A Two-Stage Hybrid Framework for Player Position and Overall Rating Prediction in Soccer

Mehmet Gür
Artificial Intelligence Engineering
TOBB Economy and Technology University
Ankara, Turkey
mehmetgur@etu.edu.tr

***Abstract*—Player rating prediction is a significant task in modern football analytics, often relying on labeled position data and handcrafted features. In this study, we propose a two-stage framework that predicts players' positional roles and subsequently estimates their overall rating scores based on those roles. Using the publicly available European Soccer Database, we first extract players' average in-game coordinates from the *Match* table and apply K-Means clustering to uncover both general (GK, Defense, Midfield, Forward) and detailed (e.g., Side Midfield, Center Forward) positional groupings. We also develop a rule-based classification system using weighted player attributes to infer tactical roles.**

In the second stage, we train four regression models—Ridge Regression, XGBoost, Random Forest, and ElasticNet—within role-aware setups to predict overall ratings. Our best-performing configuration (M6-Ridge) achieves high R^2 and low RMSE values. Feature importance analysis and positional heatmaps reveal domain-aligned skill patterns, such as the importance of finishing for forwards and reflexes for goalkeepers.

Our results show that role-aware modeling significantly improves rating prediction and provides interpretable insights for scouting and performance evaluation systems.

***Index Terms*—sport analytics, player rating prediction, positional clustering, role-based modeling, ridge regression, feature importance, interpretable machine learning, model benchmarking**

I. INTRODUCTION

A. *Motivation*

Player rating is a widely used metric in football video games and analytical platforms to quantify the overall quality and performance potential of a player. These ratings often depend on a range of underlying player attributes, such as passing, shooting, dribbling, and defensive capabilities. However, a player's role on the field (e.g., forward, defender, goalkeeper) significantly influences which attributes are most critical for performance. For example, finishing is more relevant to forwards, while positioning and reflexes are crucial for goalkeepers.

Traditional rating systems typically assign a single rating formula across all player types or use handcrafted weightings that may not accurately reflect real-world positional requirements. This can overlookthe nuanced relationship between specific skills and overall performance, potentially leading to suboptimal evaluations in both entertainment and real-life analytics contexts.

In this study, we propose a two-stage hybrid framework that first infers a player's field position and then predicts their overall rating in a role-specific context. The first stage utilizes both unsupervised clustering and rule-based heuristics to assign general and detailed position labels. In the second stage, we train multiple regression models—including Ridge Regression, Random Forest, XGBoost, and ElasticNet—within position-aware experimental setups to estimate the overall rating from player attributes.

This multi-model analysis allows us to compare predictive performance, interpretability, and alignment with domain knowledge across diverse regression techniques. To ensure interpretability, we visualize feature importance using both coefficient-based bar charts and position-wise skill heatmaps. This analysis not only validates expected domain knowledge (e.g., the importance of finishing for forwards), but also helps uncover subtle skill patterns that might be overlooked by heuristic-based or generalized evaluation systems.

B. *Literature Review*

The seminal "European Soccer Database" [1] has served as a cornerstone for numerous studies in football analytics. While many works focus on match outcome prediction or team-level dynamics, relatively few explore the linkage between a player's on-field role and their individual skill profile. Recent systematic reviews highlight the increasing use of machine learning in soccer, yet they also identify gaps in fine-grained, role-aware player modeling [2], [3].

Several studies have applied clustering algorithms, particularly K-Means, to infer player roles or archetypes from spatial and performance data [4]–[7]. Concurrently, the task of predicting player ratings—often using regression models—has also gained traction [8], [9], with frameworks such as adjusted plus-minus being adapted for football contexts [10], [11].

Our work uniquely bridges these two streams by proposing a two-stage hybrid methodology. In contrast to prior studies, we first compare spatial clustering and rule-based heuristics for position detection, and then use the resulting labels to build position-specific regression models for rating prediction. This hierarchical framework enables a more robust, interpretable, and context-sensitive analysis of player performance compared to existing approaches in the literature.

II. Dataset and Methodology

A. Data Source and Preprocessing

This study utilizes the publicly available *European Soccer Database* hosted on Kaggle [1], originally extracted from EA Sports FIFA video game logs between 2008 and 2016. The dataset is structured and includes multiple tables such as `Match`, `Player`, and `Player_Attributes`. Its hybrid nature—containing both structured numerical features (e.g., shooting, passing, marking) and semi-structured elements (e.g., categorical work rate)—necessitates a robust preprocessing workflow.In total, the dataset includes more than 40 player-specific numeric attributes, covering technical, mental, and physical skills.

