A Study on Olympic Medal Prediction Based on Random Forest and Polynomial Regression

Liyu Gu, Jieliang Ouyang *

School of Computer Science and Artificial Intelligence & Aliyun School of Big Data & School of Software, Changzhou University, Changzhou, China, 213159

* Corresponding Author Email: jieliangouyang@gmail.com

Abstract. The Olympic medal table serves as a symbol of a nation's sporting competitiveness, and fluctuations in medal counts reflect underlying patterns within complex data. This study aims to uncover the mathematical logic behind the competition for medals and to identify the intrinsic relationships between these patterns and historical data. Based on attributes that reflect a country's medal-winning potential, a predictive model is constructed to estimate medal counts with a high degree of accuracy. Specifically, six key factors influencing medal counts are identified and integrated into a random forest regression model. A clustering model based on comprehensive indicators is applied to classify and quantify countries according to their Olympic performance. Polynomial regression is then employed to forecast relevant data for the 2028 Olympic Games. These forecasts are subsequently used to predict the number of medals in 2028 using the trained random forest model, and prediction intervals are established based on the MAPE error range.The model's performance is evaluated using MSE, MAE, and R2 metrics for gold, silver, bronze, and total medal predictions. Results show low prediction errors and strong goodness-of-fit, with R^2 values of 0.93588, 0.96743, 0.96554, and 0.83254, respectively. These outcomes indicate that the model demonstrates high predictive accuracy and robustness across all medal categories.

Keywords: Medal Prediction, Clustering Model, Random Forest, Polynomial Regression.

1. Introduction

The Olympic medal table is a crucial symbol reflecting a nation's overall sporting strength and competitiveness. Accurate prediction of medal counts holds not only academic value for researchers but also practical significance for policymakers, sponsors, and sports strategists [1].

In recent years, researchers have proposed more sophisticated models to improve prediction accuracy. Peters T. et al. explored the relationship between socio-economic indicators and Olympic success, concluding that Poisson regression best explained medal outcomes, while XGBoost delivered superior predictive performance [2]. Christoph Schlembach et al. proposed a two-stage random forest model based on socio-economic indicators to predict medal distributions for the Tokyo 2020 Olympics [3]. These efforts highlight the increasing role of machine learning and optimization techniques in Olympic medal forecasting.

Beyond medal prediction, the random forest algorithm has found widespread application across various fields. For example, Bunn [4] applied a random forest model to predict outcomes in collegiate women's lacrosse, achieving an accuracy of 86.0%, while Zhao [5] combined graph convolutional networks with random forest feature extraction to enhance NBA game prediction, reaching a 71.54% success rate. Both studies identified key performance indicators as critical to improving model accuracy. In environmental science, random forest and boosted regression tree (BRT) models have been applied to predict stream water quality using diverse watershed and climatic features [6]. In aviation, Zhen Guo et al. combined random forest regression with the Maximal Information Coefficient (MIC) to predict flight departure delays, outperforming linear regression and neural network models [7].

Similarly, polynomial regression models have demonstrated strong capabilities in modeling nonlinear relationships. Jiahang Liu et al. developed a polynomial regression model with physical constraints to predict the complex motion of knuckleballs in soccer free-kicks, accurately reproducing real match trajectories [8]. Belany P et al. conducted a comparative analysis between polynomial

regression and artificial neural networks in predicting lighting energy consumption in office buildings, concluding that polynomial regression performs well in structured data scenarios [9]. In the domain of autonomous driving, Tabelini L et al. proposed an innovative lane detection method that outputs polynomial representations of lane markings through deep polynomial regression, achieving both real-time performance and accuracy comparable to state-of-the-art models [10].

