Group 5 Project Proposal
CS6220 - Big Data Systems and Analytics
Tusheet Goli, Tejas Pradeep, Aman Jain, Akshay Pramod, Jeffrey Chang

# Machine Learning to Analyze Player and Team Performance in Soccer

## Introduction

The project aims to create a soccer team performance evaluator. The project evaluates team performance using more holistic data based on match stats and player stats rather than simple metrics such as nationality or club. The project plan is to use various Machine Learning models and techniques such as Graph Neural Networks and Ensemble Learning to evaluate a team of 11 soccer players. Games such as EA Sports' FIFA use simple metrics such as player's nationality and league to analyze team chemistry and team scores. The goal of the project is to use basic metrics as a foundation to train models to use more complex data to make better predictions about team chemistry and team score.

## Motivation and Objectives

The main motivation for this project comes from the shortcomings of traditional simpler models to analyze soccer team chemistry. Approaches used by popular games such as FIFA are very rudimentary and use simple data such as player nationality and league. Such an approach leads to a very shallow analysis of team chemistry and by extension team performance as player match data and particular player attributes are not taken into consideration. The motivation behind the project is to improve on these shortcomings of games like FIFA and introduce a more future-proof data-based model. Current-day gaming consoles and PCs have gotten strong enough to run complex prediction algorithms with ease, and hence having such a model is very possible.

The main goal of the project is to build a more complex framework to analyze player and team chemistry and performance using collected match data from millions of soccer matches and to use existing soccer Transfermarkt data to evaluate the worth and potential impact of a player. The stretch goal for this project is to develop a framework that could potentially be used by soccer managers to predict the effect a new player would have on a team using quantitative analysis and a sophisticated ML model. The project shall aim to develop the model and evaluate it based on historic records of good teams vs bad teams and good signings vs bad signings. The model can be further retrained to stay consistent with these historical records. The goal is to repeat the process of training, evaluating, and retraining still we arrive at a model with consistently accurate predictions.

## Related Works

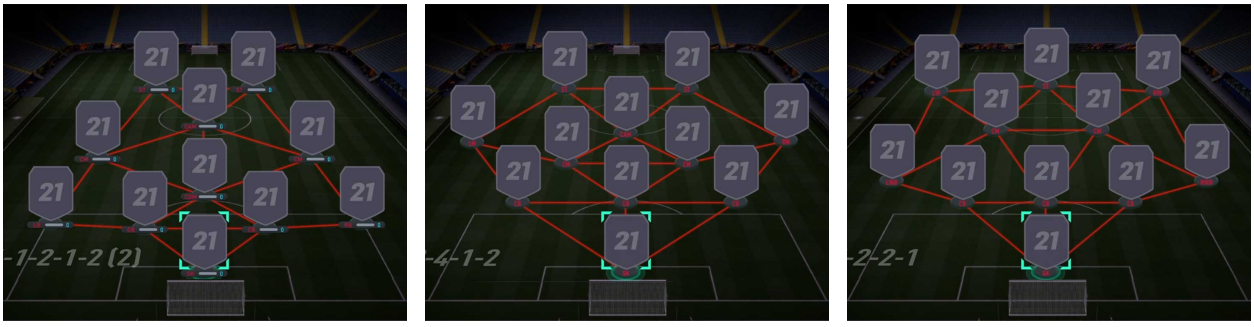
We have researched a lot of different frameworks and models to help establish a meaningful representation of team chemistry and performance in soccer. Our foundational framework comes from paper [1] that aimed to solve this problem using a simple Graph Neural

Network approach (GNN). Using layers of GNN architectures like GraphConv, GCNConv, and GATConv, the models were trained on the starting 11 for over 400+ clubs. There are also papers that utilized a PlayerRank and TeamRank framework [2] to separate individual attributes and team-based performance to assess the overall performance fit of the additional player in the team. These individual attributes can be conglomerated into meaningful scores using an ML-based ranking algorithm proposed by Kumar in paper [3].

In addition to in-game and physical attributes that constitute performance metrics [4], we would also like to integrate a psychological viewpoint regarding best-fit using a multitude of metrics like communication skills, player attitude, loyalty to the team, passion/motivation, etc. as pointed in Gershgoren's and Zepp's papers [5, 6]. These are some of the key aspects lacking from paper [1] that we would like to better capture through our newer models and training techniques to get a better estimate of player and team performance. We have reviewed prior research in Graph Neural Networks (GNNs) and Graph Attention Networks (GATs) and how we can incorporate the self-attention and neighborhood features of these networks that can be incorporated into the edge and link (chemistry) predictions of the models to help assess player-to-player links and fit [7, 8, 9].

