Sports Games Modeling and Prediction using Genetic Programming

1st Shengkai Geng
Department of Computer Science
Memorial University of Newfoundland
St. John's, NL, Canada
sgeng@mun.ca

2nd Ting Hu
School of Computing
Queen's University
Kingston, ON, Canada
ting.hu@cs.queensu.ca

***Abstract*—Sports games are largely enjoyed by fans around the globe. Plenty of financial assets, such as betting, need a reference to determine which team is more likely to win. In addition, club coaches and managers can benefit from using a analytical tool that suggests more efficient and suitable strategies to win. Genetic programming is a powerful learning algorithm for prediction and knowledge discovery. In this research, we propose to use genetic programming to model and predict the final outcome of NBA playoffs. We use the regular season performance statistics of each team to predict their final ranks in the Playoffs. Historical data of NBA teams are collected in order to train the predictive models using genetic programming. The preliminary results show that the algorithm is able to achieve a good prediction accuracy, as well as to provide an importance assessment of various performance statistics in determining the probability of winning the final championship.**

***Index Terms*—Genetic Programming, Prediction, Sports, Basketball**

I. Introduction

Sports is playing a significant role in people's daily lives today. More and more people are enjoying watching sports games. Fans are also interested in predicting the result of every game or tournament. Modeling and predicting sports games help sports teams to better understand and adapt strategies in order to win [1], as well as help fans to enjoy sports games better.

Studies have shown that the global sports market experienced a significant growth in the past several decades [2] and is predicted to continue growing [3]. In addition, since being legalized in more than 20 states in America, sports betting is becoming popular. According to a recent Statista survey, over 50% of US citizens admitted to placing a bet on a sports event at least once in their lives [4]. There is a significant amount of online revenue on sports betting [5]. Sports predictive models are used to forecast a game's result to help betting companies to define betting odds.

Machine learning [6], an approach to artificial intelligence, provides an effective way to analyze large volumes of data. Machine learning algorithms build a model using sample data, known as "training data", to make predictions or decisions without being explicitly programmed to perform the task [7].Genetic programming (GP) [8] can be used to solve a wide range of machine learning problems [9]–[12]. It is an algorithm inspired by the Darwinian principles of natural selection and evolution [13], [14].

National Basketball Association (NBA) is one of the most popular sports leagues in the world, composed of 30 teams. It was founded in 1946 and already became one of the four major professional sports leagues in the United States and Canada and is widely considered to be the premier men's professional basketball league in the world. A great number of sports fans are enjoying to watch NBA matches, especially playoffs, which would determine the final championship. Almost 70 years of NBA history provide a rich set of statistics for every team. Each statistic provides a performance evaluation for a team or a player, such as points per game, the number of rebounds, and the number of assists. These statistics can be collected and used to make predictions about a game or the whole season.

Many research studies have been conducted on basketball games prediction [15]–[20]. They used very different settings and methodologies. Most of them focused on single-game prediction based on seedings, betting odds, or performance. In addition, most predictive models from existing studies are difficult to interpret, i.e., users can not understand how a prediction has been made or what features are more influential. For sports game prediction, users are often not experts on artificial intelligence or machine learning, so providing an interpretable model as well as insights into the data are important [21].

In this paper, we propose a GP algorithm to analyze the statistics of NBA in the past 35 years. We build a model that can predict every team's performance in the playoffs based on their statistics in the regular season. Meanwhile, we provide results to help coaches and managers to get a good review of their team; they gain insights into how to improve their teams' chance to win the final championship. We provide results on which statistics are more influential in predicting a team's rank in the playoffs and present the final predictive model as a symbolic expression. We aim to provide easier to understand machine learning results for sports game prediction.

II. Data and Methods

In this section, we first describe the collection of data. We then introduce the design of the GP algorithm, followed by its

