# Athlete performance in collegiate basketball: Predicting match line-up (Visualization on Dashboard)

***Abstract*—Predicting the ideal match lineup for a given basketball match is crucial for increasing a team's chances of success. Traditionally, the lineup selection process has heavily relied on the subjective assessments of coaches, the manual study of player performances, team dynamics, and in-game observations. These methods, while valuable, are time-consuming and may not fully utilize the wealth of player statistics at our disposal. This paper aims to tackle these limitations by introducing a more efficient and data-driven approach. This paper aims to improve the lineup selection process using machine learning techniques and data-driven decisions that can benefit coaches. The data used in this paper consists of 3111 records collected from 17 athletes over a time period of one season in Division-1 Women's basketball. It includes various attributes classified into categories ranging from training, sleep, recovery, subjective stress, in-game statistics, and countermovement jumps. We used state-of-the-art imputation techniques, like MICE, to impute data to handle missing values and enhance the quality of training data for improved model training. Factor Analysis on clean data was done to reduce dimensionality of factors and to calculate importance of each factor for prediction of Reactive Strength Index Modification(RSI mod).**

***Index Terms*—MICE, Extreme Gradient Boosting, RSI_mod, Game Score, Feature Selection**

## I. Introduction

Elite achievement in the fast-paced, intensely competitive world of collegiate basketball can be achieved as a consequence of careful planning and cutting-edge predictive analytics, in addition to skill and hard effort. The mission "Athlete Performance in Collegiate Basketball: Predicting Match Lineup (Visualisation on Dashboard)" captures the point where Machine Learning and sports technology expertise come together to change how we perceive games in the present world. The effectiveness of predictive analytics in sports is demonstrated by the most recent studies that highlight the intricate impacts of player readiness. The relevant research "Impact of Sleep and Training on Game Performance and Injury in Division-1 Women's Basketball Amidst the Pandemic" is a prime example. Using Machine learning techniques, sleep patterns, training data, and subjective well-being data are combined to predict sports performance outcomes and potential injuries. Furthering the discussion, the paper "A Holistic Approach to Performance Prediction in Collegiate Athletics: Player, Team, and Conference Perspectives" assesses how individual and overall team-level performance indicators complement one another. By incorporating data from the Reactive Strength Index Modified (RSImod) to the Player Efficiency Rating (PER), predictive sports activity analytics may now play a significant role in performance optimization, capturing the dynamics of individuals and groups. Novel machine learning techniques, such as the Extreme Gradient Boosting (XGB) classifier and ensemble methods, are not only instruments in this project but also triggers that cause a paradigm change from traditional heuristic judgment based on a coach's subjective analysis to data-driven accuracy.

This paper deals with the integration of such developments and also pushes the boundaries. The main goal is to create an ML-driven system that visualizes and forecasts the most effective match lineups, ensuring that each athlete's best performance is utilized at the right time. Once a purely retrospective tool, data now serves as a crucial component for creating weekly game plans that are both predictive of future outcomes and reactive to previous performance.

The revolutionary dashboard visualization, an interface that combines complex statistics into clear, useful insights, lies at the heart of this device. With this dashboard, coaches are given the ability to go above the surface of information and interpret its meaning about player preparation, team cohesiveness, and tactical matchups versus opponents. This transformative technology inspires a new way of thinking about sports analytics, opening up possibilities for deeper insights and more informed decision-making.

A machine like this completely changes the role of the instructor; instead of being a parent who relies on instinct, the teacher or coach, in this scenario, becomes a strategist who is supported by a virtual arsenal. From "what felt proper" to "what statistics suggest is optimal," each decision's motivation changes.

In summary, this paper uses machine learning techniques to forecast the Reactive Strength Index (RSI mod), an individual key performance indicator, and match lineups. It then presents the results on a visual dashboard so coaches can make better decisions for collegiate basketball tournaments.

## II. Methodology

The available dataset that was provided needed integration and was distributed in different .csv files with instructions on how to merge them, which were performed by our team.

Firstly, we have to write all the RSI data of different weeks into one sheet. Then, we combined all the data related to sleep, training, questionnaires and games into one .csv file with the help of a date column from two different files. The

exact process needs to be applied for season 3 data. Once the dataset was combined, we performed data cleaning on the dataset. There were a lot of empty cells, which we replaced with NA values. There were some data types in the dataset that needed to be corrected. We performed all the steps to eliminate empty cells and change the data type of the data points.

After data cleaning, a lot of data points were missing from the dataset. So, we performed Multiple Imputation by Chained Equations(MICE) on the dataset to impute the missing values in the dataset. MICE uses multiple methods to fill in the missing values and then combines the answers from all the methods to fill in the value. MICE is quite flexible and reduces bias while predicting the missing values. After imputing the dataset, we run the feature selection on the dataset to find the correlation of each feature and extract the most correlated feature.Along with this analysis we were able to identify how strong this connection is between the input parameters and the target variable. Variables with high correlation coeffficient were considered useful for the forecasting of RSI mod and kept for further data analyses. We made a dataset consisting of our important features. Afterwards based on the rsi_mod values we divided the each feature in different rsi_level ranging from 0 to 3.

