Is the Midfielder a Key Player for the Matchup?

"If you have control of the midfield, you have control of the game, and you have more chances to win" - Xabi Alonso

1 Background

Soccer is a sport involving interactions between teammates and opponents. The cooperative and competitive actions displayed by the players yield a reliance on data analysis to improve team goals. Within this framework, each soccer player takes on a position based on their tactical formation for each matchup. Midfielders act as the engine of the team, connecting offense and defense. When they participate in the offensive build-up with their teammates, their exceptional ball control enables them to escape intense pressure, deliver decisive passes to teammates, and create shooting opportunities to score goals.

2 Goal

Motivated by this background and the theme of this competition, we focus on midfielders and their efficiency when on the opponent's side of the field. The follow-up question is:

How can we quantify the efficiency of midfielders when attacking on the opponent's side of the field?

Toward this goal, leveraging women's soccer tracking and event data from SkillCorner and statistical models, we introduce real-time ball possession probability for each player and use this to measure the efficiency of a midfielder's ball control, shot selection, short-distance passes, and long-distance passes. We will predict the proposed probability of ball possession based on the level of pressure on the ball. The level of pressure will be defined using the distances between the ball and players and/or between offensive and defensive players, calculated similarly to a gravity formula. As the distance between the defender and the ball carrier decreases, the level of pressure will increase, impacting the expected ball possession probability as well as the efficiency among all the metrics.

3 Methods

3.1 Proposed metrics

First, we introduce real-time ball possession probability $P_{i,t}$ for the i^{th} player at time point t in a matchup. As shown in Figure 1, the basic concept of $P_{i,t}$ is predicted based on the distances between the ball and the offensive and defensive players. For example, $P_{i,t}$ represents a low ball possession probability when the midfielder possessing the ball is surrounded by multiple defenders at a short distance. Therefore, $P_{i,t}$ can be considered as an indicator of real-time pressure on the ball. We believe that midfielders achieve the best results when making decisions such as passing or shooting when their ball possession probability is highest, i.e., when they handle the ball with less pressure. Based on the proposed ball possession probabilities for all players $P_{1,t}, P_{2,t}, ..., P_{22,t}$ on the pitch at each moment in the game, we propose a new set of metrics to quantify the efficiency of on-ball play for midfielders when attacking the opponent's side of the field. To this end, we specify intervals to indicate on-ball play (i.e.,'on-ball play area' in Figure 1) by defining 'Start of on-ball play' and 'End of on-ball play' (not explained here due to page constraints). Then, we focus on the real-time ball possession probabilities for the player within the 'on-ball play area' to calculate the following proposed metrics,

1. **Ball Control EFF:** A function of the average ball possession probability within the 'on-ball play area' and the ball possession probability at the 'End of on-ball play',
2. **Shot Selection EFF:** A function of the ball possession probability at the 'End of on-ball play' and the probability of a shot on target,
3. **Short-distance Pass EFF:** A function of the ball possession probability at the 'End of on-ball play' and the ball possession probability for a teammate receiving a short-distance pass, defined as a pass shorter than a quarter of the field,
4. **Long-distance Pass EFF:** A function of the ball possession probability at the 'End of on-ball play' and the ball possession probability for a teammate receiving a long-distance pass, defined as the pass longer than a quarter of the field.

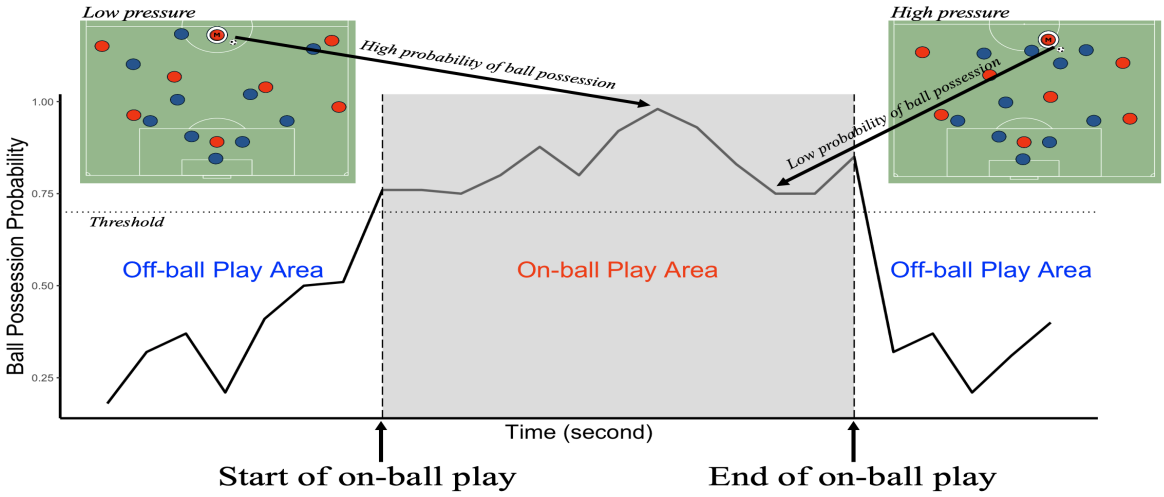

Figure 1: Example of a player's real-time ball possession probability.

These metrics provide deeper insights into how midfielders influence each play by utilizing distances between the ball and players, or among all players on the pitch, to offer information about pressure. The metrics can be measured in real-time to track the player's performance during the matchup, such as first-half ball control EFF and second-half short-distance pass EFF.

3.2 Data Analysis and timeline

A crucial first step is to predict the real-time ball possession probabilities for all players at each moment during the matchup. We will generate variables such as the speed of the ball and players, the angle of attack, the level of pressure based on distances between the ball and players, and the distances among all players on the pitch. These variables will be considered as model inputs, with a ball possession-related variable as the output for machine learning algorithms and statistical models such as random forest and XGBoost, targeting a multiclass (22 classes representing players) classification. The dataset will be split into training and testing sets, with the training dataset used for model development and the testing dataset for model validation. Next, we will identify the formulas for the four proposed metrics to evaluate the efficiency of play based on the pressure. For example, let $P^{on}_{10,t}$ be the ball possession probability for a midfielder at the end of on-ball play, and $P_{11,t}$ be the ball possession probability for a teammate who will receive the pass. Then, a possible formula of short-distance Pass EFF for the midfielder can be calculated as $P^{on}_{10,t} \times P_{11,t}$. If the player within the on-ball play passes to a teammate under less pressure and/or the teammate receives the pass under less pressure, the resulting EFF measure would be higher, indicating high short-distance Pass EFF for the midfielder. We will propose optimal combinations for the four proposed metrics using real-time ball possession probabilities for all players. Table 1 presents the timeline for data analysis in this project.

Table 1: Data analysis timeline

	December				January			
	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4
Data preprocessing	✓	✓						
Exploratory analysis		✓	✓	✓				
Statistical modeling/Model validation			✓	✓	✓	✓	✓	
Interpretation/Presentation					✓	✓	✓	✓

4 Expected Outcomes and Contribution

What makes our template unique, is that it provides both teams with data on their midfielders and opponents. This data will aid teams in future games, evaluating talent, and designing new formations to create the best opportunities possible. We assume that successful passes, shots, or dribbles under high pressure are more effective, while failed attempts under high pressure are less effective, as better options are typically available in such situations. Therefore, this template is a universal data-driven efficiency ranking of midfielders in Women's Professional Soccer based on the data provided by SkillCorner.