(Pr(Non - Penetrative) \times V(Non - Penetrative)\Big), \tag{2}$$

in which penetrative and non-penetrative passes are complementary events (i.e., $Pr(Penetrative) + Pr(Non - Penetrative) = 1$), and $Pr(Penetrative) = xPC$. V(Penetrative) and V(Non-Penetrative) stand for the value created by performing a pass, assuming to be penetrative or non-penetrative, respectively. To quantify such values, we employ a slightly modified version of the well-known VAEP framework [3] as follows: considering the features of the current pass and the previous two events, we calculate the probability of goal scoring or conceding within the next 10 actions. However, in our setup we set the look-ahead to the next 5 actions as potential penetrative passes are usually in the attacking or established possession phases and might soon result in a goal scoring or conceding. Note that we also differentiate between penetrative and non-penetrative pass actions. Thus, the VAEP framework outputs different values according to the outcome of the actions. Finally, the VAEP framework [3] uses the following equation (3) to measure the value of each pass with the respective penetration outcome in our setup:

$$Value(p_i, x|z) = \Delta P_{scores}(p_i, x|z) - \Delta P_{concedes}(p_i, x|z), \tag{3}$$

where p_i stands for a pass action, x is the set of features of each action, and z is the penetration outcome of the pass (i.e., penetrative or non-penetrative), ΔP_{scores} and $\Delta P_{concedes}$ stand for the change in probability of scoring and conceding a goal within next 5 actions after performing a pass, respectively.

4 Experiments and Results

In this section, we first conduct experiments with both conventional machine learning and deep learning models to find the best prediction results for each component of our decomposed model. We then use the prediction results of the winner model for use cases such as teams' performance in terms of penetrative pass completion and value creation.

4.1 Best Performing Prediction Model

In order to get the best result of prediction for each component of our decomposed model, we have trained several machine learning algorithms in the line of VAEP framework [3], setting 3 actions as the game state. For the latter two probabilities, we also included the pass being penetrative or not on top of the VAEP features. The rest of the features are listed in Table 2 for each of the actions in the game state.

The models' parameters were tuned using RandomizedSearchCV, and Pearson correlation was used to select the features (threshold was 0.8), and the highly correlated features were dropped. We chronologically split the dataset into 80% of train, 10% of validation for hyperparameter tuning and model selection, and the remaining 10% as hold-out data for testing. Table 1 shows the evaluation on test set using two classification metrics: ROC AUC, and log-loss (binary cross-entropy).

According to Table 1, the CNN-LSTM model (c.f., Figure 2) outperforms all other models since it can capture both the spatial (through convolutions layers) and temporal (through recurrent layers) nature of the dataset. For the training process, we used batch size of 64, on Google Colab using Keras sequential models after 10 epochs. Therefore, we continue the analysis of the results estimated by this model.

4.2 Does a Penetrative Pass Affect Goal Scoring or Conceding?

We aim to investigate the effect of penetrative and non-penetrative passes on the probability of goal scoring and conceding within the next 5 actions of the

Table 1: Conventional machine learning and deep learning models evaluation on the test set

Model	xPC		Pr(Scoring)		Pr(Conceding)	
	AUC	Loss	AUC	Loss	AUC	Loss
Logistic Regression	0.75	0.45	0.71	0.47	0.66	0.88
XGBoost	0.77	0.41	0.76	0.42	0.71	0.44
Perceptron	0.79	0.40	0.76	0.41	0.74	0.42
CNN	0.81	0.39	0.79	0.40	0.77	0.42
LSTM	0.83	0.35	0.81	0.39	0.80	0.40
LSTM with dropout	0.84	0.35	0.81	0.39	0.81	0.39
CNN-LSTM	0.88	0.32	0.85	0.35	0.84	0.36

