Ensemble Learning-Based Optimal Line-Up Prediction for Basketball

Anant Kedia AU2220040
Maanit Shah AU2220043
Ayush Chaurasia AU2220032
Samarth Bankar AU2220001
Viraj Bhatia AU2220038

***Abstract*—This project aims to propose a comprehensive and explainable framework for profiling basketball athletes using machine learning. The study leverages game data from the Northeast Conference (NEC) and Metro Atlantic Athletic Conference (MAAC) from 2020 to 2024, focusing on clustering and predictive modelling. Fuzzy C-means (FCM) clustering and Spectral Clustering with Feature Selection (SC-FS) were applied to identify player profiles based on key performance metrics. Additionally, an ensemble classification model was used for predicting wins based on team line-ups. Following this, a model utilizing the win prediction model and player positions is proposed for predicting optimal team line-ups.**

***Index Terms*—Machine learning, Sports analytics, Clustering algorithms, Predictive modelling**

I. Introduction

In the ever-evolving landscape of sports analytics in basketball, machine learning-based approaches have emerged as promising tools for enhancing athlete evaluation and team strategy development [1]. Studies in the field have demonstrated the efficacy of machine learning algorithms in predicting sports outcomes and assessing player performance.

Multiple recent studies have demonstrated the aforementioned. For instance, Wang et al. [2] utilise various different ML approaches to provide a comprehensive evaluation of their performance in predicting game outcomes. However, the inputs in their approach fail to capture real-world factors like team dynamics. Similarly, Chen et al. [3] use a hybrid model that combines various approaches including K-Nearest Neighbours (KNN), Extreme Gradient Boosting (XGBoost) and Stochastic Gradient Boosting (SGB). While this approach improves accuracy, the model's complexity increases, adversely affecting interpretability.

Several approaches have been explored to tackle the problem of explainability. Ouyang et al. [4] integrate Shapley Additive Explanations (SHAP) with XGBoost as an effective approach to improve explainability. However, this integration does increase the computational complexity and fails to generalise performance indicators. Islam et al. [5] propose iXGB, an approach to enhance XGBoost's interpretability by approximating decision rules from its internal structure and generating counterfactuals. Nevertheless, the approximation of decision rules may not capture all the complexities of the model.

All of the approaches discussed also fail at capturing some real-world factors. Hu et al. [6] quantify player interactions at synergies to enhance the XGBoost prediction model. However, quantifying synergies for a larger set of players is difficult. Also, the method is extremely sensitive to outliers.

This project focuses on tackling some of the limitations discussed above. The main objective of this project is to propose a comprehensive framework for profiling basketball athletes and identifying the optimal lineup for a team against a given opponent.

II. Methodology

We used a dataset consisting of the Northeast Conference (NEC) and Metro Atlantic Athletic Conference (MAAC) matches from 25/11/20 to 25/03/24. These conferences are affiliated with the National Collegiate Athletic Association (NAAC) Division-1 [7] [8]. Each match is represented by the two teams playing it, and the teams' respective information such as the athletes participating in the match and standard box statistics representing their performance. Furthermore, the teams' total score in the match and whether they won or lost is present in the data. Finally, an aggregate score of each athlete's performance, called Hollinger's Game Score, is present, calculated as

$$\begin{aligned} Game\ Score = PTS &+ (0.4 \cdot FGM) - (0.7 \cdot FGA) \\ &-\big(0.4 \cdot (FTA - FTM)\big) + (0.7 \cdot OREB) \\ &+ (0.3 \cdot DREB) + STL + (0.7 \cdot AST) \\ &+ (0.7 \cdot BLK) - (0.4 \cdot PF) - TO, \end{aligned} \tag{1}$$

where the abbreviations are the standard basketball box scores.

A. Fuzzy C-Means Clustering for Player Position Profiling

In basketball analytics, player positions are not rigidly defined, and athletes often exhibit characteristics of multiple roles. Traditional classification methods assign players to a single position, but in reality, a player may have overlapping skills suited for different roles. Understanding a player's suitability for multiple positions is crucial for team line-up decisions, player recruitment, and tactical planning. A fuzzy classification approach allows for a more nuanced position assignment, making it a more effective method for athlete profiling.

We used Fuzzy C-Means (FCM) clustering to obtain soft position memberships, as player positions are not rigidly defined, and athletes usually take on multiple roles.