Preprocessing and Feature Engineering Pipeline:

1) **Data Restructuring:** The `Match` table stores player coordinates in a wide format (e.g., `home_player_X1...X11`), which was converted into a long-format structure. This enabled per-player aggregation and analysis across matches.
2) **Coordinate Averaging:** For each player, average field coordinates `avg_X` and `avg_Y` were computed over all match appearances. These spatial statistics served as the foundation for clustering-based role inference.
3) **Temporal Filtering:** To capture players' most current form, only the latest record per player was retained from the `Player_Attributes` table.
4) **Data Integration:** The coordinate features were merged with player attributes using the `player_api_id` as a foreign key, resulting in a unified instance-per-player table.
5) **Missing Data Handling and Cleaning:**
   - *Categorical Imputation:* Missing values in attributes like `attacking_work_rate` were filled with the modal class ("medium").
   - *Normalization of Categorical Strings:* Text variables were lowercased, and values like "none" were replaced with "medium" for semantic alignment.
   - *Numerical Cleansing:* Any records with missing numeric attributes were dropped using `dropna()` to ensure model compatibility.
6) **Attribute Type Annotation:**
   - All technical skill scores (e.g., `finishing`, `marking`) were treated as *interval* variables.
   - Categorical features (e.g., `preferred_foot`, `work_rate`) were considered *nominal.*
   - Player coordinate features were regarded as *ratio* variables.
7) **Feature Normalization:** All numerical features were normalized using `MinMaxScaler` to ensure consistent model behavior across different regression techniques.

B. Methodology Overview

Our proposed framework follows a **two-stage hybrid pipeline**.

1) Stage 1: Position Classification: We assign positional labels using two fundamentally different strategies:

- **Clustering-Based (Models M1–M4):** Using K-Means clustering on average player coordinates, a two-step process was applied:
  1) Y-axis clustering into 4 general roles: Goalkeeper (GK), Defender (DEF), Midfielder (MID), Forward (FWD).
  2) X-axis splitting to further distinguish central vs. side variants—yielding 7 detailed positional roles.
- **Rule-Based (Models M5–M8):** We developed hand-crafted scoring functions using weighted combinations of role-relevant features. The functions differ for general and detailed roles and were manually tuned using descriptive statistics from the dataset.

2) Stage 2: Position-Aware Performance Prediction: In this stage, we predict the `overall_rating` of players using the positional labels from Stage 1. Four regression models were implemented: Ridge, Random Forest, XGBoost, and ElasticNet. Rather than using PCA or ICA for dimensionality reduction, we leveraged *domain-driven feature subsets* per role type. For instance:

- Goalkeepers used only `gk_*` features.
- Midfielders included passing, ball control, and vision.
- Forwards emphasized finishing, shot power, and positioning.

Model Matrix Design: We trained eight configurations (M1–M8) as shown in Table I, representing combinations of:

- **Position Labeling Method:** Clustering-based vs. Rule-based
- **Role Granularity:** General (4 roles) vs. Detailed (7 roles)
- **Feature Scope:** Role-specific features vs. All features

TABLE I: Eight Experimental Model Configurations (M1–M8)

Model ID	Labeling Method	Label Type	Feature Set
M1	Clustering-Based	General (4 roles)	Role-Specific Features
M2	Clustering-Based	General (4 roles)	All Features
M3	Clustering-Based	Detailed (7 roles)	Role-Specific Features
M4	Clustering-Based	Detailed (7 roles)	All Features
M5	Rule-Based	General (4 roles)	Role-Specific Features
M6	Rule-Based	General (4 roles)	All Features
M7	Rule-Based	Detailed (7 roles)	Role-Specific Features
M8	Rule-Based	Detailed (7 roles)	All Features

III. Results

A. K-Means Clustering Visualization

To initiate our two-stage framework, we first performed K-Means clustering to identify spatial position groups based on players' average in-match coordinates. This approach enabled both general and detailed spatial categorization, which laid the foundation for our role-based analysis.