Inspired by these studies, this paper proposes a random forest-based framework for predicting Olympic medal counts. Six key influencing factors are extracted through the analysis of historical Olympic data, and a clustering-based random forest regression model is constructed. Polynomial regression is further used to estimate variables that are not directly available for future Olympic Games. Finally, this framework is applied to predict the medal counts for each country in the 2028 Los Angeles Olympics. The proposed model aims to deliver accurate, robust, and interpretable predictions for national Olympic performance. (Data sourced: https://www.comap.com/membership/member-resources/item/models-for-olympic-medal-tables.)

2. Predicting Olympic Medals with a Random Forest Approach

2.1. Analysis of Influencing Factors

To predict the number of medals a country may earn, it is essential to analyze the factors that might influence the medal count based on the existing dataset. Through this analysis, six main influencing factors were identified.

(1) The tier of a country in the Olympic Games is often strongly correlated with the number of medals it wins. Once the tier of a country is determined, the number of medals it wins tends not to change significantly.

A country category index, denoted as C_i, is defined to measure a country's overall medal-winning ability. The category assignment for each country is determined based on the clustering results discussed below.

(2) This can effectively reflect the competitive strength of athletes and the country's medal potential. For athletes participating in the competition, if they have previous medal experience, it indicates they are world-class competitors. Therefore, the more historical medals athletes from a country have in the year of the competition, the greater the country's potential to win medals. The different meanings of gold, silver, and bronze medals are also taken into consideration.Given that gold signifies the highest level of achievement and silver and bronze are of lesser value, different weights are accordingly allocated to each medal type. For a country i, the total historical medal count of its athletes is:

$$M_{athl,i} = \sum_{j \in Country\ i} (\omega_1 \cdot M_{gol,j} + \omega_2 \cdot M_{sil,j} + \omega_3 \cdot M_{bro,j}) \tag{1}$$

(3) Although the total historical participation of athletes is not as important as their historical medal count, it can still reflect the countrys medal potential to some extent. The more competitions athletes have participated in, the more experienced they are. For a country i, the total historical participation of its athletes is:

$$T_{athl,i} = \sum_{j \in Country\ i} T_j \tag{2}$$

Among them, T_j represents the historical number of participations of athlete j.

(4) When athletes compete in a host country, it can bring a series of positive effects. The host country may boost athletes' confidence and morale. Additionally, the host country may be more inclined to select events that favor its strong sports. In contrast, athletes from non-host countries might face issues such as time zone differences, unfamiliar diets, and other acclimatization challenges,

which can affect their performance. These factors highlight that the host country effect often significantly enhances a nation's ability to win medals. A binary variable H_i is defined here:

$$H_i = \begin{cases} 1, \textit{if country i is the host nation,} \\ 0, \textit{otherwise.} \end{cases} \tag{3}$$

(5) The number of events is often positively correlated with a country's medal count. The more events there are, the higher the chances of winning medals. If there is only one event, the level of competition can be imagined. The total number of events is denoted by E_i which is the same for each country in every competition.

(6) Competitions have uncertain factors, such as unexpected upsets by seeded athletes or the rise of underdogs. A random disturbance term is defined to simulate the uncertainty in the competition, which follows a normal distribution: $\varepsilon_i : N(0, \sigma^2)$.

The above factors represent those identified as influencing the number of medals a country wins. To calculate the category factor for each country, this study has established the following clustering model.

2.2. Classification of Different Types of Countries Based on Comprehensive Index

Through analysis of the complete medal counts of all Summer Olympics from 1896 to 2024 and comparison of the total historical medals, significant disparities between countries have been observed. The medal-winning capabilities of each country are assessed and classified based on the following three aspects.

(1) The total number of medals held by a country is the most direct indicator of its ability to win medals. The more medals, especially gold medals, a country holds, the stronger its medal-winning capability. For each country, let $M_{gol,i}$, $M_{sil,i}$, and $M_{bro,i}$ represent the historical total number of gold, silver, and bronze medals, respectively. The weighted total medal count is:

$$M_{total,i} = \omega_1 \cdot M_{gol,i} + \omega_2 \cdot M_{sil,i} + \omega_3 \cdot M_{bro,i} \tag{4}$$

(2)The change in the number of medals often reflects whether a countrys medal acquisition is on the rise or decline. The larger the trend of increasing or decreasing medals, the faster the countrys medal count is changing.

