Using Evolutionary Algorithm and Neural Network to Predict the Best Time Allocation for NBA Teams

Kuang-Yi Tai
Department of Computer Science
National Yang Ming Chiao Tung University
HsinChu City 300, Taiwan
yitai77@gamil.com

Yu-Zheng Ma
Department of Computer Science
National Yang Ming Chiao Tung University
HsinChu City 300, Taiwan
lugmanma@gmail.com

ABSTRACT

In an intense environment in the National Basketball Association (NBA), accurately allocating players' playing time is a key factor in determining the success of the team. This study introduces a method that utilizes an evolution algorithm and neural network (NN) to produce the strongest time allocation. In this work, we first construct and fine-tune a regression model capable of predicting the win rate of each team. Moreover, we propose an evolutionary algorithm (EA) with a distinctive fitness function to reproduce the fittest individual, indicating the best time allocation. The results demonstrate that the proposed method can effectively distribute a plausible of amount time to each player based on his average performance.

CCS CONCEPTS

• Theory of computation~Design and analysis of algorithms~Mathematical optimization~Discrete optimization~Optimization with randomized search heuristics~Evolutionary algorithms
• Computing methodologies~Machine learning~Learning paradigms~Supervised learning~Supervised learning by regression

KEYWORDS

Evolutionary algorithms, machine learning

1. INTRODUCTION

In the high-pressure world of professional basketball, especially in the National Basketball Association (NBA), strategic decisions are the key to the success of the team. In these decisions, it is essential for the athletes to allocate time. Without a good time allocation, there is no way to win. It involves a series of complex factors such as player performance, endurance, team dynamics, and tactical cooperation against opponents. Traditionally, these decisions are mostly dependent on experienced coaches and staff. They often rely on intuition and basic statistical analysis, which makes their judgments have many limitations and even have personal subjective thinking. However, the increasingly emerging field of sports analysis provides new ways for data-driven decisions. The data performance has a very great right to speak. That is, only believes that the judgment and answers given by the data, try not to participate in the subjective consciousness as much as possible. This new type of deep learning and evolution computing model is expected to bring revolutionary changes in professional sports strategies.

This article introduces an advanced method that combines the evolution algorithm with the regression model of neural networks (NN) to optimize the time allocation of NBA players. We assume that data can bring our model judgment a very important foundation and knowledge. To explore this, we first prepared a comprehensive dataset, including detailed players and team statistics information and many advanced data from three NBA seasons. These advanced data allow analysts or models to learn a lot of implicit information from them. Therefore, the dataset we compiled can lay the foundation for our subsequent analysis and modeling.

The core of our method is to develop a regression model, which aims to predict the win rate of the team from various statistical characteristics. The architecture of this model includes continuous layers and non-linear activation functions, which aim to capture the inherent complex relationships in sports performance data. At the same time, we also obtain the actual win rate of each team for the corresponding years, so we can compare the results and fine-tune the model. In addition, the stability and predictive accuracy of the model are strictly evaluated by leave-one-out cross-validation (LOOCV) to ensure its reliability and prevent overfitting.

As a supplement to the forecast model, we use an evolution algorithm (EA) to optimize the player's playing time allocation for the rosters of Los Angeles Lakers (LAL) and Logs Angeles Clippers (LAC) in 2023-24 NBA season. This EA simulates and develops a variety of potential strategies guided by a fitness function consisting of model prediction and a penalty term, which restricts the behavior of EA to make it more suitable for real-world situations. Furthermore, We increase the important factor and fatigue that will affect the performance of players in reality. When the player has played more than the average time, the performance of the extra time will drop by 70%. This process of iteration enables us to explore a wide range of time allocation scenarios and focus on those strategies that are expected to maximize the performance of the team. Figure 2 summarizes the framework of our method.

The combination of these advanced computing technologies, NN and EA, represents a novel method in sports analysis. By making full use of the advantages of these two methods, we aim to provide a strong strategic decision-making tool for the NBA teams. This article introduces our methodology, the results of our methods, and the discussion on its limitations and promising prospects.