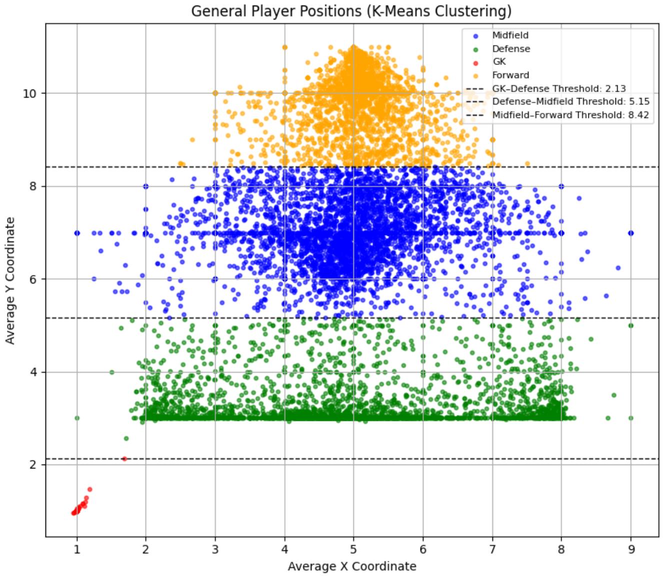

Fig. 1: General Player Positions based on K-Means clustering and Y-threshold segmentation.

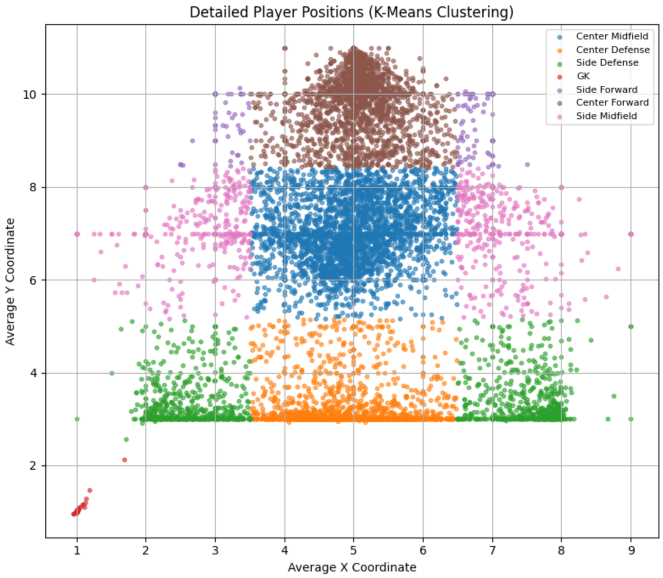

Fig. 2: Detailed Player Positions derived from K-Means clustering (seven-class spatial segmentation).

As shown in Figure 1 and 2, the clustering model achieves a strong and interpretable separation of player roles along the Y-axis of the pitch. The general segmentation captures 4 zones: Goalkeeper, Defense, Midfield, and Forward, while the detailed one further splits central and side areas into 7 distinct roles.

B. *Comparison of Positional Labeling Approaches*

To assess how two different position labeling approaches perform, we compared the general and detailed role distributions derived from KMeans-based spatial clustering and rule-based statistical heuristics. The following tables present the percentage distributions:

TABLE II: General Position Distribution (%) - KMeans vs Rule-Based

Position	KMeans	Rule-Based
Defense	33.04%	39.01%
Midfield	35.94%	34.97%
Forward	22.36%	17.31%
GK	8.67%	8.71%

TABLE III: Detailed Position Distribution (%) - KMeans vs Rule-Based

Position	KMeans	Rule-Based
Center Midfield	28.83%	11.61%
Center Defense	18.55%	24.61%
Center Forward	20.94%	12.18%
Side Midfield	7.11%	14.75%
Side Defense	14.49%	14.16%
Side Forward	1.42%	13.98%
GK	8.67%	8.71%

These discrepancies highlight fundamental differences between the two methodologies. The KMeans model tends to cluster players more heavily into central roles such as *Center Midfield* and *Center Forward*, likely due to dense spatial overlap in central pitch zones. In contrast, the rule-based model, which incorporates weighted player attributes, produces a more balanced distribution by assigning players more frequently to roles such as *Side Midfield* and *Side Forward*. This contrast indicates that spatial data alone may sometimes overlook lateral tactical roles.