We first extracted seven key performance metrics from the dataset: Field Goal Percentage (FG%), Points Scored

(PTS), Three-Point Percentage (3P%), Assists (AST), Rebounds (REB), Steals (STL), and Blocks (BLK). Two of these statistics, FG% and 3P%, were calculated using the formulas:

$$\text{FG\%} = \frac{\text{FGM}}{\text{FGA}}, \quad \text{3P\%} = \frac{\text{3PM}}{\text{3PA}} \tag{2}$$

Given a membership matrix U and weight matrix W from FCM clustering, where u_{ij} represents the degree to which player i belongs to cluster j, we computed position-specific memberships using:

$$P = UW \tag{3}$$

This transformation ensures that each player has a soft membership score for all five basketball positions, reflecting their suitability for multiple roles.

B. Spectral Clustering with Feature Selection

We implemented a clustering procedure proposed by Liu et al. called Spectral Clustering with Feature Selection (SC-FS) [9]. We aimed to find patterns within the athlete's individual box statistics. Furthermore, we chose specific cluster values to aim for finding strength and weakness patterns within individual athletes and teams as well.

The algorithm specifically tackles the problem of clustering high dimensional data, which requires more effort than simpler distance-based methods such as K-Means or DBSCAN. We implement spectral clustering (which is a graph-based clustering method) initially taking all the features into account. This provides us with a preliminary clustering of the data, although unreliable. Then, we compute the Within Cluster Sum of Squares (WCSS) of clusters made based on all subsets of parameters. We also take into account the Silhouette scores of these clusterings. This procedure provides us with a metric to measure which subsets of parameters provide the best clusterings.

C. Win Prediction using Ensemble Learning

1) Feature Engineering: Raw box score statistics provide a snapshot of a team's performance in a single game but may not effectively capture trends or contextual nuances. By engineering features the model can better understand a team's current form and momentum, which are critical factors in predicting game outcomes.

First, possessions were computed using the equation:

$$Possessions = FGA - OREB + TO + (0.4 * FTA) \tag{4}$$

Possessions were then used to compute the offensive and defensive efficiency of the teams. Offensive efficiency ($ORtg$) is the number of points a team scores in 100 possessions [10]. It is computed using the equation:

$$ORtg = (PTS/Possessions) * 100 \tag{5}$$

In a game, the offensive efficiency of a team is equal to the defensive efficiency of the opponent team [10]. Thus, the defensive efficiency ($DRtg$) was computed as opponent points allowed per 100 possessions using the equation:

$$DRtg = (OpponentPTS/Possessions) * 100 \tag{6}$$

For every game, a feature vector was assembled that included both the team's and opponent's last 5-game averages for $ORtg$, $DRtg$, and Game Score, as they provide a comprehensive overview of a team's efficiency. The outcome of the game (win or loss) was used as the target variable. This approach ensured that each data point reflected the comparative form of both teams leading up to the matchup.

2) Prediction Model: Ensemble learning is a machine learning paradigm where multiple models, often referred to as "base learners" are trained to solve the same problem and combined to obtain a better predictive performance.

Stacking is an ensemble technique that combines multiple classification models using a meta-classifier [11]. In this approach, base models are trained on the entire training dataset, and a meta-model is trained on the outputs (predictions) of these base models as features. The meta-model learns how to best combine the base models' predictions to improve overall performance [12].

Here, a stacking ensemble approach was employed to predict the winner between two teams in a basketball match. The base learners included Random Forest Classifier, XGBoost Classifier, and Logistic Regression. A Logistic Regression model served as the meta-learner.

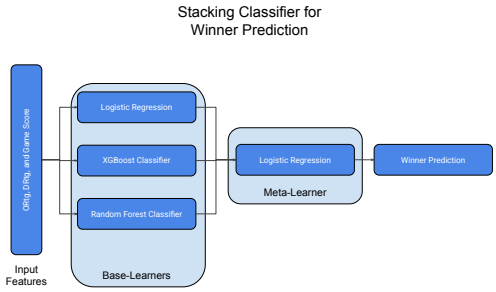

Fig. 1. Stacking Classifier used for Winner Prediction

D. Optimal Line-Up Prediction

The ensemble learning-based win prediction model and the position memberships from the fuzzy c-means clustering were used to develop an optimal line-up prediction model, which provides the optimal line-up for a team given the opponent team.

For the team in question, players with complete data and assigned positions are selected. All possible valid combinations of five players, one from each position (PG, SG, SF, PF, C), are generated. For each line-up, average $ORtg$ and Game Score are calculated. Using the trained ensemble model, the win probability against the specified opponent is predicted

for each line-up. The line-up with the highest predicted win probability is selected as the optimal line-up.