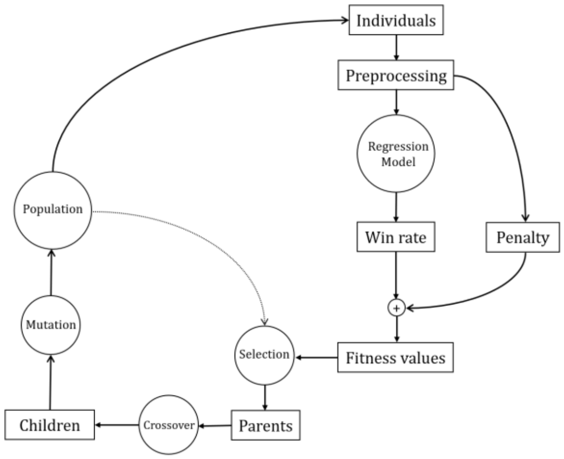


Figure 1: The architectural framework of the method.

2. DATASET

The dataset contains statistics of all players in the past three NBA seasons (2020-21, 2021-22, 2022-23) and the team roster of our targeted teams, LAL and LAC, in 2023-24 NBA season from [1], [2], in addition to the win rate of each team in each season from [3]. In other words, while using statistics of players in 2022-23 season, we build LAL and LAC with the roster in 2023-24 season.

Key statistics encompass traditional metrics such as points scored (PTS), assists (AST), offensive (ORB) and defensive rebounds (DRB), blocks (BLK), steals (STL), turnovers (TOV), and personal fouls (PF). Additionally, advanced metrics like two-point attempts (2PA), three-point attempts (3PA), free throw attempts (FTA), effective field goal percentage (eFG%), two-point percentage (2P%), three-point percentage (3P%), free throw percentage (FT%), player efficiency rating (PER), and win shares per 48 minutes (WS/48) are included.

2.1 Data Preprocessing

The pre-processing of this dataset includes four steps to ensure the consistency, correlation and analysis preparation of the data.

Firstly, we remove the data of some potential tanking teams with a win rate threshold of 0.35. Because these teams intentionally perform poorly in order to obtain higher draft picks in the future, their statistics are meaningless, which may contaminate the dataset and affect the quality of our model. Secondly, the number of players on each team has been reduced to only nine players to prevent the statistics of non-rotating players from misleading the model. Thirdly, in a 48-minute NBA game, a team must have five players playing on the court. Therefore, we linearly scale the statistics of all players on each team to make the total amount of time they play for 240 minutes. This step is necessary since the previous removal of some players brings about a lopsided reduction of time for each team. At last, we create the statistics of a team with the weighted sum of the statistics of each player on the team. Combing with the third step and this one, the statistics of each team can be expressed as:

$$\boldsymbol{Stats}_i' = \sum_{i=1}^{9} \frac{\sum_{j=1}^{9} \boldsymbol{MP}_j}{\mathbf{240}} \times \frac{\boldsymbol{MP}_i}{\mathbf{48}} \times \boldsymbol{Stats}_i$$

where $\boldsymbol{MP}_i$ is the amount of time the $\boldsymbol{i}$th player plays for on average. In addition, $\boldsymbol{Stats}$ should be replaced with different kinds of statistics, including PTS, AST, ORB, DRB, BLK, STL, TOV, PF, 2PA, 3PA, FTA, eFG%, 2P%, 3P%, FT%, WS/48.

3. REGRESSION MODEL

3.1 Architecture

We mainly hope to build a model. Its structure can analyze the complex interactive relationships of our dataset, and predict the winning rate of the team with a good prediction. In Figure 2, the architecture of this model is a sequential structure that is arranged by a linear layer, ReLU as the activation function, and batch normalization. The last linear layer outputs a continuous value, reflecting the prediction of the win rate of the team with team statistics as the input features.