**Proposed Work**

We plan on improving the model from the existing work by using some of the techniques we have learned in class. The existing work arbitrarily chose a graph structure for the GNN model based on the most common soccer formation, a 1-4-3-3 formation. However, there is room to make the model more robust by trying different underlying graph structures based on less common formations such as a 1-4-1-2-1-2, or 3-4-3. Introducing a variety of underlying graph structures will give us a more holistic view of how players interact with each other. Thus, we are lifting our search space from a one-dimensional space of players only, to a two-dimensional space of players and formations. From these results, we can then gain more nuanced and robust insight into team strengths.



*Examples of different soccer formations and underlying graph structures.*
*In order from left to right: 1-4-1-2-1-2, 1-3-4-1-2, 1-5-2-2-1*

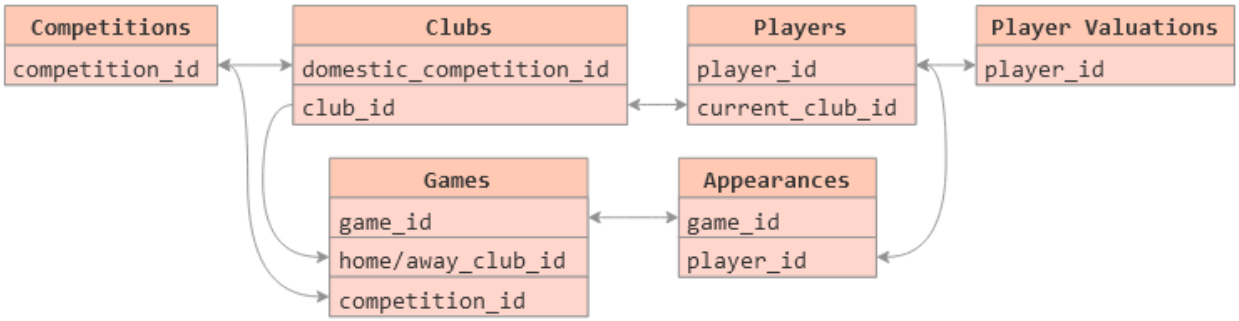
By running our Graph Neural Network model on tens, or even hundreds, of different valid formations, we can use Ensemble learning to then evaluate team strength more accurately.

The underlying motivation behind evaluating a team's strength based on different formations is reflective of real life: different teams use different formations in different situations to leverage their players' relative strengths and hide their relative weaknesses compared to their opponents.

The purpose of Ensemble learning is to provide a concrete way to synthesize the data from all the different GNN results derived from the different formations. We want to identify the formations that are the most indicative of real-world results, as some formations are far more common in real life (such as the 1-4-3-3 formation) compared to others. Thus, when evaluating a team, we want to give weighted priority to the formations that are more common than the ones that are less common. To figure out the weights assigned to a formation, we will use ensemble learning to compare the input space of formations to the output space of actual ranking.

**Dataset**

The dataset we are planning on using for this project is the Football Transfermarkt Dataset (https://data.world/dcereijo/player-scores). This is an up-to-date dataset obtained from the real-time FIFA website. This dataset has data for 55,000+ games, 400+ clubs, 20,000+ players, 1,000,000+ player appearance records, and 300,000+ player market valuations across the world. This dataset is in the form of 5 independent csv files (players, clubs, games, competitions, appearances). We plan on using the various attributes in this dataset to help us with the task of assessing player/team performance.




**Evaluation and Testing Methods**

We can evaluate our hyperparameter tuning of our GNN based on whether or not we see a performance improvement based on our original model. The main difficulty behind evaluating our model and defining the "performance improvement" is that we do not have the ability to actually move players from one team to another and measure how well they work together or how many games they win now. We can define a performance improvement based on how well the team chemistry conforms to a FIFA Ultimate Team Builder model. We know that this is a very simplistic metric to use (league/nationality) but it's a good starting point.

Another potential way we can measure the success of the model is by using historical team data and winning records, under the assumption that greater team chemistry is associated with a better team record. Suppose team A has a 4-10 record in 2019 and an 8-6 record in 2020 with better team chemistry in their starting 11 in 2020 than in 2019. Then we could gauge this as

a “success” for our model, performing across various starting 11s across different seasons. Similarly, we can also use data from real games to identify whether our model is performing well or not. For example, if Team A beat Team B, then our model should evaluate Team A’s starting 11 more strongly than Team B’s starting 11. By having this baseline of which teams win against which other teams, we can evaluate our model’s performance based on how accurately the model is able to evaluate a higher performance of the winning team.

The method we will be using to test our GNN will be similar to the evaluation. We face similar difficulties as evaluation, but we can attempt to use similar solutions. We will test how well the teams we predict perform based on the FIFA Ultimate Team Builder models compared to our own predictions. Their model is simplified compared to what we want to find but still has some accuracy. We will also be splitting up our data between different years and different games to create a training dataset, as well as a testing dataset to see how well we can predict actual results. This will allow us to compare what we expect to happen when players move teams to what actually happens. The closer our prediction of team strength (evaluated based on game results) is to reality, the better our model is. Thus, we will measure this difference and use it as our test result.