C. *Agreement Rate Between Labeling Methods*

To quantify the consistency between the two labeling methods, we computed the agreement rates for both general and detailed positions:

- **General role matching accuracy:** 80.52%
- **Detailed role matching accuracy:** 56.16%

While general roles show a relatively high match rate, the notable drop in detailed role agreement suggests that the two methods diverge significantly in nuanced role identification. These findings imply that although KMeans provides useful spatial segmentation, rule-based labeling may better capture real-world functional roles—especially in tactical and attribute-driven contexts.

D. *Model Evaluation and Regression Results*

To evaluate the predictive performance of our models, we designed eight distinct configurations (M1–M8), combining two role-labeling strategies (KMeans vs Rule-Based) and four regression algorithms (Ridge, ElasticNet, XGBoost, Random Forest). Each model was assessed using average RMSE and R^2 scores, both for general roles (4 categories) and detailed roles (7 categories).

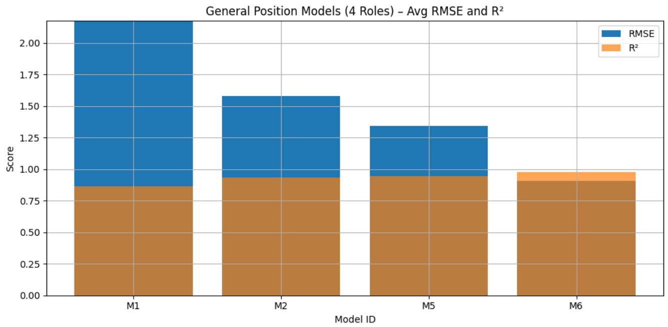

Fig. 3: General Position Models (4 Roles) – Avg RMSE and R^2

As seen in Figure 3, rule-based models (M5–M6) outperformed their KMeans-based counterparts (M1–M2) in general position predictions. Specifically, model M6 (Rule-Based + Ridge Regression) achieved the lowest RMSE and highest R^2, indicating superior alignment between features and overall rating when roles are accurately labeled through rule-based heuristics.

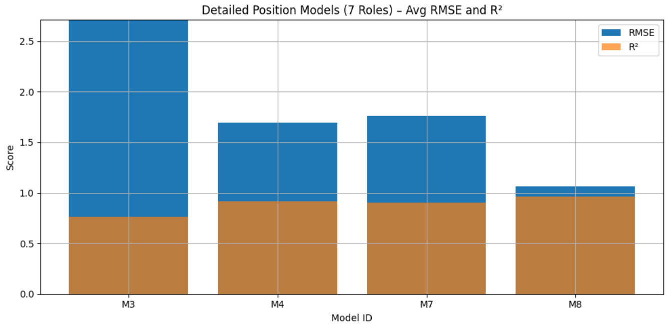

Fig. 4: Detailed Position Models (7 Roles) – Avg RMSE and R^2

For detailed roles, the performance gap became even more pronounced (Figure 4). Model M8 (Rule-Based + Ridge) maintained a high R^2 and low RMSE, while M3 (KMeans + Ridge) yielded the worst results, again confirming that noisy or incorrect position labeling deteriorates predictive accuracy.

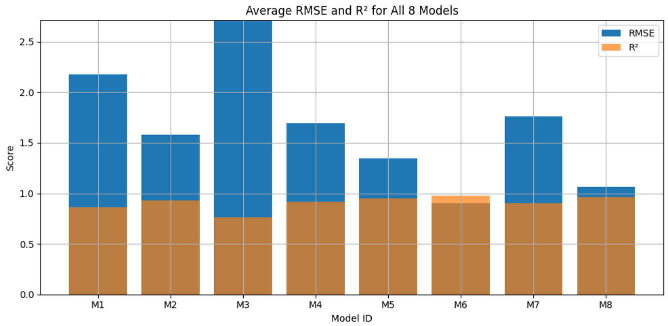

Fig. 5: Average RMSE and R^2 for All 8 Models

Figure 5 provides a comprehensive view of all models across both labeling strategies. A consistent pattern emerges: rule-based role labeling leads to stronger regression performance, regardless of the algorithm used. KMeans-based models not only exhibit higher RMSE but also reduced R^2 values, indicating weaker model fit.