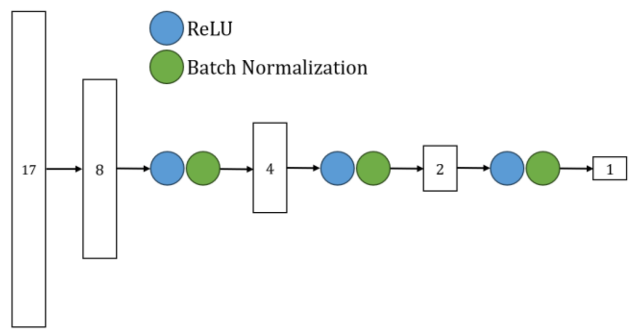


Figure 2: The architecture of the regression model.

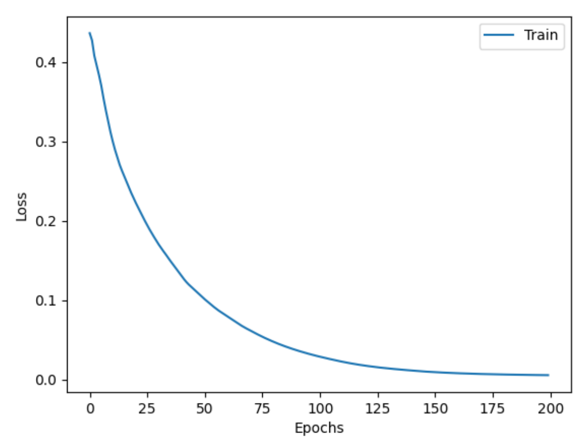


Figure 3: The loss vs. epoch graph of our model with the entire preprocessed dataset as input.

3.2 Evaluation

The evaluation of our regression model is to ensure its accuracy and reliability. We use mean square error (MSE) as the loss function and adopt leave-one-out cross-validation (LOOCV). This method is a rigorous test method, especially for a dataset with a limited number of samples (only for nearly three years).

In Figure 3, the model is trained on the entire dataset after the evaluation. This final model is preserved to serve as a vital part of the fitness function of the following EA.

4. EVOLUTIONARY ALGORITHM

EA, combined with the trained regression model, can select parents and reproduce the next population from generation to generation. In this method, we equip it with a novel and distinctive fitness function to simulate real-world situations.

4.1 Representation

The representation of individuals is set to be a list of nine continuous values, each of which represents the amount of time the corresponding player should play for in a game. In this way, we ensure a large spectrum of genotype space.

4.2 Fitness Function

The fitness function aims to determine the effectiveness and potential of the team under a specific time allocation of players. However, we do not restrict the individual representation in any way. Thus, it is certainly possible for a team to have no players on the court for a particular period of time, or allow players to play more than 48 minutes, which is the duration of a game. Before directly using each individual to calculate weighted statistics as the input of the win rate predictor, penalizing invalid individuals is essential to maintain authenticity. Consequently, the fitness function can be divided into four main steps.

First, for each individual, we calculate the sum of the nine elements. Since they represent the playing time of each player, we must ensure the sum is equal to 240 (48 minutes x 5 players on the court). To penalize invalid individuals and force EA to evolve more valid ones, the penalty term can be denoted as:

$$\boldsymbol{Penalty_{Individual_i}} = \left|\frac{\mathbf{240} - \sum_{j=1}^{9} \boldsymbol{Individual_j}}{\mathbf{240}}\right|$$

where $\boldsymbol{Penalty_{Individual_i}}$ is the penalty term of each individual and $\boldsymbol{Individual_i}$ corresponds to the $\boldsymbol{i}$th element of the individual. It is worth noticing that we divide the result by 240 and take the absolute value. The division normalizes the term to make it between 0 and 1, ensuring that it is within the same scale of win rate. In addition, invalid individuals including scenarios where the sum of all players' playing time is either lower or higher than 240 minutes. Taking the absolute value can prevent the fitness value from subtracting negative penalty values.

