Prediction Of Medals Based on Machine Learning and OLS Statistical Regression Models

Jianing Zhang, Qifei Yan, Lingge Shi *, Xize Chen

School of pharmacy, Shenyang Pharmaceutical University, Shenyang, China, 110016

* Corresponding Author Email: shilingge9@gmail.com

Abstract. This paper constructs a predictive model that integrates variables such as the number of athletes and the growth rate of medals. It employs regression and classification techniques from the field of machine learning, utilizing algorithms including Random Forests, Support Vector Machines, and Logistic Regression and the performance of random forest in the last three Olympic Games was compared to determine the optimal performance of random forest in the classification task. Additionally, it applies Bayesian Change Point Detection and Ordinary Least Squares statistical regression models to analyze the performance trajectories of the Chinese women's volleyball team, the United States women's volleyball team, and the Romanian gymnastics team. Based on the data predicted by the model, the study proposes to increase investment in sports science research, attract talent, strengthen logistical support, and develop special policies. At the same time, error analysis and sensitivity analysis were carried out in this study, and the limitations in the data and model and the sensitivity of the model to certain features were pointed out.

Keywords: Olympic Games; Medal Prediction; Machine Learning Models; OLS Statistical Regression Model.

1. Introduction

As the largest and most influential sporting event globally, the Olympic Games' competition events are determined by the International Olympic Committee (IOC), while the host nation also wields a degree of influence [1], concurrently, as the host of the 2028 Los Angeles Olympics, the United States has incorporated baseball/softball, six-a-side lacrosse, cricket, squash, and flag football into the sports program for the 2028 Olympic Games [2]. Having the capacity to propose the inclusion of events that showcase its unique characteristics or strengths. However, due to the absence of certain historical data [3], rendering medal prediction particularly challenging. To address this issue, this study designs a model that integrates machine learning algorithms with OLS Statistical Regression Model, aiming to predict future Olympic medal standings, including those for 2028.

Cao et al. investigated the relationship between urban nighttime lighting and regional development disparities using an OLS statistical regression model [4]. Building upon this, Liu et al. employed the same method to further analyze the regional disparities in China's economic development and their characteristics of stochastic convergence [5]. Additionally, you et al. utilized this model to examine the correlation between GDP and carbon emissions [6]. Chen et al. state that machine learning is at the core of artificial intelligence and is the fundamental approach to endowing computers with intelligence [7]. Zou has implemented an automated full inspection of PCB hole information using machine learning, enabling a streamlined production process [8]. Building upon this foundation, Wang et al. have also employed machine learning to achieve non-destructive, optically precise inspection of mechanical components [9]. Although numerous academic applications of machine learning models currently exist, due to the complexity of these models, there is presently no research on the application of machine learning to medal prediction.

To sum up,this paper constructs a model that employs machine learning and OLS statistical regression Model, leveraging historical Olympic medal standings and competition data to iteratively refine its predictions of Olympic medal counts for 2028 and beyond. The model aims to provide strategic guidance to organizers for optimizing competition activities.

2. Model building

2.1. Basic assumptions

Our team establish the reliability of the data to ensure the authenticity of historical Olympic data across nations and the credibility of predictions and evaluations. By analyzing the medal distribution among different countries, our team hypothesize that factors such as cultural differences, government support, and medical coverage, which are universally applicable across all nations, do not influence medal acquisition. Our team assume that the fundamental structure of the 2028 Los Angeles Olympics will remain consistent with previous years, with no significant changes in event categories or competition rules.

2.2. Modeling

We initially processed the raw data as follows: Firstly, organize all the country information to ensure data consistency between different tables. Secondly, specific delegations' splitting and merging scenarios were addressed, along with the corresponding variable adjustments.

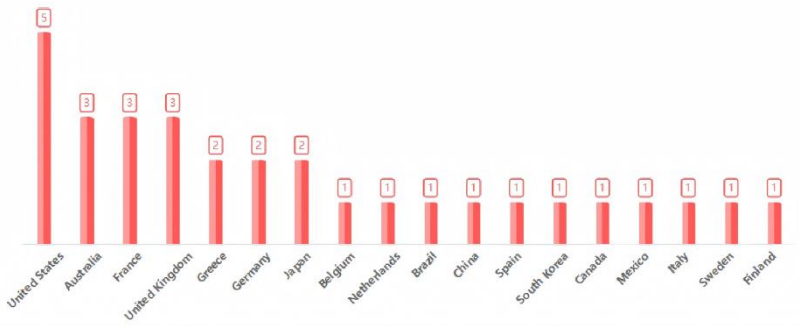

Figure 1. Count of host countries for the Summer Olympics

Figure 1 shows that the United States ranks first as the host country 5 times, which also establishes the United States as a sports power.

Summer Olympics are held every four years. To account for the timeliness of our analysis, we have established models using data from the most recent three editions (2016-2024) and the most recent ten editions (1988-2024) of the Games. This dataset encapsulates key information on these summer Olympic events. Specific distinctions have been made for certain "Special" Olympic delegations. For instance, in the 1992 Summer Olympics, a joint team composed of athletes from countries such as Russia and Belarus participated.In such a case, the data from these countries are correspondingly consolidated.