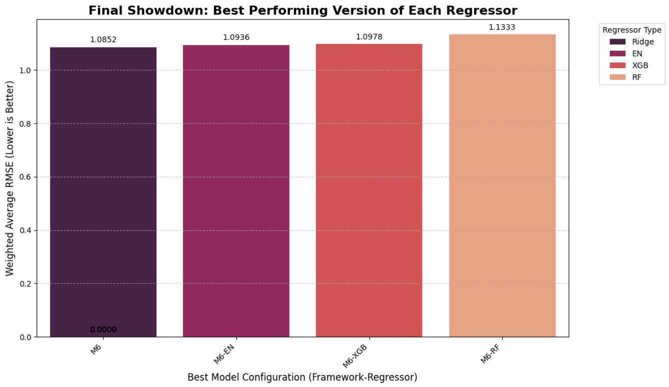

Fig. 6: Final Comparison: Best Performing Variant of Each Regressor

In the final analysis (Figure 6), we compared the best-performing version of each regressor based on weighted RMSE across all role types. Ridge regression with rule-based labeling (M6) emerged as the top performer (RMSE = 1.0852), narrowly outperforming ElasticNet and XGBoost variants. Random Forest lagged slightly behind, suggesting that linear models with regularization may better capture the positional nuances embedded in the player attributes.

E. Feature Importance Analysis

Understanding which skills most significantly contribute to player ratings across positions is critical for both performance modeling and talent evaluation. Using the coefficients from our best-performing model (M6 - Ridge), we analyzed the top influential features per role.

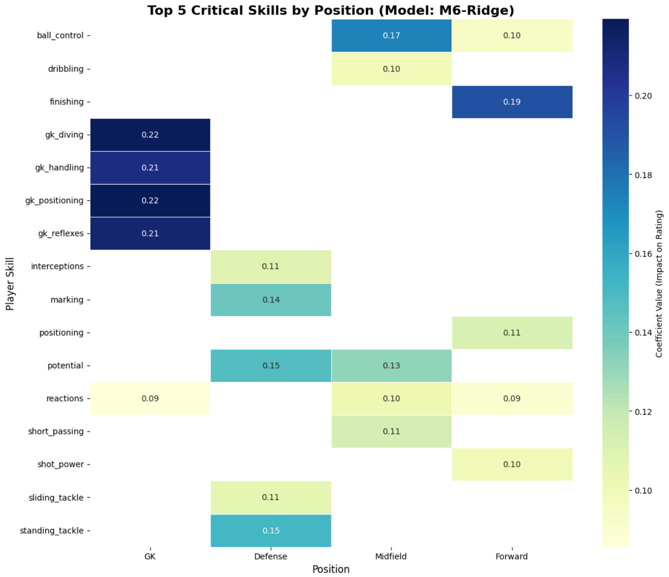

Fig. 7: Top 5 Critical Skills by Position (Model: M6-Ridge)

As shown in Figure 7, goalkeeper ratings are most strongly influenced by skills such as *gk_diving*, *gk_positioning*, and *gk_reflexes*, all with high coefficient values. Defensive roles emphasize *standing_tackle*, *marking*, and *potential*, while midfielders rely more on *ball_control* and *short_passing*. Forwards show significant dependence on *finishing* and *ball_control*.

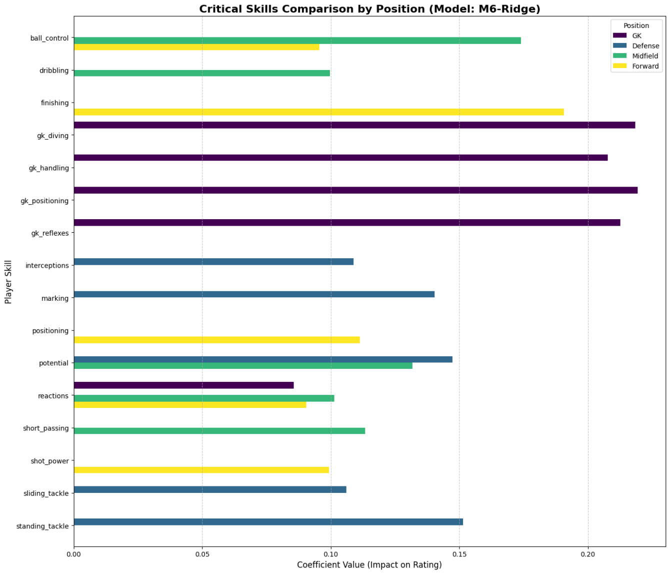

Fig. 8: Critical Skills Comparison by Position (Model: M6-Ridge)

Figure 8 highlights key skill contributions across different positions. As expected, goalkeeper-specific attributes (*gk_diving*, *gk_positioning*, *gk_reflexes*) are highly influential for goalkeepers but irrelevant elsewhere.