Second, even though in real-world situations, some extremely talented players can play for 48 minutes in a game, such as LeBron James in 2017-18 NBA season playoffs, we still regard them as anomalous scenarios. If an element of the individual is greater than the corresponding player's average playing time, we reduce the exceeding amount of time by 70 %. This can be represented as:

$$\boldsymbol{Exceeding_i} = \boldsymbol{MP_i} - \boldsymbol{Individual_i}$$

$$\boldsymbol{Individual'_i} = \boldsymbol{MP_i} + \mathbf{0.3} \times (\boldsymbol{Excceeding_i})$$

where $\boldsymbol{Exceeding_i}$ denotes the exceeding amount of time the $\boldsymbol{i}$th player plays for. Therefore, although the $\boldsymbol{i}$th player takes up $\boldsymbol{Individual_i}$ of 240 minutes, his performance will decrease by 70% for the rest of the game after he plays for more than $\boldsymbol{MP_i}$. This fatigue impact can effectively maintain $\boldsymbol{Individual_i}$ not significantly greater than $\boldsymbol{MP_i}$.

Third, the calculated and modified individual is used to construct the team statistics, which serves as the input of the predictor. It can output a continuous value indicating the predicted win rate of the team under such time allocation. Lastly, having the win rate subtract the penalty term can help EA select and reproduce both valid and fittest individuals for future generations.

4.3 Parameters

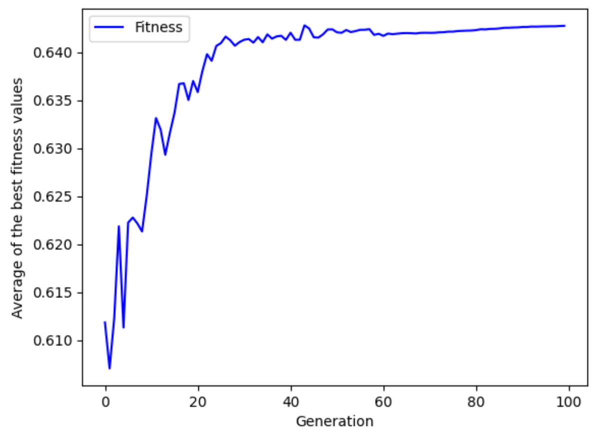
Parameters, including the number of independent runs, generations, and population size, are set up with a balance between efficiency and reliability, preventing EA from getting considerably time-consuming or defective. We execute EA for ten independent runs, each containing a hundred generations with a population size of a hundred per generation. Furthermore, our EA typically consists of three pivotal operations, tournament selection, uniform crossover, and Gaussian mutation. The tournament selection ensures EA can select the fittest as parents to reproduce offspring through uniform crossover In addition, given the utilization of continuous values, Gaussian mutation is a reasonable and deliberate method to introduce randomness into the algorithm.

5. RESULTS

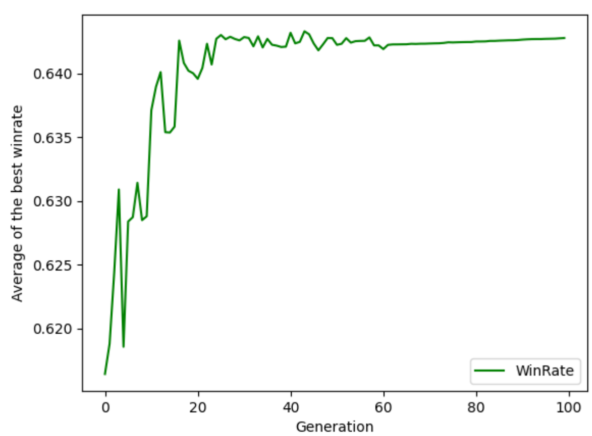
The results of our method demonstrate the average of the ten fittest individuals with the highest fitness value from ten independent runs, ensuring the stability and robustness of the outcomes. Figures 4 and 5 represent the evolutionary progress of EA throughout generations for LAL and LAC respectively. We calculate the average fitness value, win rate, and penalty of ten fittest individuals from ten independent runs in each generation. The statistics have shown EA effectively reproduces better and more valid individuals for future populations. Moreover, in Figure 6, the time allocations suggested by EA meet our expectations. All-star players, like LeBron James and Anthony Davis in LAL or Paul George and James Harden in LAC, receive a fair amount of playing time, indicating our method can produce reasonable time allocations for both teams.