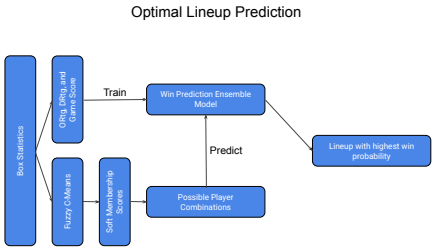

Fig. 2. Optimal Lineup Prediction

III. Results

A. Fuzzy C-Means

We applied Fuzzy C-Means clustering to group players into five clusters, corresponding to the five traditional basketball positions: Point Guard (PG), Shooting Guard (SG), Small Forward (SF), Power Forward (PF), and Center (C). However, from the five clusters generated, four had meaningful centroids and an additional cluster emerged, consisting of players with minimal impact. To map the clusters to meaningful basketball positions, we constructed a weight matrix based on domain knowledge and cluster centroids. By analysing the statistical tendencies of each cluster's centroid, we observed that some clusters leaned more toward specific positions. For example, clusters with high assists and moderate scoring were associated with PGs, while clusters with strong rebounding and blocking tendencies were linked to PFs and Cs. This knowledge guided the formation of the following position weight matrix:

$$W = \begin{bmatrix} 0.1 & 0.5 & 0.4 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2 & 0.5 & 0.3 \\ 0.2 & 0.3 & 0.4 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.6 & 0.1 & 0.2 & 0.1 & 0.0 \end{bmatrix} \tag{7}$$

Using W, the position-specific membership scores were calculated for each player.

B. Spectral Clustering with Feature Selection

We found that every clustering taking 3 or more features into account produced negative silhouette scores which is undesirable. The highest scores were produced by single parameter clusterings which are not useful as they simply represent percentile divisions. Finally, the few 2-feature clusters with good silhouette scores used the "Wins" feature which does not allow much for interpretation about the athletes and teams.

We concluded that the box statistics as is are not viable for much analysis, and performed the feature engineering steps described in subsection II-C1 for predictions.

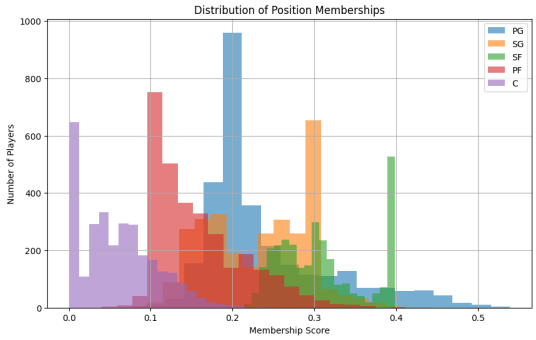

Fig. 3. Membership Histogram for the 5 Positions

C. Win Prediction using Ensemble Learning

The dataset was divided into training and testing subsets using an 80/20 split. Accuracy and Area Under the Receiver Operating Characteristic Curve (ROC AUC) were used as performance metrics to evaluate the performance of the ensemble model.

The ensemble model achieved an accuracy of 71.43%, indicating that it correctly predicted the outcome of approximately 71% of the games in the test dataset. The model obtained an ROC AUC score of 0.778, demonstrating its capability to distinguish between winning and losing teams. An ROC AUC score of 0.778 suggests that the model has a good ability to discriminate between the two classes across various threshold settings.

D. Optimal Line-up Prediction

To identify the optimal starting line-up for a given team against a specific opponent, the model evaluated all possible combinations of players, ensuring one player per position (PG, SG, SF, PF, C). Each line-up's average offensive efficiency and game score were calculated based on the players' recent performances (last five games). These metrics, along with the opponent's corresponding statistics, were input into the trained ensemble model to estimate the probability of winning. The line-up with the highest predicted win probability was selected as optimal.

```
Optimal Lineup:
             Athlete Position
28320  Sajada Bonner       PG
28317  Nalyce Dudley       SG
11665     Ann Porter       SF
28323   Faith Pappas       PF
28322    Kelsey Wood        C
```

Fig. 4. Optimal Line-up for Sacred Heart Pioneers against the Arizona State Sun Devils

For instance, when determining the optimal line-up for the Sacred Heart Pioneers against the Arizona State Sun Devils,

the players selected by the model for each position are given in Fig.4.

IV. Discussion

The integration of Fuzzy C-means clustering provided soft membership scores for traditional basketball positions. This approach aligns with the fluid nature of modern basketball, where players often exhibit skills across multiple positions. The resulting profiles offer a more nuanced understanding of player capabilities, which could assist recruiters in identifying talent that fits specific team dynamics.

The ensemble learning-based win prediction model achieved a respectable accuracy of 71.43% and an ROC AUC score of 0.778, validating its capability to assess team performance and predict outcomes. By incorporating engineered features ($ORtg$, $DRtg$, and Game Score from recent games), the model moves beyond static box scores and captures that capture the dynamics of the game.