**Plan of Activities**

The resources we plan on using are:

- Software:
    - GitHub (Gatech): We will use Gatech GitHub as our primary source control.
    - Jupyter Notebook or Google Colab: This will be used for creating our ML models and running our code.
    - MySQL: Since the original dataset is a collection of CSV files, we plan on using MySQL to convert this to a .sqlite file to make data selection and processing fast.
    - Python Flask Server: We will use python Flask to host our frontend code.
- Hardware:
    - Heroku DB Hosting or AWS or GCP or DigitalOcean Server: We might use one of these services to host our .sqlite database that will help feed data to our ML training as well as render our frontend. We are leaning towards Heroku DB and we plan on opting into a paid plan to avoid getting rate limited.
- Programming Languages:
    - Backend:
        - Python: We plan on using Python for backend coding since we are hosting our application on Python Flask. We will also use this for our ML models which will be written on Juptyer Notebook or Google Colab.
        - SQL: We will use SQL to run queries on our compiled .sqlite database.
    - Frontend (Visualization):
        - ReactJS: Our team members have decent experience with ReactJS and we plan on building the entire frontend application and viz tool using this.

| Weeks | Dates | Tasks |
|---|---|---|
| Week 6 | Sept 26 - Sept 30 | **Project proposal document and submission** |
| Week 7 | Oct 1 - Oct 7 | Data collection and cleaning |
| Week 8 | Oct 8 - Oct 14 | Convert dataset to .sqlite database and host it on a cloud service (Heroku DB), setup python flask server application |
| Week 9 | Oct 15 - Oct 21 | Create novel team/player performance score/algorithm, create and train MVP ML models |
| Week 10 | Oct 22 - Oct 28 | Analyze the performance of models and tune weights and hyperparameters |
| Week 11 | Oct 29 - Nov 4 | Have MVP of frontend visualization tool |
| Week 12 | Nov 5 - Nov 11 | **Prepare for project workshop presentation** |
| Week 13 | Nov 12 - Nov 18 | Improve model to the final state (better accuracy), obtain results and analysis from experiment test runs, finetune team/player performance algorithm and other weights/hyperparams |
| Week 14 | Nov 19 - Nov 25 | Improve frontend visualization tool and bring it to its final state |
| Week 15 | Nov 26 - Dec 2 | **Project demo submission** |
| Week 16 | Dec 3 - Dec 6 | **Project final report document and code package submission** |

**GitHub Repository**

We created a private repository on Gatech GitHub to store our data, code, and analysis. Link - https://github.gatech.edu/tgoli3/gt-bds-f22-team5

## References

### Papers


[1] T. Goli, E. Gu, P. Khorana, S. Manjesh, T. Pradeep, A. Srinivas. “Analyzing Player and Team Chemistry/Performance in Soccer” (2022).

[2] Pappalardo, Luca, Paolo Cintia, Paolo Ferragina, Emanuele Massucco, Dino Pedreschi, and Fosca Giannotti. “PlayerRank: Data-Driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach: ACM Transactions on Intelligent Systems and Technology: Vol 10, No 5.” ACM Transactions on Intelligent Systems and Technology, September 1, 2019. https://dl.acm.org/doi/10.1145/3343172.

[3] Garnier, Paul, and Theophane Gregoir. “Papers with Code - Paper Tables with Annotated Results for Evaluating Soccer Player: From Live Camera to Deep Reinforcement Learning.” The latest in Machine Learning, January 2021. https://paperswithcode.com/paper/evaluating-soccer-player-from-live-camera-to/review/.

[4] S. Ghar, S. Patil, and V. Arunachalam, "Data-Driven football scouting assistance with simulated player performance extrapolation," 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), 2021, pp. 1160-1167, DOI: 10.1109/ICMLA52953.2021.00189.

[5] Gershgoren, Lael, Itay Basevitch, Aaron Gershgoren, Yaron S. Brill, Robert J. Schinke, and Gershon Tenenbaum. "Expertise in soccer teams: A thematic inquiry into the role of shared mental models within team chemistry." Psychology of Sport and Exercise 24 (2016): 128-139.

[6] Zepp C, Kleinert J. Symmetric and complementary fit based on prototypical attributes of soccer teams. Group Processes & Intergroup Relations. 2015;18(4):557-572. doi:10.1177/1368430214556701

[7] Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. "Graph attention networks." arXiv preprint arXiv:1710.10903 (2017)

[8] Zhang, Muhan, and Yixin Chen. "Link prediction based on graph neural networks." Advances in neural information processing systems 31 (2018).

[9] Lee, John Boaz, Ryan Rossi, and Xiangnan Kong. "Graph classification using structural attention." In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1666-1674. 2018.


### Datasets


“Football Data from Transfermarkt - Dataset by Dcereijo.” data.world, February 19, 2022. https://data.world/dcereijo/player-scores.