978-1-7281-6929-3/20/$31.00 ©2020 IEEE

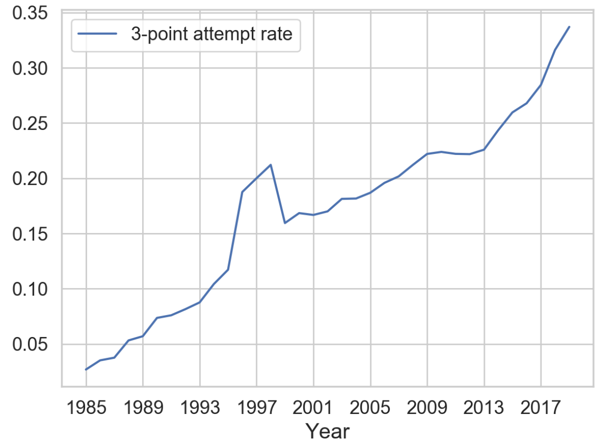


Fig. 1. The average 3-point attempt rate in the past 35 years

Name	Description
SOS	Strength of schedule: a rating of strength of schedule, the rating is denominated in points above/below average, where zero is average.
ORtg	Offensive rating: an estimate of points produced or scored per 100 possessions.
DRtg	Defensive rating: an estimate of points allowed per 100 possessions.
Pace	Pace factor: an estimate of possessions per 48 minutes.
FTr	Free throw attempt rate: number of free throw attempts Per field goal attempt.
3PAr	3-point attempt rate: percentage of FG attempts from 3-point range.
TS%	True shooting percentage: a measure of shooting efficiency that takes into account 2-point field goals, 3-point field goals, and free throws.
eFG%	Effective field goal percentage: this statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal.
TOV%	Turnover percentage: an estimate of turnovers committed per 100 plays.
ORB%	Offensive rebound percentage: an estimate of the percentage of available offensive rebounds a team grabbed.
FT/FGA	Free throws per field goal attempt.
OFG%	Opponent effective field goal percentage.
OTOV%	Opponent turnover percentage.
DRB%	Defensive rebound percentage of opponents: an estimate of the percentage of available defensive rebounds a team grabbed.
OFT/FGA	Opponent free throws per field goal attempt.
Attend	Total attendance at the team's primary arena in one season.

TABLE I
NBA STATISTICS AS FEATURES FOR LEARNING

configurations.

A. Data Collection

NBA provides plenty of different statistics for every team and player. These statistics almost cover every season since the league was founded in 1946. However, not all statistics are suitable for the learning process. For instance, Fig.1 shows the change of 3-point in the past 35 years. Before the 21st century, 3-point was considered as a method with low efficiency to get points. With the increasing abundance of tactics, 3-point starts to become popular among all teams. NBA teams choose to shoot more 3-points in order to get better offensive performance. Changes such as this make simple measures less effective for the use of a predictive feature. A key tenet for many modern basketball analysts is that basketball is best evaluated at the level of possessions [22]. During a single game, both teams have approximately the same number of possessions, because they alternate possession. However, over the course of a season and entirely history, teams play at very different paces, which can dramatically color their points scored and points allowed per game. Hence, per-possession and per-minute statistics are more useful than per-game statistics for analyses.

In order to have measures that reflect such style changes over time, we collect the advanced statistics of the last 35 years (from 1984 to 2018) from Sports Reference [23]. Advanced statistics provide a more in-depth means to look at box score, and more accurately evaluate the skills and productivity of a player or a team. For example, offensive rating measures how many points a team can get in 100 possessions, instead of points per game, which is a basic statistic. We choose 16 advanced statistics as features for the training of the GP algorithm. All the features are described in Table I.

We aim to predict the results of the entire NBA playoffs using one trained model. Given the input statistics of a season, this GP model outputs the results of the playoffs. We design a ranking system based on the tournament schedule of the NBA playoffs (see Table II). Each NBA season includes the regular season and the playoffs. During the regular season, each team plays 82 games, 41 home and 41 away against other teams. A team faces opponents in its own division four times a year (16 games). Each team plays six of the teams from the other two divisions in its conference four times (24 games), and the remaining four teams three times (12 games). Finally, each team plays all the teams in the other conference twice apiece (30 games). Playoffs begin in April after the conclusion of the regular season with the top eight teams in each conference, regardless of divisional alignment, competing for the league's championship title. The playoffs system is shown in Fig.2. The team who won the final championship was the first in that season. Who lost the final was the second place. Moreover, west teams and east teams were not equally good in most seasons. For instance, Fig.3 shows the difference, calculated by subtraction, of their average offensive rating and defensive rating from 1984 to 2018. Therefore, we rank teams in the two conferences separately, except for the final series where