Linear regression is employed to calculate the trend of each country's total medal count over time.The regression equation is $M_{i,t} = \alpha_i + \beta_i t + \varepsilon_t$.where α_i is the intercept, ε_t is the error term, and β_i is the slope, representing the change in the medal count over time. Let T represent the total number of years considered for the calculation, and use the least squares method to estimate β_i:

$$\beta_i = \frac{T\sum_t (t \cdot M_{i,t}) - \sum_t t \cdot \sum_t M_{i,t}}{T\sum_t t^2 - (\sum_t t)^2} \tag{5}$$

(3) Medal volatility is used to measure the stability of a countrys performance. High volatility often indicates instability, which may not be a good sign. The standard deviation σ_i is used to represent the volatility of the medals (where $\overline{M_i}$ is the average number of medals for country i over all years):

$$\sigma_i = \sqrt{\frac{1}{T}\sum_t (M_{i,t} - \overline{M_i})^2} \tag{6}$$

Finally, a comprehensive index is constructed to cluster countries using a weighted method:

$$Z_i = \omega_{total} \cdot M'_{total,i} + \omega_{slope} \cdot \beta'_i - \omega_{volatility} \cdot \sigma'_i \tag{7}$$

Where $\omega_{total}, \omega_{slope}, \omega_{volatility} > 0$. A negative weight is assigned to the volatility indicator σ'_i because high volatility typically indicates instability. Subsequently, all countries are clustered using the K-means method based on the above criteria. With the construction of these influencing factors, medal counts of countries in future Olympic Games can now be predicted based on the existing data.

2.3. Predictive Model Based on Random Forest

The number of medals does not usually have a linear relationship with the factors discussed above. To predict the medal count, a random forest regression model is adopted due to its ability to represent nonlinear relationships using decision trees. Suppose that prior to a given Olympic Games, historical data for N countries are available, represented as $D = \{(X_1, G_1), \text{L}, (X_N, G_N)\}$, where $X_i = (C_i, M_{athl,i}, T_{athl,i}, H_i, E_i, \varepsilon_i)$ and G_i is the corresponding gold medal count. Let there be M decision trees. The prediction formula of the random forest is the average value of the predictions from all the trees:

$$\hat{G}_i = \frac{1}{M} \sum_{m=1}^{M} f_m(X_i) \tag{8}$$

Based on existing data, silver and bronze medal counts are similarly predicted, and the results are denoted as $\hat{S}_i$ and $\hat{B}_i$. The predicted total medal count is then:

$$\hat{T}_i = \hat{G}_i + \hat{S}_i + \hat{B}_i \tag{9}$$

For years with known total medal count data, the Mean Absolute Percentage Error (MAPE) between predicted and actual values is calculated as follows:

$$MAPE = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{\hat{T}_i - T_i}{T_i} \right| \tag{10}$$

Due to many uncertain factors in the competition, to better describe the prediction results, the prediction of a single value can be converted into a prediction interval:

$$[\hat{T}_{L,i}, \hat{T}_{R,i}] = [\hat{T}_i \cdot (1 - MAPE), \hat{T}_i \cdot (1 + MAPE)] \tag{11}$$

Based on the above formulas, a prediction interval for each country's medal count in future Olympic Games can be provided. Prior to this, it is necessary to ensure that all required information for the 2028 Olympics, such as the values of various factors for the participating countries, is available.

Among the above factors, country categories and host country status are clearly defined for the 2028 Olympics. However, since athlete information for the 2028 Games is not yet available, historical medal counts or total athlete participation cannot be directly calculated. Similarly, the total number of events in the 2028 Olympics remains unknown, necessitating the development of predictive models for this data. For instance, because the total number of events exhibits neither strong seasonality nor autocorrelation and the dataset is limited, polynomial regression is selected to forecast future values.