We run the XGBoost on the dataset with different hyperparameters. The best hyperparameters we got were Learning Rate-0.1, max_depth-3, n_estimators:300, the accuracy of the XgBoost was 0.8571, and the F1 score was 0.8567. We also tried to run a random forest model on the dataset, and the parameters we got were max_depth-7, min_sample_split-10, n_estimators-200. The accuracy we got was 0.83, and The F1 score was 0.83. From the data, we concluded that XGBoost was better than random forest in this dataset.

| Algorithm | LR/MD/NE | Accuracy | F1 Score |
|---|---|---|---|
| XGBoost | 0.1 / 3 / 300 | 0.8571 | 0.8567 |
| Random Forest | 7 / 10 / 200 | 0.8331 | 0.8314 |

TABLE I
COMPARISON OF XGBOOST AND RANDOM FOREST ALGORITHMS WITH FEATURES AND PARAMETERS

## III. NEXT STEPS

- Use XGBoost to predict the RSI mod of n+1 week with different combination of previous weeks features.
- With the help of game score, sleep record and rsi of the each player. We would predict the the lineup for next match that contains the player with overall highest score.
- All the data of all players will be visible to the coach on the dashboard prepared by us.

## IV. RESULT

## V. CONCLUSION

In this paper, we addressed the challenge of predicting the optimal match lineup for collegiate basketball using data-driven approaches. We used Extreme Gradient Boosting (XGBoost), and employing data preprocessing techniques such as Multiple Imputation by Chained Equations (MICE). Our methodology involved data cleaning, feature selection, and model training, resulting in an XGBoost model with an accuracy of 85.71% and an F1 score of 85.67%, outperforming Random Forest in our dataset. Future steps include enhancing the predictive capabilities by incorporating time series to predict player performance in upcoming games, ultimately aiding coaches in making informed lineup decisions.

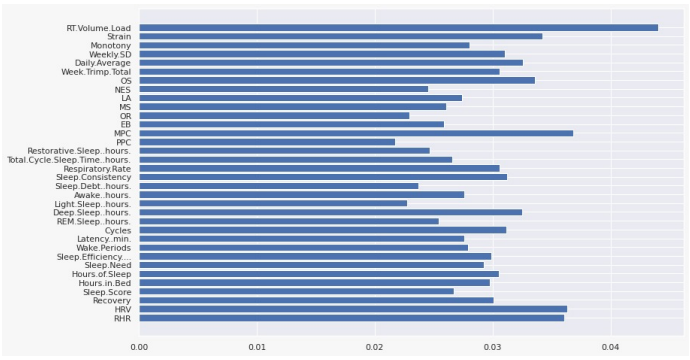


Fig. 1. Feature Selection

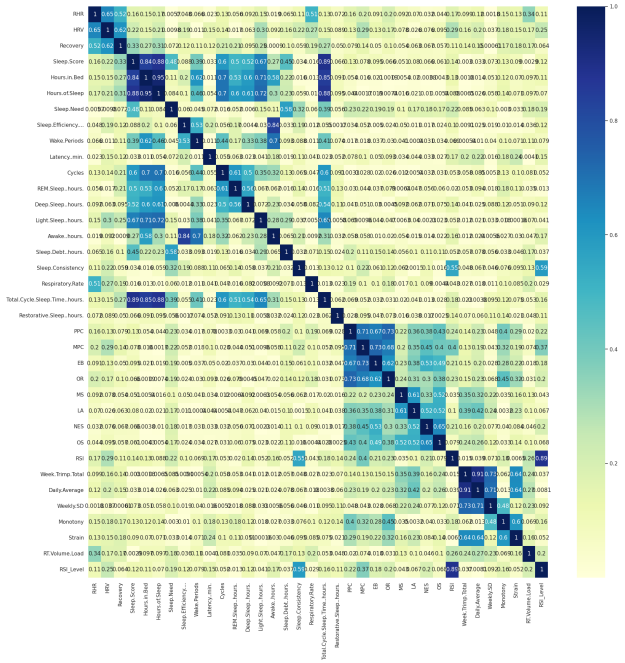

Fig. 2. Correlation between each feature.

## REFERENCES


[1] Taber, C.B., Sharma, S., Raval, M.S. et al. A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives. Sci Rep 14, 1162 (2024). https://doi.org/10.1038/s41598-024-51658-8.

[2] Impact of sleep and training on game performance and injury in division-1 women's Basketball Amidst the Pandemic S Senbel, S Sharma, MS Raval, C Taber, J Nolan... - Ieee Access, 2022