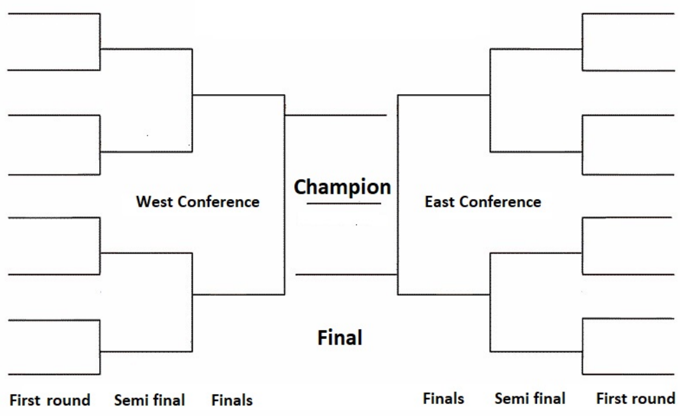


Fig. 2. NBA playoffs bracket

the east conference finalist plays against the west conference finalist. In addition, it is more important to correctly predict higher ranks, therefore, we set different weights for different ranks.

Rank	Number of teams	Weight	Explanation
1	1	512	The final champion
2	1	32	Who entered the finals but lost
3	1+1	16	Who entered the conference finals but lost
4	2+2	8	Who entered the conference semi-finals but lost
5	4+4	4	Who entered the Playoffs but did not win any series
6	7+7	2	Who did not enter playoffs

TABLE II
THE RANKING SYSTEM

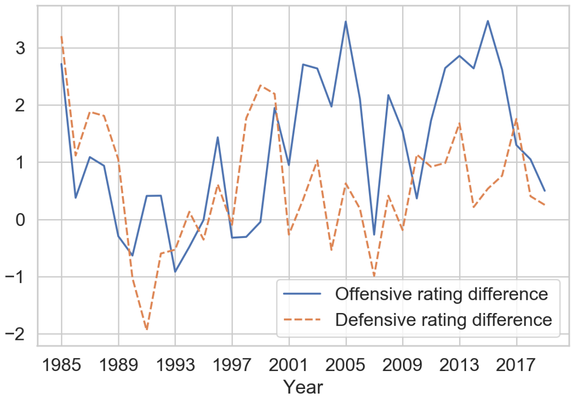


Fig. 3. Difference of offensive rating and defensive rating between west conference and east conference over last 35 years

We use a 7-fold cross validation, where we divide the dataset into 7 groups, and each group contains the statistics of 5 years (seasons). We run the GP algorithm 10 times for using one group as the test data and the rest 6 groups as training data. Therefore, we collect 70 unique runs of the algorithm in total.

B. GP Algorithm Design

The goal of our GP algorithm is to predict the ranking of all teams in the playoffs using their regular season statistics as the input. In this research, we use the tree-based GP [24]. The terminal set includes the 16 features described in Table I, and 10 constant numbers, 1 to 10. The function set includes addition, subtraction, multiplication, and protected division.

To evaluate the fitness of one GP tree, we compare the predicted ranking and the actual ranking of all the seasons in the training data.

The ranking error of one team can be calculated as:

$$e_{team}(t) = r^{'}(t) \times weight(r(t)), \qquad (1)$$

where $r^{'}$ is the rank difference, t is the corresponding team, $r(t)$ is that team's actual rank, and $weight(r(t))$ is the weight of that rank position.

For example, if the team Lakers is predicted with a rank 3, i.e, stopped after the conference final, but it was actually the final champion, the error would be $|3-1| \times 512 = 1024$.

Then, the error of one season can be computed as:

$$e_{season}(s) = \sum_{t} e_{team}(t), \qquad (2)$$

where t is all the teams in season s.
Finally, the fitness of one GP tree is defined as:

$$fitness = \sum_{s} e_{season}(s), \qquad (3)$$

where s is all 30 seasons in the training dataset.

The parent selection method used in this research is tournament selection. We randomly pick a group of individuals from the population and compare their fitness. We choose two individuals with the best fitness as the parents.

We perform subtree crossover by exchanging the subtrees starting at two randomly selected nodes in the given parents. After creating two new individuals, we use them to replace two individuals with the worst fitness in the tournament parent selection. For mutation, we randomly pick a mutation point, and replace the subtree using a randomly generated tree. We use stagnation termination, i.e., we stop the evolution process if the best fitness has not changed for 300 consecutive generations.