6. DISCUSSION

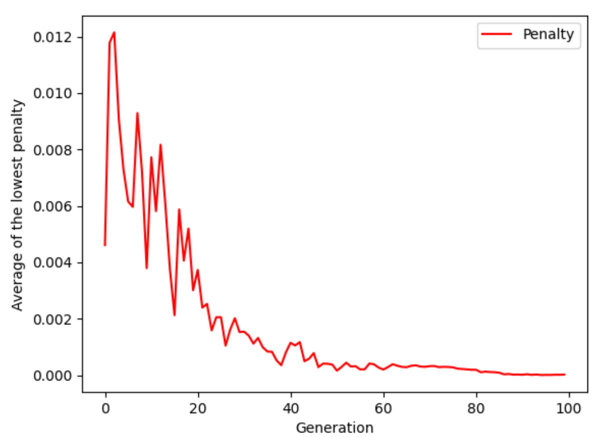
Although our method can allocate some all-star players significant amounts of playing time, other extremely talented players, such as Kawhi Leonard, cannot stress considerable contribution on the team. Moreover, in real-world situations, LAL and LAC have been benching Christian Wood and P.J. Tucker for



(a) Fitness Values

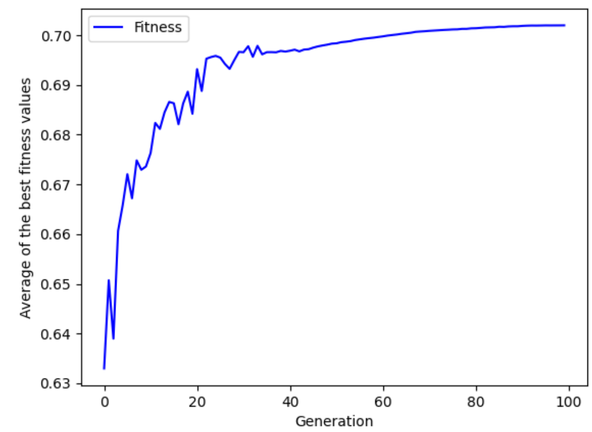


(b) Win Rates

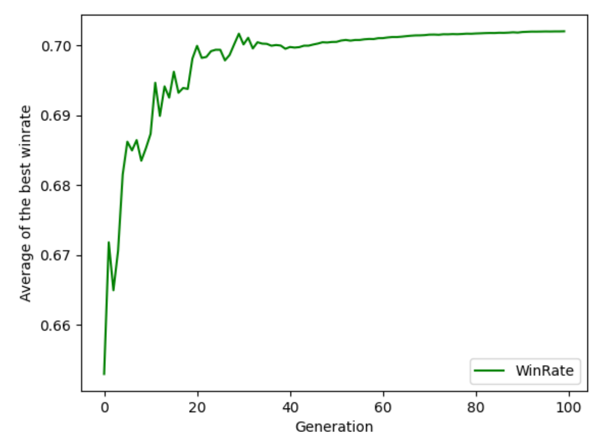


(c) Penalty Terms

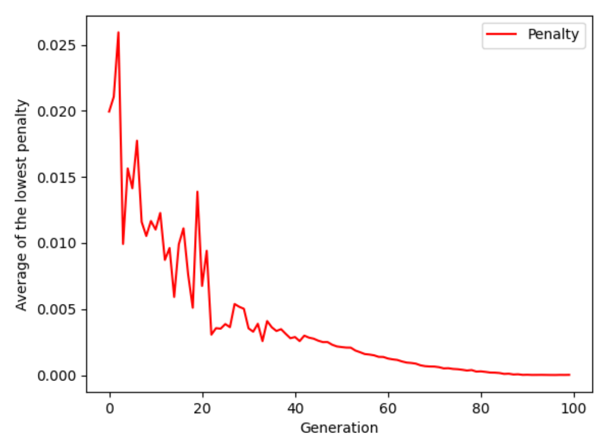
Figure 4: The evolutionary progress of EA throughout generations for LAL.