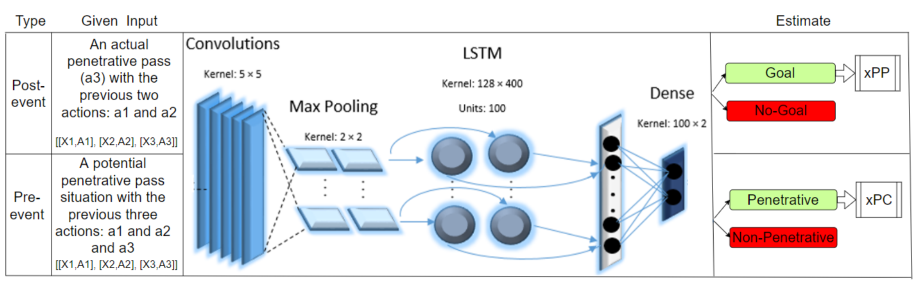
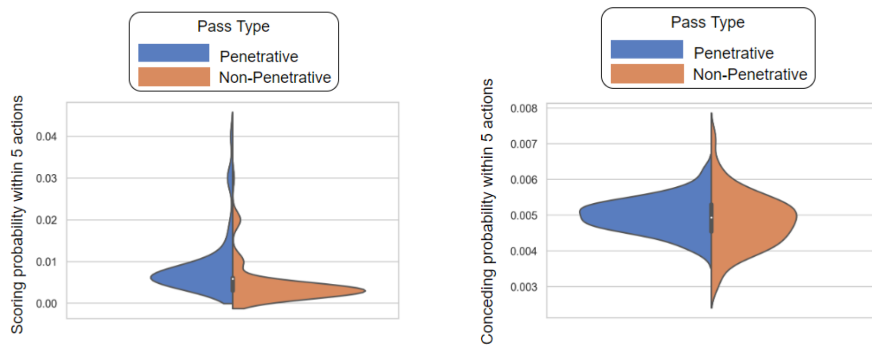


Fig. 2: CNN-LSTM network architecture. Input game states represent both features vector (X), and one-hot vector of actions (A).

Table 2: The game state features list according to their usage for each of the pre-event metrics (e.g., xPC), and post-event metrics (e.g., goal scoring and conceding probabilities from a penetrative pass). We order players using the permutation invariant sorting scheme proposed in [7] in which the ball holder is selected as the anchor in the first position, and the rest of the players are numbered according to their distances to the anchor.

State feature name	Description	Pre-event?	Post-event
Teammates location	(x,y) location of teammates in the visible area	✓	✓
Opponents location	(x,y) location of opponents in the visible area	✓	✓
Team ID	action is performed by a teammate or opponent?	✓	✓
Distance to goal	Euclidean distance from action location to center of the goal line	✓	✓
Max angle of view	maximum angle that is created by the ball with any two adjacent opponent players	✓	✓
Max distance to opponents	maximum Euclidean distance between adjacent opponent players in front of the ball holder	✓	✓
Min distance to teammates	minimum distance between the ball holder and any of the teammate players	✓	✓
Time remaining	time remained from action occurrence to the end of match half	✓	✓
Goal difference	actual difference between the expert team and opponent goals	✓	✓
Action result	successful or unsuccessful	-	✓
Body ID	is the action performed by head or body or foot?	-	✓
Height in case of a pass	ground, high, low	-	✓
Technique in case of a pass	in/out swinging, straight, through-ball	-	✓

current pass. The violin plots in Figure 3 depict the probability of goal scoring on left, and conceding in right, for all penetrative and non-penetrative passes of the UEFA EURO 2020 dataset. We can see that although the penetrative passes are boosting the probability of goal scoring in comparison to the non-penetrative passes in Figure 3a (with a median of 0.006 for penetrative and 0.003 for non-penetrative), the non-penetrative passes do not seem to have much effect on boosting the goal conceding probability in Figure 3b. That is because the non-penetrative passes are not necessarily turnovers and might keep the possession by a side-way pass to a teammate, for instance.



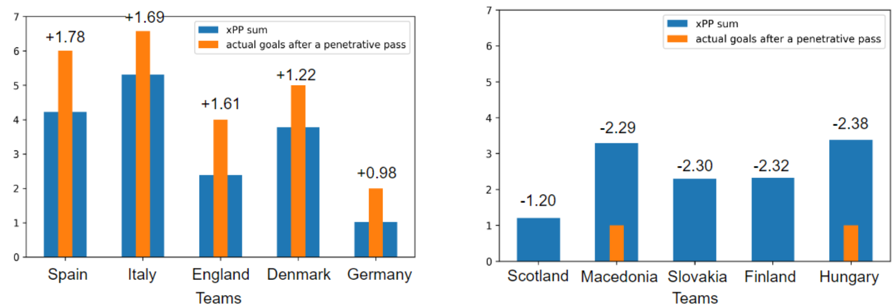
(a) Scoring a goal (b) Conceding a goal

Fig. 3: Probability of scoring or conceding a goal within next 5 actions

4.3 Teams' Penetrative Performance Analysis

In this section, we evaluate the teams participating in UEFA EURO 2020 in terms of the three introduced metrics: xPP, xPC, and xPPP.