C. Parameter Tuning

To have a good performance, getting a set of most suitable parameters (the population size, the mutate rate, and the tournament size) is essential. Because the final championship is the most important in the prediction, we test 72 different combinations of parameters and record the best predictive

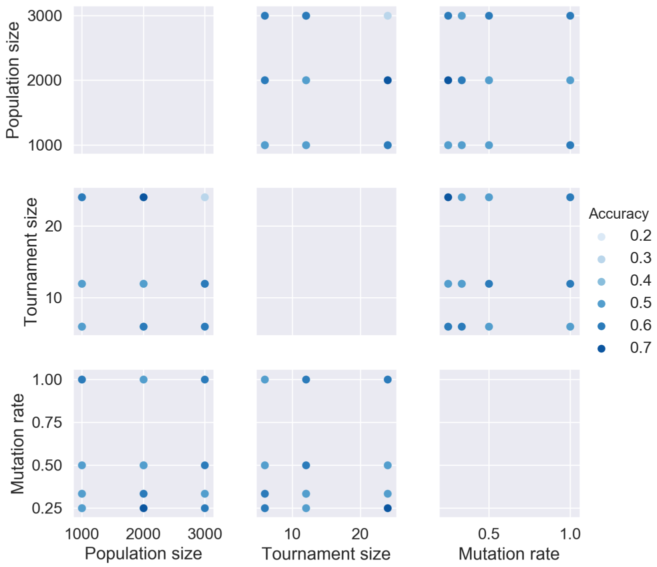


Fig. 4. The accuracy of different parameter settings

Parameters	*Setting*
Representation	Binary tree
Tree initialization	Randomly generate a tree with a max depth of 5
Terminal set	16 statistics, numbers from 1 to 10
Function set	Add, Sub, Multi, Protected Div
Population size	2000
Crossover probability	1
Parent selection	Tournament selection
Crossover/Mutation method	Subtree replacement
Maximum tree depth	8
Mutation rate	1/3
Tournament size	24
Fitness function	Rank error
Stop condition	Stagnation termination
Max generation	3000

TABLE III
PARAMETER SETTING

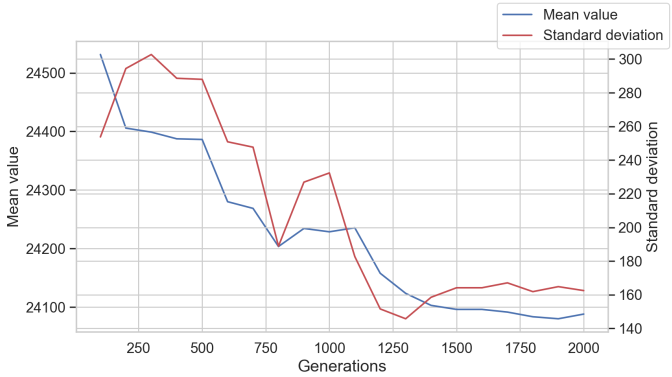


Fig. 5. The mean fitness and standard deviation during evaluation

accuracy of the final championship of every run. Each combination is tested 3 times and we record the mean accuracy of the best models. Fig. 4 shows the parameter comparison results. The best parameter setting of this GP algorithm is a population size of 2000, a mutate rate of 0.3, and a tournament size of 24. The complete parameter setting is shown in Table III.

III. RESULTS

In this section, we first show the overall performance of this GP algorithm. We then present the prediction result. Last, we show the feature importance analysis result.

A. Performance of GP

Fig. 5 shows the mean best fitness and the standard deviation at each generation of one representative run. It can be seen that the error of each run is decreasing as well as the standard deviation.

We also analyze the success rate of the GP algorithm. Out of 70 best evolved models (from 70 independent runs), 52 models have a predictive accuracy higher than 40%, and 21 models have an accuracy higher than 60%.

B. Prediction Results

From the 70 best evolved predictive models, we choose the best prediction model with the best fitness and the best final championship prediction accuracy. This model has a fitness of 24064, and a final championship prediction accuracy of 0.8. This model (GP tree) can be represented as a mathematical expression shown as follows.