2.4. Prediction of 2028 Indicators Based on Polynomial

Polynomial regression extends the modeling of nonlinear relationships. Let Y_t denote the observation at time t, it can be described using an n-degree polynomial as follows:

$$Y_t = \mu_0 + \mu_1 t + \text{L} \ \mu_n t^n + \varepsilon_t \tag{12}$$

where $\mu_0, \mu_1, \mathrm{L}\ , \mu_n$ are the polynomial coefficients to be estimated, and ε_t represents the random error, usually assumed to follow a normal distribution. The coefficients are estimated using the least squares method:

$$s(\mu) = \sum_t [Y_t - (\mu_0 + \mu_1 t + \mathrm{L}\ \ \mu_n t^n)]^2 \tag{13}$$

The solution for the estimated parameters is:

$$\hat{\mu} = (X^T X)^{-1} X^T Y \tag{14}$$

Finally, historical medal counts, total participation, and the total number of events for future years can be predicted. Once all necessary input data are obtained, the previously established random forest model can be used to forecast the medal tally for upcoming Olympic Games.

3. Medal Prediction Model Based on Random Forest

3.1. Country Clustering Results

Using comprehensive indicators, a K-means algorithm was developed to classify countries. Figure 1 displays the clustering results for all participating countries.

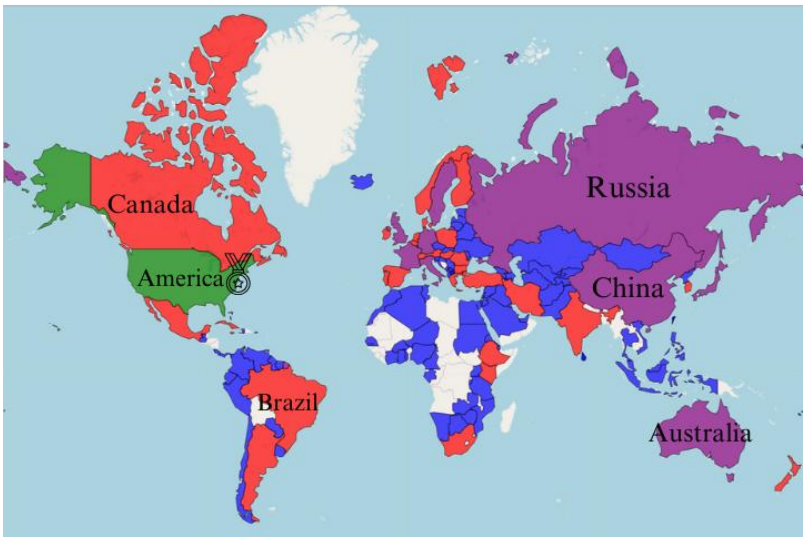

Figure 1. Classification map of different categories of countries

Based on the results of clustering analysis, the United States is categorized as a separate group, highlighting its leading position in the history of the Olympics. This position is maintained because the United States has never missed an Olympic Games since its inception, and its performance has been stable in recent Olympic Games.

Countries like China and Russia are categorized in the second tier. China only resumed participation in the Olympics in 1980 and made significant progress after 2000. Russias performance has been even more volatile, particularly due to the dissolution of the Soviet Union and suspensions in the past two Olympic Games, which have affected its position in the medal rankings.

Countries like Canada and Brazil belong to the third tier. These countries have more unstable Olympic results, either due to infrequent participation or, despite competing, they rarely win medals. There are also countries that participated but did not win any medals.

3.2. Polynomial Regression Calculation of Input Features

Apart from national categories, several other factors influence the number of medals. Polynomial regression was employed to estimate the values of the following input features.