For outfield players:

- **Ball control** and **finishing** are critical for midfielders and forwards.
- **Standing tackle** and **marking** play a major role in defensive ratings.
- **Potential** and **reactions** show moderate impact across multiple roles, suggesting general importance.

Overall, the model effectively aligns skill relevance with position-specific expectations.

IV. Conclusion and Future Work

This study presented a two-stage hybrid framework to predict soccer players' overall ratings based on estimated positional roles. Due to the absence of ground-truth position labels in the dataset, we explored two alternative methods for position estimation: a statistical rule-based labeling and a KMeans-based clustering approach using average player coordinates. These position labels were then used to train regression models to predict players' overall ratings.

Experimental results revealed that the **M6: Rule-Based(All features) + Ridge Regression (General Position)** configuration achieved the lowest RMSE and highest R^2 score among all models, indicating that the statistical rule-based approach—despite its handcrafted nature—can effectively guide performance modeling when designed with position-specific features. While the KMeans-based method offers a spatially grounded alternative, its predictions showed notable divergence from the rule-based labels, especially in detailed roles, and underperformed in the regression stage.

A. Discussion on Limitations

A major limitation of this study is the lack of a validated ground truth for player positions. As a result, both the rule-based and KMeans labels serve as approximations, which may introduce noise into the regression models. Moreover, since overall ratings are predicted based on these approximated positions, the evaluation is indirectly affected by any error in role estimation. This limits the interpretability of absolute model accuracy and necessitates caution when generalizing results.

B. Potential Applications

The proposed two-stage hybrid framework holds promise for real-world applications in professional scouting, performance monitoring, and sports analytics.

In practice, clubs and analysts often lack access to comprehensive performance metrics. In such cases, spatial tracking data—readily available from GPS wearables or broadcast feeds—can be used to estimate tactical roles and initiate rating predictions, especially when enriched with statistical summaries. This approach could enable semi-automated player evaluation systems in scouting platforms, particularly for lower leagues or youth teams with limited manual labeling resources.

C. Future Work

Future efforts will focus on enhancing the positional labeling quality through manual annotation by domain experts. This would provide a reliable ground truth for evaluating clustering-based or rule-based positional inference, enabling more objective comparisons. Furthermore, interpretability methods such as SHAP can be applied to the existing regression models (Ridge, XGBoost, ElasticNet, Random Forest) to better understand the impact of each feature on predicted ratings across different roles. These improvements would strengthen the overall validity and transparency of the proposed two-stage framework.

References

[1] H. Mathien, "European soccer database," https://www.kaggle.com/datasets/hugomathien/soccer/data, 2016.

[2] P. Central, "Machine learning application in soccer: A systematic review," *PMC*, 2022.

[3] ——, "Identifying soccer players' playing styles: A systematic review," *PMC*, 2023.

[4] U. Di Giacomo, F. Mercaldo, A. Santone, and G. Capobianco, "Machine learning on soccer player positions," *International Journal of Decision Support System Technology*, vol. 16, no. 2, pp. 58–74, 2024.

[5] S. Azzami, H. Hadi, F. Alzami, C. Irawan, and A. Nurhindarto, "Clustering and profiling analysis for fifa football player using k-means," *Journal of Informatika*, 2025.

[6] A. Johansson, "Clustering soccer players: Investigating unsupervised learning on player positions," Uppsala University Thesis, Tech. Rep., 2022.

[7] S. Akhanli and C. Hennig, "Clustering of football players based on performance data and aggregated clustering validity indexes," *arXiv preprint arXiv:2204.07943*, 2022.

[8] Awwalm, "Fifa overall player ratings prediction using regression methods," https://github.com/awwalm/FIFAPredictor, 2020.

[9] U. C. Portuguesa, "Predictive models of the performance of professional football players," https://repositorio.ucp.pt/handle/10400.14/36762, 2021.

[10] I. McHale, P. Scarf, and D. Folker, "Plus-minus player ratings for soccer," *ResearchGate*, 2017.

[11] S. S. A. Conference, "Regularized adjusted plus-minus models for evaluating and scouting soccer players," https://arxiv.org/abs/2401.17832, 2024.