$$
\begin{aligned}
Y = {} & (((11 + ((ORtg \times ((((OTOV - 1 - (((OFT/FGA \\
& \div 3) \div \frac{5}{2}) \div ((((8 - TOV) - 3PAr) - (((16 + eFG \\
& - ORB) - ((DRtg \div 5) + DRtg + SOS)) + ((OFG \\
& \times 5) + \frac{9}{4}))) + 10))) \div 3) + eFG) + 3)) - 3)) + (3- \\
& ((-Pace \times 7) \div OFG))) - (((OFG \div FTR) - (3+ \\
& ((((((10 \times ((((FT/FGA - 6) - (7 \times Attend)) \times (-\frac{1}{6})) \\
& + 10 + Pace)) + OFT/FGA) \div (TOV \times 4)) \times (36\div \\
& (7 \times Pace))) - (((((\frac{72}{7} + (((3PAr - TOV) \times \frac{4}{5})\times \\
& ORtg + OFT/FGA - ((3PAr + DRtg + 5 - SOS \\
& - eFG) \times (48 - ORB - eFG)))) - 23) \times ((8 - TOV) \\
& + 5)) \div OFT/FGA)) - 80))) \div (4 + (8 - eFG)))) - ((7 \\
& - Pace - DRtg) \times (-\frac{147}{3}) - TOV)
\end{aligned}
\tag{4}
$$

In this formula, DRtg(defensive rating), eFG(efficient field goal percentage), FT/FGA(free throws per field goal attempt) are the features with the highest number of occurrences. These high occurrences indicate that to win the championship, good defense performance and efficient shoots are important.

The best model is able to correctly predict the final champion as well as the ranking of the entire playoffs. One example of the prediction is shown in Table IV and Fig. 7. We use the best model to give every team in the 2008 season a score based on their statistics in the regular season, which is shown in Table IV. Those scores can be seen as the rating of the corresponding team. Then we rank these value following the rules of NBA playoffs bracket. The comparison of the actual ranking and the predicted ranking is shown in Fig. 7 It can be seen that the ranking of most of the teams in 2008 season

Team	Conference	Score
Boston Celtics	East	361.263489821
Los Angeles Lakers	West	355.708922399
Detroit Pistons	East	354.613450819
Utah Jazz	West	354.506982964
New Orleans Hornets	West	350.125536215
San Antonio Spurs	West	348.740139577
Phoenix Suns	West	347.331848175
Houston Rockets	West	347.26664762
Dallas Mavericks	West	347.108114323
Orlando Magic	East	344.603219486
Denver Nuggets	West	341.348027922
Toronto Raptors	East	337.292089407
Philadelphia 76ers	East	328.034888749
Cleveland Cavaliers	East	325.410094292
Washington Wizards	East	325.310003045
Atlanta Hawks	East	318.46345551

TABLE IV
THE PREDICTION SCORE OF EACH TEAM IN 2008 SEASON

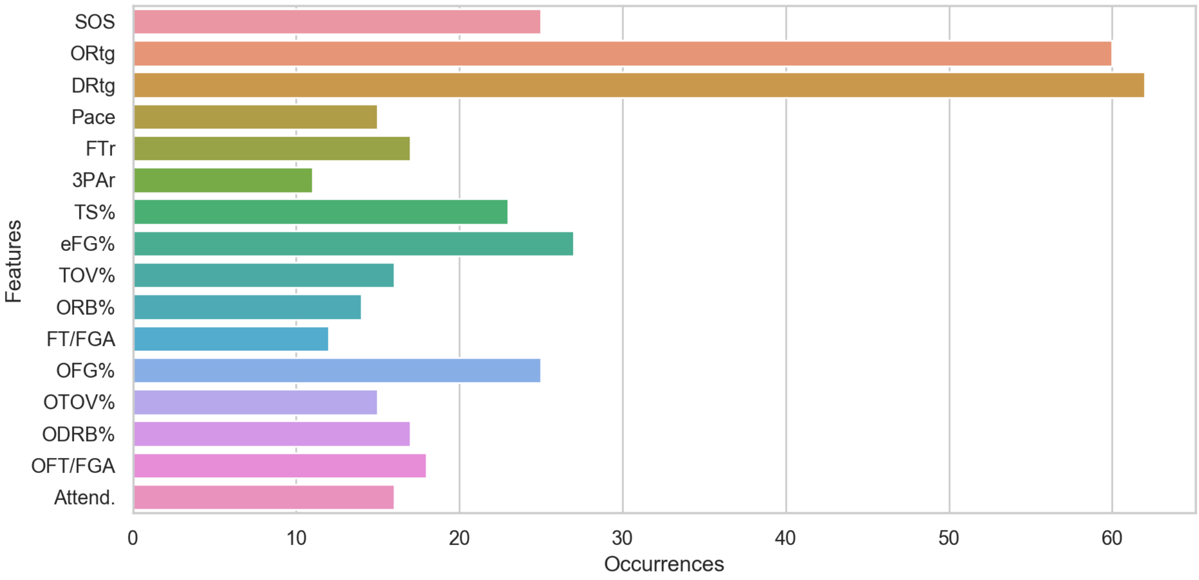
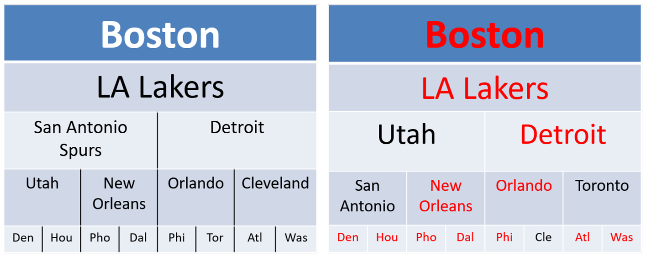