Optimal line-up prediction demonstrated a practical application of the model's predictive power. By simulating all valid five-player combinations and evaluating them against an opponent's profile, the model delivers tactical insights previously unavailable through traditional metrics. This data-driven approach can enhance coach decision-making, especially in match-up planning.

Our work on clustering demonstrated that the data is insufficient to provide meaningful unsupervised patterns as the method considered every possible subset of parameters and was unable to find non-trivial ones with meaningful clustering.

A. Limitations

Our model does not explicitly account for factors such as player fatigue, injuries, home-court advantage, travel distance, or player synergy. While some of these factors are difficult to quantify or may require play-by-play data, which is not available with the desired degree of detail, they can significantly impact team performance. Furthermore, the fuzzy position memberships, while more nuanced than traditional classifications, still rely on a predefined weight matrix derived from domain knowledge, which introduces a degree of subjectivity.

Additionally, our win prediction model's 71.43% accuracy, while respectable, still leaves substantial room for improvement. Basketball outcomes are inherently difficult to predict due to the sport's dynamic nature and the impact of in-game adjustments, psychological factors, and random variation.

B. Future Research Directions

Building on the foundation established in this study, several promising research directions emerge. The framework could be extended to incorporate spatio-temporal data from player tracking systems, which would provide a more granular understanding of player movements, interactions, and tactical patterns. Moreover, graph-based approaches could better capture the network effects of player-to-player interactions.

Furthermore, the fuzzy position profiling approach could be refined by incorporating defensive metrics and advanced statistics beyond the standard box score. The position weight matrix could also be learned from data rather than prior knowledge, potentially revealing emerging position archetypes that traditional basketball taxonomy doesn't capture.

V. Conclusions

This study presents a comprehensive machine learning framework for basketball analytics that addresses key limitations in existing approaches. The fuzzy position memberships developed through our clustering approach offer a more realistic representation of player roles than traditional classifications, acknowledging the versatility required in modern basketball. Our ensemble learning model for win prediction achieved a competitive accuracy of 71.43%. By systematically evaluating all valid player combinations against specific opponents, our approach for optimal line-up prediction provides coaches with objective recommendations that complement their expertise and intuition. Despite its limitations, our framework makes significant contributions to basketball analytics by offering practical tools for player evaluation and line-up optimisation.

References

[1] R. P. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Applied Computing and Informatics*, vol. 15, no. 1, pp. 27–33, 2019.

[2] J. Wang, "Predictive analysis of nba game outcomes through machine learning," 01 2024, pp. 46–55.

[3] W.-J. Chen, M.-J. Jhou, T.-S. Lee, and C.-J. Lu, "Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association," *Entropy*, vol. 23, no. 4, 2021.

[4] Y. Ouyang, X. Li, W. Zhou, W. Hong, W. Zheng, F. Qi, and L. Peng, "Integration of machine learning xgboost and shap models for nba game outcome prediction and quantitative analysis methodology," *PLOS ONE*, vol. 19, no. 7, pp. 1–25, 07 2024. [Online]. Available: https://doi.org/10.1371/journal.pone.0307478

[5] M. R. Islam, M. U. Ahmed, and S. Begum, "iXGB: Improving the interpretability of XGBoost using decision rules and counterfactuals," in *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024)*, vol. 3, 2024, pp. 1345–1353. [Online]. Available: https://www.scitepress.org/Papers/2024/124740/124740.pdf

[6] H. Hu, G. Dimitrov, D. Menn, and S. Wu, "NBA player performance prediction based on XGBoost and synergies," in *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022. [Online]. Available: https://courses.cs.washington.edu/courses/cse547/23wi/old_projects/23wi/NBA_Performance.pdf

[7] "The Northeast conference." [Online]. Available: https://northeastconference.org/sports/2024/8/22/gen_aboutnec_2425.aspx

[8] "The MAAC." [Online]. Available: https://maacsports.com/sports/2017/6/20/GEN_0620175706.aspx

[9] T. Liu, Y. Lu, B. Zhu, and H. Zhao, "Clustering high-dimensional data via feature selection," *Biometrics*, vol. 79, no. 2, pp. 940–950, 2023.

[10] NBAstuffer, "Offensive efficiency in basketball explained," 2017. [Online]. Available: https://www.nbastuffer.com/analytics101/offensive-efficiency/

[11] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Machine learning*, vol. 54, pp. 255–273, 2004.

[12] Scikit-learn, "Stackingclassifier — scikit-learn 1.6.1 documentation," 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingClassifier.html