The total number of events is measured by counting all events included in the medal tally for each Olympic Games.

The historical medal count of athletes in the competition year is obtained by identifying athletes who have previously won medals, summing the total number of medals they earned, and then aggregating these values by the country each athlete represents.

The historical participation count of athletes is calculated by summing the number of times athletes from each country have participated in past Olympic events, based on the records from the athlete dataset.

3.3. 2028 Olympic Medal Table Prediction

Since 1960, the number of Olympic events has gradually increased, especially with more team events and women's events, which has significantly changed the medal distribution pattern of each Olympic Games. During the same period, the global level of sports has greatly improved, and athletes from many countries have shown more outstanding performances in the Olympic Games, leading to an increase in the number of medals and participants. Therefore, more recent data is crucial for predicting future Olympic medal distributions.

In conclusion, data after 1960 provides information that better reflects current sports trends. As a result, these data points are assigned higher weight in modeling to more accurately reflect the medal distribution of future Olympic Games. The predicted data for the 2028 Olympics is shown in the Table 1.

Table 1. Olympic Medal Statistics for 2028

Year	NOC	Num Medalists	Total Medals	Total Participations	Total Events	Category
2028	CHA	49	72	242	331	3
2028	DEN	11	16	87	331	0
2028	NED	31	42	156	331	0
2028	FIN	4	5	64	331	0
2028	NOR	9	11	49	331	0
2028	ROU	20	34	128	331	0
2028	EST	2	2	23	331	4
2028	FRA	39	51	264	331	3

Subsequently, training is conducted using data from 1896 to 2024 with a random forest model. This model incorporates multiple influencing factors, including country classification, historical medal counts, and athlete participation numbers, when making splits. If a particular feature, at a certain threshold, leads to the greatest impurity reduction, it is selected for splitting at that node. Each decision tree then produces a prediction for the number of medals, and the random forest aggregates these predictions by averaging them. This ensemble learning approach is more robust than relying on a single decision tree and effectively reduces the risk of overfitting. Polynomial regression was used to predict the historical medal count, total participation, and total number of events for athletes in the year 2028. By combining this data with country classification and host country status, these values were input into the previously trained model to obtain predicted medal counts for various countries at the 2028 Los Angeles Olympic Games, as shown in Table 2.

Table 2. 2028 Olympic Medal Table

NOC	Gold	Silver	Bronze	Total
USA	[40,44]	[39,42]	[37,41]	[117,127]
CHN	[28,32]	[30,33]	[23,26]	[81,91]
GER	[23,26]	[21,25]	[20,23]	[64,74]
RUS	[18,22]	[14,17]	[14,17]	[49,59]
ITA	[16,20]	[16,19]	[13,16]	[42,53]
GBR	[14,17]	[12,15]	[11,14]	[36,47]
FRA	[13,17]	[13,16]	[10,14]	[37,47]
AUS	[13,17]	[18,21]	[20,24]	[51,61]

The United States is predicted to win 46 gold medals in 2028, compared to 40 in 2024, showing a noticeable increase. With the added advantage of being the host country, the U.S. is expected to perform better in 2028.

China is predicted to win 40 gold medals in 2028, with a decrease from 40 to 32. It is expected to perform worse than in 2024. However, despite this decrease, China still maintains an advantage in many sports.Additionally, the United Kingdom and Germany are expected to perform better, while Russia and Italy may perform worse than before.

3.4. Evaluation of Prediction Model Accuracy

To assess the accuracy of the model's predictions, the R^2 coefficient is used by applying the model to the known medal data from 2024:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(T_i - \hat{T}_i)^2}{\sum_{i=1}^{N}(T_i - \bar{T})^2} \tag{15}$$

Additionally, the MSE and MAE metrics are introduced to evaluate the model's performance, with their values presented in Table 3.