Fig. 6. Feature importance results



(a) Actual ranking (b) Predicted ranking

Fig. 7. Actual and predicted ranking of NBA 2007-2008 season.Those grid with red color represent the rank position which is correctly predicted.

Rank position	Accuracy			
	All	>0.4	>0.6	> 0.8
Final Champion	0.42	0.5	0.65	0.8
Conference Champion	0.50	0.51	0.49	0.68
Lost in conference final	0.26	0.26	0.3	0.48
Lost in semi-final	0.33	0.33	0.37	0.37

TABLE V
MEAN PREDICTING ACCURACY OF ALL RANKING PLACES PREDICTED BY MODELS AND GROUPED BY THEIR CHAMPION ACCURACY

is correctly predicted. The model successfully predicts the final champion Boston Celtic, the runner-up Lakers, and the conference finalist Detroit.

We also investigate the accuracy of predicting each exact rank position for all 70 runs. Table V shows the prediction accuracy of each place in the playoffs. We group all models by their champion accuracy to four groups, all models, models with champion accuracy over 40 percent, models with champion accuracy over 60 percent, and models with at least 80 percent champion accuracy. It can be seen that the models which were generated by this GP method not only have excellent performance in predicting the final champion but also can be used to predict other places, especially for those models with good performance on predicting the final champion.

Note that the prediction is based on regular season statistics. However, unpredictable events may happen during the playoffs, such as player injuries. Besides, staffs and players change in a team during the regular season, which is very common, and could reduce the reliability of the statistics of that team. So, finding a perfect model is extremely difficult.

C. Feature Importance

We record the frequency of every feature occurring in the best evolved predictive models of the 70 GP runs. The result can be used to analyze which features are more critical to the playoffs. This can be very helpful for coaches and team managers. Fig. 6 shows the result of feature importance assessment.

Here are some highlights of our observations:

- The most important factors are the offensive rating and defensive rating, which are the best statistics to represent the offensive and the defensive ability of a team.
- Although offense may seem a bit more important than defense in this result, limiting opponent's shooting percentage also plays an essential role.
- The low occurrence of 3-Point Attempt Rate Percentage indicates that shooting more 3-point may not be very helpful.

- A novel observation is that attendance is an important factor. The support of fans helps their team to win.

IV. Conclusion

In this research, we developed a new method to predict sports results using GP. Our GP algorithm can evolve highly accurate prediction of the entire playoffs ranking, as well as provide insights into what factors are more influential on winning the game. Our GP algorithm can be used to predict a variety of sports games, as long as it has a similar games bracket as NBA playoffs which is most sports leagues are using.

Future extension of this research includes considering more features and samples in addition to the 16 statistics and 35 years of samples used in this research. Statistics of the last games of a regular season may reflect the playoffs better than the statistics of the whole regular season. Finding a way to reflect injury, as a factor with a significant impact in sports, could improve the prediction results.

References

[1] S. Nunes and M. Sousa, "Applying data mining techniques to football data from european championships," in *Actas da 1ª Conferência de Metodologias de Investigação Científica (CoMIC'06)*, 2006.

[2] A. Kearney., "Global sports market - total revenue from 2005 to 2017 (in billion u.s. dollars)." 2019, https://www.statista.com/statistics/370560/worldwide-sports-market-revenue/, Last accessed on 2019-08-30.