(a) Fitness Values

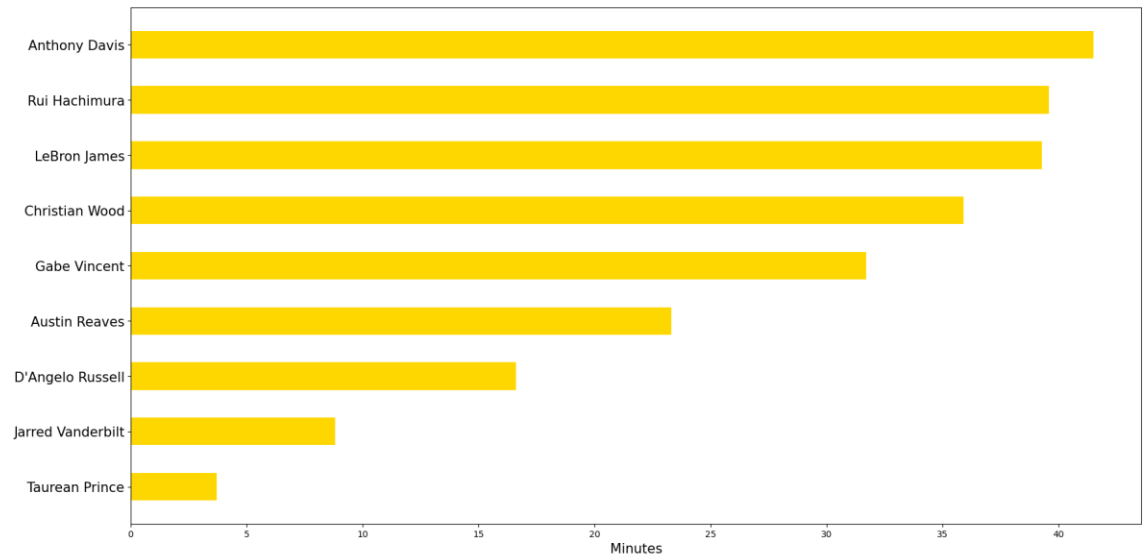


(b) Win Rates

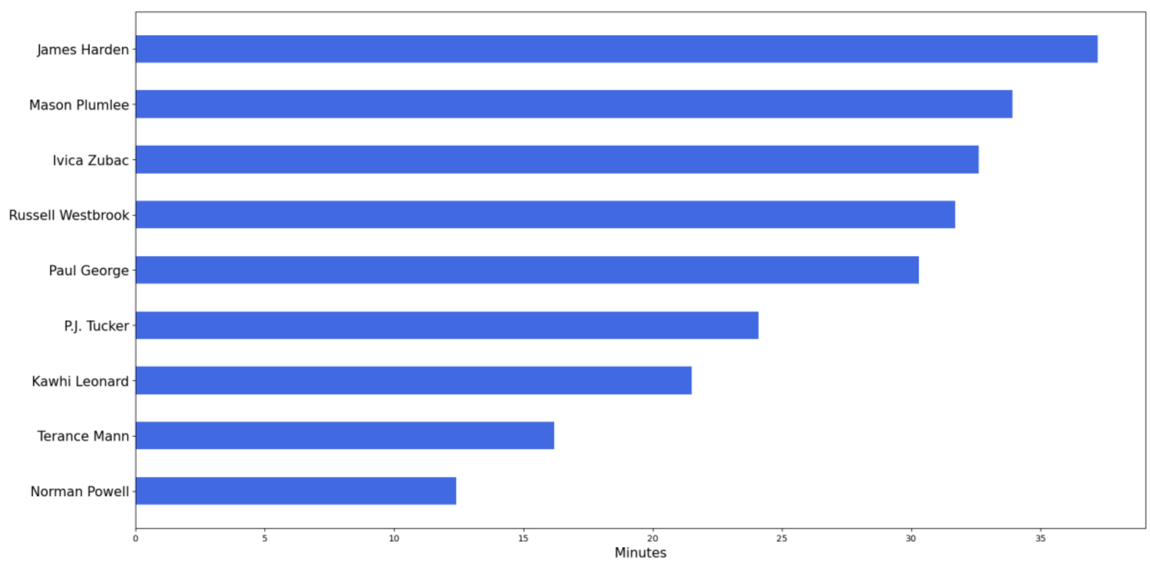


(c) Penalty Terms

Figure 5: The evolutionary progress of EA throughout generations for LAC.