Table 3. Error Metrics

Name of the error	Value
MSE	[1.4084, 0.67395, 0.72165, 72.197]
MAE	[0.36374, 0.30370, 0.33237, 2.7117]
R^2	[0.93588, 0.96743, 0.96554, 0.83254]

From the evaluation indicators in Figure 2, it is evident that the model performs relatively well in predicting gold, silver, and bronze medals, but there is a larger error in predicting the total number of medals. This is because the total medal count is derived by summing gold, silver, and bronze medals, which causes the errors in the first three categories to accumulate, leading to a larger error in the total medal count. Furthermore, if the medal count for certain countries fluctuates more complexly, this could make predicting the total medal count more challenging.

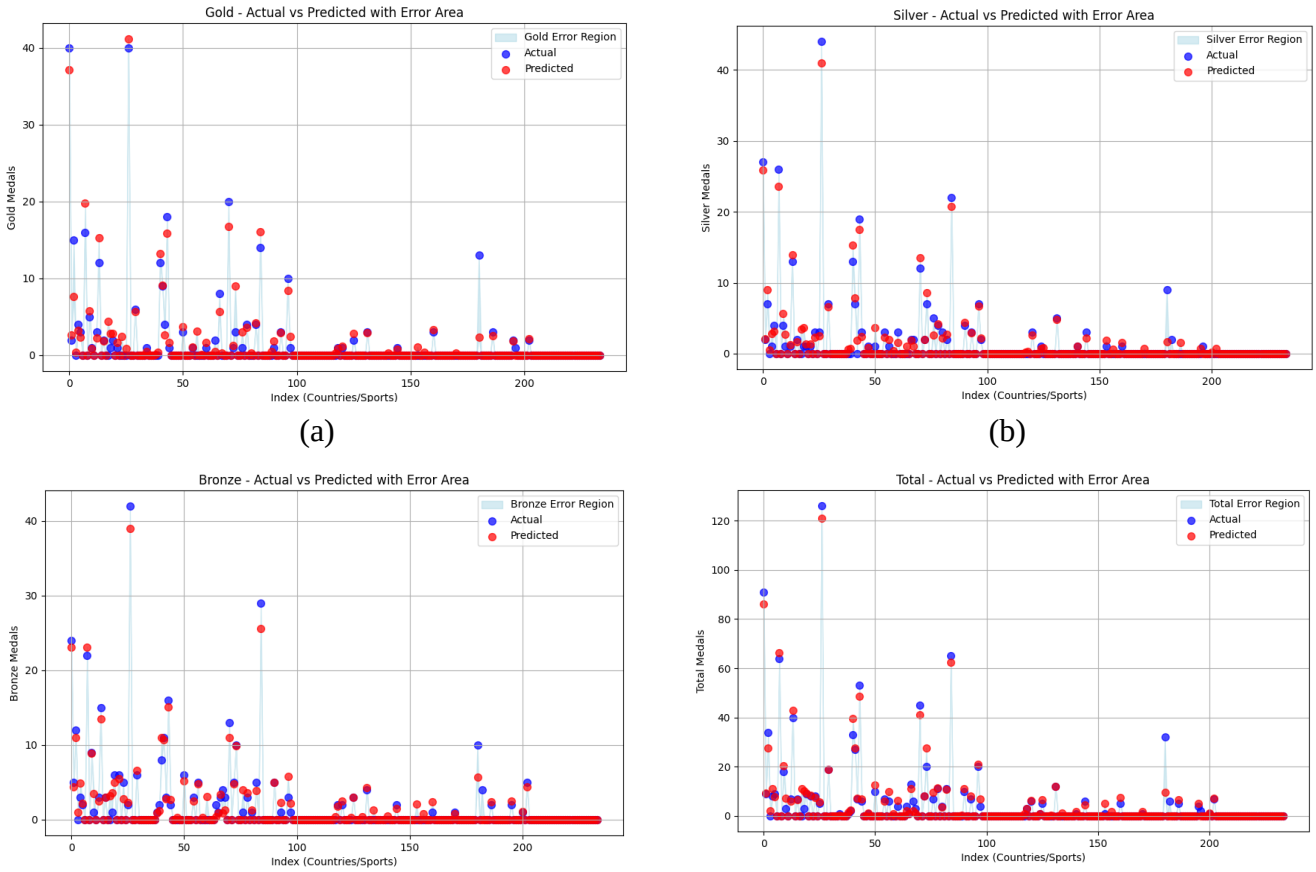

Figure 2. Prediction and Actual Medal Count Error

4. Conclusions

This study presents a comprehensive framework for predicting Olympic medal counts by integrating random forest regression and polynomial regression methods. Based on an in-depth analysis of historical Olympic data, six key influencing factors were identified: country classification, historical medal achievements, athlete participation, host nation status, number of events, and competition uncertainties. A clustering model was used to categorize countries according to their Olympic competitiveness, laying the groundwork for more accurate modeling.

To support prediction for the 2028 Olympic Games, polynomial regression was employed to forecast unavailable input variables such as the number of events and athletes' historical participation. Subsequently, these values were fed into the trained random forest model to generate predictions. Experimental results demonstrate the model's strong performance across different medal types. Evaluation metrics for gold, silver, bronze, and total medal counts yielded low error rates: MSE values of [1.4084, 0.67395, 0.72165, 72.197], MAE values of [0.36374, 0.30370, 0.33237, 2.7117], and R^2 values of [0.93588, 0.96743, 0.96554, 0.83254], respectively. These results indicate high predictive accuracy and robustness, particularly for individual medal types.