[3] PwC., "Sports sponsorship market size in north america from 2006 to 2022 (in billion u.s. dollars)." 2018, https://www.statista.com/statistics/194221/total-revenue-from-sports-sponsorship-in-north-america-since-2004/, Last accessed on 2019-08-30.

[4] S. Survey, "Have you ever placed a bet on a sports event?" 2017, https://www.statista.com/statistics/661672/betting-on-sports-events-in-us/, Last accessed on 2019-11-30.

[5] P. P. Betfair, "Online revenue of selected sports betting / daily fantasy sports companies in the united states in 2017 (in million u.s. dollars)," 2018, https://www.statista.com/statistics/917442/us-sports-betting-fantasy-sports-online-revenue/, Last accessed on 2019-08-30.

[6] D. Michie, D. J. Spiegelhalter, C. Taylor *et al.*, "Machine learning," *Neural and Statistical Classification*, vol. 13, 1994.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. springer, 2006.

[8] R. Poli, W. B. Langdon, N. F. McPhee, and J. R. Koza, *A Field Guide to Genetic Programming*. Lulu Enterprises, UK Ltd, 2008.

[9] T. Hu, K. Oksanen, W. Zhang, E. Randell, A. Furey, and G. Zhai, "Analyzing feature importance for metabolomics using genetic programming," in *European Conference on Genetic Programming*. Springer, 2018, pp. 68–83.

[10] T. Hu, K. Oksanen, W. Zhang, E. Randell, A. Furey, G. Sun, and G. Zhai, "An evolutionary learning and network approach to identifying key metabolites for osteoarthritis," *PLoS Computational Biology*, vol. 14, no. 3, p. e1005986, 2018.

[11] Y. Zhang, T. Hu, X. Liang, M. Z. Ali, and M. N. S. K. Shabbir, "Fault detection and classification for induction motors using genetic programming," in *Genetic Programming, Proceeding of EuroGP 2019*, ser. Lecture Notes in Computer Science, L. Sekanina, T. Hu, N. Lourenço, H. Richter, and P. García-Sánchez, Eds., vol. 11451, Springer, Cham, 2019, pp. 178–193.

[12] Y. Zhang, Y. Chen, and T. Hu, "Classification of autism genes using network science and linear genetic programming," in *Genetic Programming. Proceeding of EuroGP 2020*, ser. Lecture Notes in Computer Science, T. Hu, N. Lourenço, E. Medvet, and F. Divina, Eds., vol. 12101, Springer, Cham, 2020, pp. 279–294.

[13] T. Bäck, D. B. Fogel, and Z. Michalewicz, *Handbook of Evolutionary Computation*. CRC Press, 1997.

[14] L. Davis, "Handbook of genetic algorithms," 1991.

[15] S. B. Caudill, "Predicting discrete outcomes with the maximum score estimator: The case of the NCAA men's basketball tournament," *International Journal of Forecasting*, vol. 19, no. 2, pp. 313–317, 2003.

[16] B. Loeffelholz, E. Bednar, K. W. Bauer *et al.*, "Predicting nba games using neural networks," *Journal of Quantitative Analysis in Sports*, vol. 5, no. 1, pp. 1–17, 2009.

[17] H. Manner, "Modeling and forecasting the outcomes of nba basketball games," *Journal of Quantitative Analysis in Sports*, vol. 12, no. 1, pp. 31–41, 2016.

[18] J. W. Rosenfeld, J. I. Fisher, D. Adler, and C. Morris, "Predicting overtime with the pythagorean formula," *Journal of Quantitative Analysis in Sports*, vol. 6, no. 2, pp. 1–19, 2010.

[19] H. O. Stekler, A. Klein *et al.*, "Predicting the outcomes of ncaa basketball championship games," *Journal of Quantitative Analysis in Sports*, vol. 8, no. 1, pp. 1–10, 2012.

[20] E. Štrumbelj, "On determining probability forecasts from betting odds," *International journal of forecasting*, vol. 30, no. 4, pp. 934–943, 2014.

[21] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[22] J. Kubatko, D. Oliver, K. Pelton, and D. T. Rosenbaum, "A starting point for analyzing basketball statistics," *Journal of Quantitative Analysis in Sports*, vol. 3, no. 3, pp. 1–24, 2007.

[23] S. Reference, "Basketball reference," https://www.basketball-reference.com/, Last accessed on 2019-09-07.

[24] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT press, 1992, vol. 1.