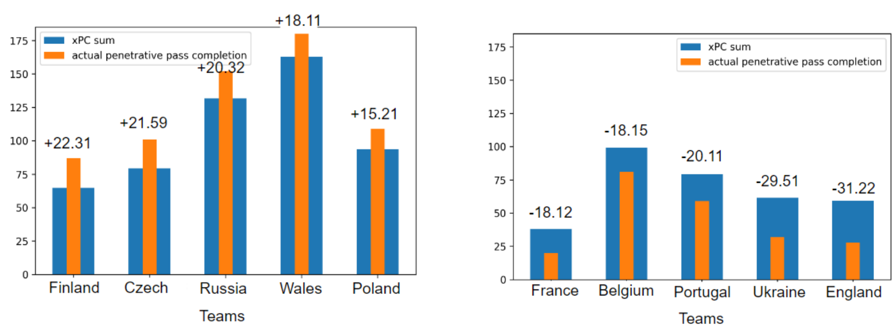
xPP: In order to evaluate how teams are performing in terms of converting a penetrative pass to a goal within the next 5 actions, we sum up their xPP through all their actual penetrative passes. We call this metric as xPP sum, and interpret it as follows: we expect the team to score xPP sum goals on average of all penetrative passes in the games. Note that this value cannot be used for team evaluation (i.e., the higher xPP sum does not imply the better performance of the team). However, we can compare this value with the actual number of goals the team scored within the next 5 actions of a penetrative pass. Now the difference between actual goals and expectations implies the team performance, in which a positive xPP-difference indicates overperformance of the team to score a goal within the next 5 actions of a penetrative pass, and a negative xPP-difference indicates underperformance. Figure 4 illustrates the 5 best and 5 worst teams in terms of xPP-difference on left and right, respectively.



(a) Best teams (b) Worst teams

Fig. 4: Teams' performance in converting a penetrative pass to a goal. Positive and negative numbers above the bars show the xPP-difference for each team.

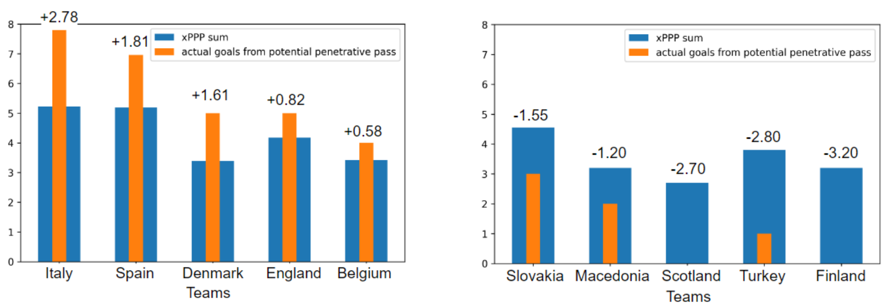
xPC: Similarly, we evaluate the teams in terms of xPC (i.e., likelihood of successfully converting a potential penetrative pass to an actual penetrative pass). For this aim, we sum up all xPC of the teams through their potential penetrative pass situations and call it as xPC sum. We interpret an xPC sum for a team as follows: we expect the team to complete xPC sum penetrative passes on average of all potential penetrative passes in the games. We then compare this value with the actual number of completions for each team for evaluation purposes. The result of the 5 best and 5 worst teams in terms of xPC-difference is illustrated in Figure 5.



(a) Best teams (b) Worst teams

Fig. 5: Teams' performance in completing a potential penetrative pass. Positive and negative numbers above the bars show the xPC-difference for each team.

xPPP: Finally, we evaluate the teams in terms of converting a potential penetrative pass into scoring/conceding situations. To do so, we sum up all xPPP of the teams through all their potential penetrative passes in the games and call it xPPP sum. We interpret xPPP sum as follows: we expect the team to score/concede xPPP sum goals on average of all potential penetrative passes it had in the games. Now we can compare it with the actual number of goals they scored within the next 5 actions of all potential penetrative pass situations, and evaluate them according to their xPPP-difference. The result of the 5 best and 5 worst teams in terms of xPP-difference is illustrated in Figure 6. While comparing teams ranking in terms of xPC in Figure 5 and xPPP in Figure 6, we observe that England and Belgium are among the worst teams in xPC (i.e., they are not good at completing a penetrative pass), but among best teams in xPPP (i.e., they are good at scoring goals from a potential penetrative pass situation).



(a) Best teams

(b) Worst teams

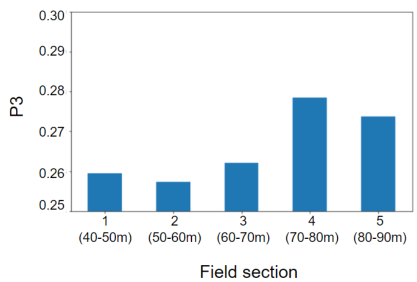
Fig. 6: Teams' performance in converting a potential penetrative pass to goal. Positive and negative numbers above the bars show the xPPP-difference for each team.