In the future, this modeling approach could be further enhanced by incorporating real-time athlete performance data, country-specific investment in sports, and event-level competitive dynamics. Moreover, the methodology established in this paper can serve as a reference for predicting outcomes in other large-scale sports competitions such as the Asian Games or World Championships, thereby expanding its practical applications beyond the Olympic Games.

References

[1] Forrest D, Sanz I, Tena J D. Forecasting national team medal totals at the Summer Olympic Games[J]. International Journal of Forecasting, 2010, 26(3): 576-588.

[2] Peters T. Winning against the odds: A socio-economic analysis of Olympic success[D]. Erasmus University, 2023.

[3] Schlembach C, Schmidt S L, Schreyer D, et al. Forecasting the Olympic medal distribution during a pandemic: a socio-economic machine learning model[J]. arXiv preprint arXiv:2012.04378, 2020.

[4] Bunn J, Reagor M K, Myers B J. Creating a random forest model to determine success in women's collegiate lacrosse[J]. Journal of Sport and Human Performance, 2025, 13(1): 1-9.

[5] Zhao K, Du C, Tan G. Enhancing basketball game outcome prediction through fused graph convolutional networks and random forest algorithm[J]. Entropy, 2023, 25(5): 765.

[6] Alnahit A O, Mishra A K, Khan A A. Stream water quality prediction using boosted regression tree and random forest models[J]. Stochastic Environmental Research and Risk Assessment, 2022, 36(9): 2661-2680.

[7] Guo Z, Yu B, Hao M, et al. A novel hybrid method for flight departure delay prediction using Random Forest Regression and Maximal Information Coefficient[J]. Aerospace Science and Technology, 2021, 116: 106822.

[8] Liu J, Liang D, Cho H. A polynomial regression model for predicting knuckleball movements in soccer free-kick[J]. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, 2025: 17543371241311676.

[9] Belany P, Hrabovsky P, Sedivy S, et al. A comparative analysis of polynomial regression and artificial neural networks for prediction of lighting consumption[J]. Buildings, 2024, 14(6): 1712.

[10] Tabelini L, Berriel R, Paixao T M, et al. Polylanenet: Lane estimation via deep polynomial regression[C]//2020 25th international conference on pattern recognition (ICPR). IEEE, 2021: 6150-6156.