(a) LAL Time Allocation



(b) LAC Time Allocation

Figure 6: The average of the ten fittest individuals with the highest fitness value from ten independent runs, demonstrating the final results of our method.

several games. However, the results consider them as vital parts of the teams. Both problems suggest that our method still requires further improvements. As a result, we propose its potential prospects and limitations.

First, we hope to expand the regression model to make the predictor more reliable. Currently, the model utilizes seventeen features and four fully connected layers. Nonetheless, some additional statistics have the potential to strengthen it. For example, just as rebounds can be divided into ORB and DRB, we can also separate WS/48 into two parts, offensive win share per minute (OWS/48) and defensive win share per minute (DWS/48). In this manner, we can have more comprehensive knowledge regarding each player's contribution on the two ends. Furthermore, Plus-Minus ($\pm$) and On/Off Court Stats (ON/OFF) can provide us with insights into players' impact on the performance of both teammates and opponents. Accordingly, our predictor will not solely focus on each player's own performance; in turn, it will take the bigger picture into account.

Second, there are five positions in a basketball game, point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C). Having one player of each position playing on the court can usually benefit the team in a great manner. Nevertheless, our method ignores the balance of time distribution between the backcourt and frontline. For instance, the result of LAL in Figure 6(a) suggests that four frontline players should play longer than any other players. Therefore, it is certainly possible to have less than two guards playing on the court, which can severely harm the team's performance, especially on the defensive end. Besides, in the contemporary NBA, the position of each player is not confined to the traditional five. The emergence of derivative positions, like point forward (PF) and swingman that can play various positions, makes the NBA even more unpredictable and variable in terms of game strategy. In addition, diverse coaching styles and team philosophies result in various roster combinations. Some head coaches prefer to utilize seven or eight players, while others may deploy ten or even twelve players in a game. It is certainly worth future investigations to allow each team to have distinct player participation numbers.

Third, in our method, we simplify the fatigue impact on players to a drop in performance by 70%. Nevertheless, we cannot quantify it in real-world scenarios since a wide variety of both mental and physical factors can account for the influence. A player can be too tired to perform well because of family issues, age problems, slight injuries, etc. Thus, our method may fail to allow all-star players to play overtime to carry the weight and hope of the team.

7. SUMMARY AND CONCLUSIONS

This paper started with the training process of a regression model that can predict the win rate of each team under a specific time allocation. The combination of data preprocessing and LOOCV ensured the stability, robustness, and accuracy of the predictor. On top of that, we proposed a distinctive and effective fitness function by utilizing the win rate predictor and introducing fatigue impact and constructed an EA to reproduce and evolve the best time allocations, which was the representation of individuals, for LAL and LAC respectively.

The emphasis of this study was to judge the team performance. The end was to consider comprehensive factors that can affect a player's performance in order to make the results more suitable for real-time scenarios. We proposed the method and implemented EA to accomplish this task, which may provide a strong strategic decision-making tool for the NBA teams.

8. REFERENCES

[1] David Abraham Smith. 2021. NBA Players Per Game and Advanced Stats 1980-2021. Retrieved from https://www.kaggle.com/datasets/davidabrahamsmith/nba-players-per-game-and-advanced-stats-19802021

[2] Jamie Welsh. 2023. NBA Per Game and Advanced Stats (2022-23 Season). Retrieved from https://www.kaggle.com/datasets/jamiewelsh2/nba-per-game-player-statistics-2022-2023-season

[3] Team Rankings. 2023. NBA Team Win % - All Games. Retrieved from https://www.teamrankings.com/nba/stat/win-pct-all-games