4.4 Field Section Analysis:

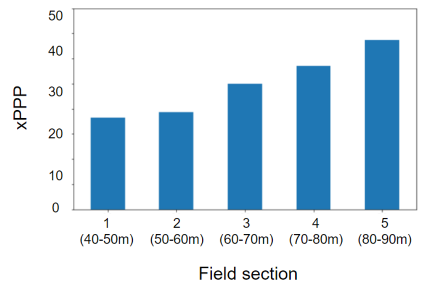
We also analyzed the impact of starting location of a potential penetrative pass to be a successful penetrative pass and how much we expect a potential penetrative pass to create a goal situation if it has been started in each field section along the touchline. To do so, we compute the $P3$ metric [13] as:

$$P3 = \frac{\text{number of successful penetrative passes}}{\text{number of potential penetrative passes}}$$

for each field section. Figure 7 illustrates the xPPP sum and the P3 for each one of the 5 sections of the field touchline (from 40 to 90 meters, each section stretching 10 meters). It can be noticed from Figure 7a that a potential penetrative pass becomes successful more often in section 4 and less often in section 2. However, considering the xPPP sum graph in Figure 7b we can observe that more goals are expected to be scored from a potential penetrative pass situation as we move forward in the field towards the opponent's goal.



(a) P3% by Field touchline sections



(b) xPPP sum by Field touchline sections

Fig. 7: Field section analysis on penetrative pass performance

5 Conclusion

Penetrative pass is an important tactic in offensive games. Being able to perform penetrative passes can lead the team in possession of the ball closer to scoring a goal. We proposed a possible application of event and positional data to measure teams' performance in terms of penetrative passes. For the best five and worst five teams in the dataset, we provide a trade-off between the ability to complete penetrative passes, and the expected value of goals they can create by doing so. Besides, we showed the impact of starting location of a potential penetrative pass to be a successful penetrative pass and how much we expect a potential penetrative pass to create a goal situation if it has been started in each field section along the touchline. However, the current study still lacks the full context of the game state due to the absence of some players' locations in the visible area of StatsBomb360 dataset. As a future direction, we aim to improve the accuracy of prediction by developing more sophisticated deep learning techniques such as graph attention networks, which could better capture the interaction between players, and use the results to infer what could lead a player to perform a penetrative pass and what is its impact on goal scoring.

Acknowledgment

Project no. 128233 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the FK_18 funding scheme. We thank Hadi Sotudeh for his valuable comments that helped improve this paper.

References

1. How impactful are line-breaking passes? `https://bit.ly/39uQe6Q`

2. Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., Matthews, I.: Large-scale analysis of soccer matches using spatiotemporal tracking data. In: IEEE International Conference on Data Mining (ICDM) (2021)
3. Decroos, T., Bransen, L., Van Haaren, J., Davis, J.: Actions speak louder than goals: Valuing player actions in soccer. In: In ACM KDD (2019)
4. Fernandez, J., Born, L.: Soccermap: A deep learning architecture for visually-interpretable analysis in soccer. In: ECML PKDD (2020)
5. Fernandez, J., Bornn, L., Cervone, D.: Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In: 13th MIT Sloan Sports Analytics Conference (2019)
6. Gyarmati, L., Stanojevic, R.: A merit-based evaluation of soccer passes. In: ACM KDD Workshop on Large-Scale Sports Analytics (2016)
7. Mehrasa, N., Zhong, Y., Tung, F., Bornn, L., Mori, G.: Deep learning of player trajectory representations for team activity analysis. In: 12th MIT Sloan Sports Analytics Conference (2018)
8. Peralta Alguacil, F., Fernandez, J., Piñones Arce, P., Sumpter, D.: Seeing in to the future: using self-propelled particle models to aid player decision-making in soccer. In: In Proceedings of the 14th MIT Sloan Sports Analytics Conference (2020)
9. Power, P., Ruiz, H., Wei, X., Lucey, P.: Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer. In: ACM KDD (2017)
10. Rahimian, P., Oroojlooy, A., Toka, L.: Towards optimized actions in critical situations of soccer games with deep reinforcement learning. In: IEEE DSAA (2021)
11. Rahimian, P., Van Haaren, J., Abzhanova, T., Toka, L.: Beyond action valuation: A deep reinforcement learning framework for optimizing player decisions in soccer. In: 16th MIT Sloan Sports Analytics Conference (2022)
12. Reina, R., Raabeb, D., Memmert, D.: Which pass is better? Novel approaches to assess passing effectiveness in elite soccer. Human Movement Science **55**, 172–181 (2017)
13. Sotude, H.: Potential penetrative pass (P3). In: STASBOMB Conference (2021)