**Introduction**

Soccer is a team sport that include multiple roles for players such as goalkeeper, defender, midfielder, and striker. Each position also requires different physical abilities and skills to perform their duties effectively. For example, a striker is responsible for scoring goals and requires better attributes on pace, shooting, and attacking positioning. On the other hand, a defender needs to stop opposing teams' attacks and requires attributes such as tackling, marking, and interceptions. Seeking ways to gain a competitive edge, the use of data analytic techniques to fit and predict a player’s possible performances is an emerging trend in soccer. For instance, the German national team utilized data analytics to inform their training and match strategies, which led to their success in the 2014 World Cup.

In this research, the classification of players into strikers or defenders based on their physical attributes and skills is performed, which is crucial in player recruitment, team formation planning, and match strategy development. For instance, a team may use this analysis to recruit a striker who fits their attacking style and has a high probability of scoring goals. Additionally, coaches can use this analysis to develop a formation that maximizes the strengths of their players and neutralizes the opposition's strengths. Therefore, we need a quick way to judge whether a player should be placed in an attacking position or at a defending position on the field given some physical data such as max speed, passing ability, dribbling skills, and pace that could be quantified and measured for each player. Also, in modern soccer, the role of whether a defending player should act as a traditional back (LB / RB) or as a modern wing back (LWB / RWB), which would require more attacking abilities, is getting more important, as teams are seeking a way to break through the traditional tactics. This research also analyzes on how should we determine whether a player is better as back or wing back in this game, given the physical quantified data.

**Statistical Method**

1. Data description

The original dataset ([players\_20.csv](https://www.kaggle.com/datasets/stefanoleone992/fifa-20-complete-player-dataset?resource=download&select=players_20.csv)[[1]](#footnote-1)) is obtained from Kaggle. This dataset contains FIFA20 player data. We pre-process the dataset with python to only obtain the data columns that we want to use in this final project and removed all data of Goalkeepers, which has a different record form than other players. The updated dataset [players\_20\_edited.csv](https://uofi.box.com/s/91m9qt18v7zhxh00d18edv7fx6x95fsw)[[2]](#footnote-2) contains 16242 non-goalkeeper soccer players and their physical data, scores on skills, and positions.

2. Classification Method

We used Support Vector Machines (SVM) as the main statistical method to classify whether a player is better to play as a striker or as a defender on the field. SVM is a well-established method for classification, which has been used in many fields, including image recognition, bioinformatics, and finance (Cortes & Vapnik, 1995; Cristianini & Shawe-Taylor, 2000). We selected the SVM method because it can effectively deal with high-dimensional data, and it is capable of finding nonlinear decision boundaries (Hsu & Lin, 2002). To improve the classification performance, we also used ensemble methods such as bagging and boosting, which combine multiple SVM models to make a final prediction (Bauer & Kohavi, 1999). In addition, we used multi-resolution tree kernels to extract features from the player positions data, which has been shown to be effective in sentiment analysis (Wang, Tang, & Liu, 2009).

In the two classifications, we applied simple Linear SVM, tuned Linear SVM, Radial Kernel SVM, and Polynomial Kernel SVM and compares in their effects of classification for the training data and k-fold cross validation is applied during training to prevent overfitting of the model and provide a more accurate estimate of the model's performance on unseen data. (Hastie, Tibshirani, & Friedman, 2009)

**Results**

(1) Classification – Striker vs Defender

(2) Classification – Back vs Wing Back

**Discussion**

(1) Classification – Striker vs Defender

(2) Classification – Back vs Wing Back

1. The original dataset players\_20.csv could also be found at https://www.kaggle.com/datasets/stefanoleone992/fifa-20-complete-player-dataset?resource=download&select=players\_20.csv if the link fails. [↑](#footnote-ref-1)
2. The updated dataset players\_20\_edited.csv could also be found at https://uofi.box.com/s/91m9qt18v7zhxh00d18edv7fx6x95fsw if the link fails. [↑](#footnote-ref-2)