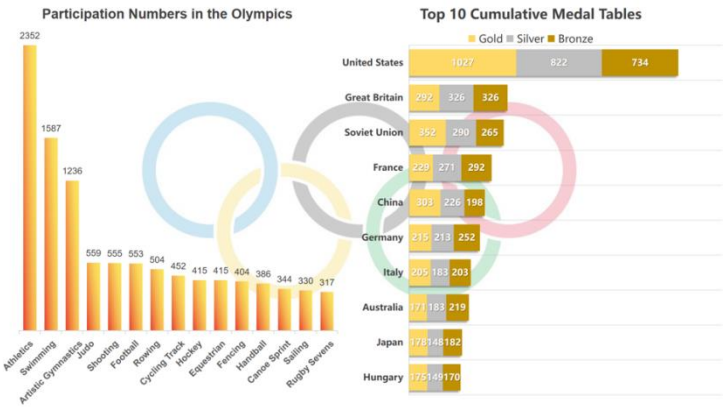

Figure 2. Participation numbers in the Olympics and top 10 cumulative medal tables

Figure 2 shows the number of participants in each event on the left, with track and field leading the way with 2,352 participants. On the right, the cumulative medal tally lists the top ten countries,

where the United States leads with 1,027 gold medals, showcasing both the scale of participation and competitive prowess in the Olympic Games.

We introduce the number of athletes, their growth rate, gold medals, the total number of gold medals, and their change rate to reflect the level of athlete participation and capture the trends in medal acquisition over time. Additionally, we flag the host country status.

To depict the long-term influence of national or regional underlying characteristics on their Olympic performance, such as temporary fluctuations in short-term achievements making the previous results incapable of fully reflecting their strengths, we incorporated team dummy variables. These primarily serve to illustrate the enduring impact of national or regional inherent traits on their Olympic outcomes. A partial data presentation is shown in Table1.

Table 1. The characteristic variables utilized in the medal prediction model

Variable Description	Definition	Data Type
Athlete_count	number of participants in the Olympic Games that year	continuous variable
Athlete_growth_rate	Growth Rate of Athlete Numbers	calculation result
Gold_X	Total number of gold medals over the past X Games.	continuous variable
Gold_growth_rate	Growth Rate of Gold Medals	calculation result
Total_X	Total number of medals over the past X Games.	continuous variable
Total_growth_rate	Growth Rate of Total Medals	calculation result
Is_Host	Whether the host country for the current year.	0/1 variable
Medal	Whether a medal has ever been won before.	0/1 variable
Assumption variables	variables related to the level of national sports development	continuous variable

It is worth mentioning that for the feature data “Assumption variables”, our team did not introduce a new data set, but instead used machine learning algorithms to optimize the medal data of each country each year and then added additional parameters.

Our team employ machine learning regression models to predict the number of gold medals and the total medal count. To quantify these uncertainties, we utilize quantile regression, training multiple models to predict different quantiles (10%, 50%, 90%), thereby providing confidence intervals for the number of gold medals and total medals for each country. The relevant formulas are presented in equation 1-2.

$$GoId\ Prediction\ Interval = [Gold_{0.1}, Gold_{0.9}] \tag{1}$$

$$Total\ Prediction\ Interval = [Total_{0.1}, Total_{0.9}] \tag{2}$$

Our team employ a machine learning classification model to predict whether a country will win a medal at the 2028 Olympics, with the model outputting the probability of each country winning a medal in 2028. We consider that when the probability value is less than or equal to 0.5, the country will not win a medal; when it is greater than 0.5, the country will win a medal.

Predictions for Olympic medals are often not based on past medal data but are influenced by factors such as athlete performance, event selection, and national strength[10]. Therefore, to avoid excessive reliance on historical medal data, we first used data from the most recent three editions for modeling. A partial data presentation is shown in Table 2.

Table 2. Example of characteristic data in the past three years

NOC	Year	Athlete_growth_rate	Gold_growth_rate	Total_growth_rate	Is_Host	Medal
USA	2024	-0.002336	0.025641	0.115044	0	1
USA	2020	0.190542	-0.152174	-0.066115	0	1
USA	2016	0	0	0	0	1

Our team use regression and classification models such as Random Forest, Support Vector Machines, and Logistic Regression, comparing their performance to determine the optimal model for practical prediction. The generation steps of the Random Forest are illustrated in Figure 3.

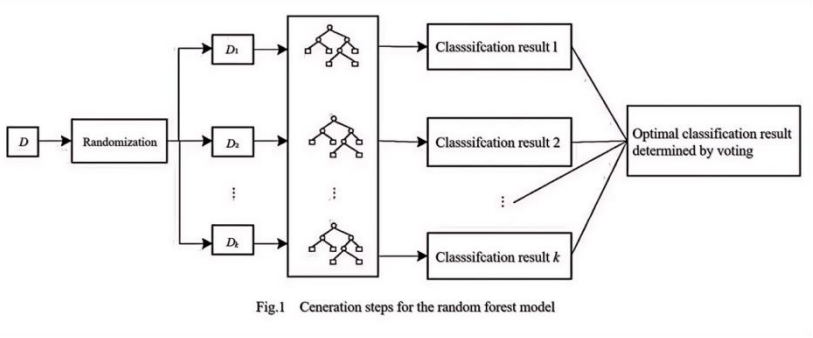

Figure 3. The generation steps of the Random Forest

Random Forest is an ensemble algorithm based on decision trees, which trains decision trees by sampling and synthesizes results for decision-making; Support Vector Machine classifies by finding the optimal hyperplane and commonly uses kernel functions to handle nonlinear problems; Logistic Regression is based on linear regression, mapping probabilities using logistic functions, and determining parameters through maximum likelihood estimation.

The VIM calculation method of the random forest routine has the Gini index, and the score statistics of the variables can be expressed. For the relevant formulas, see equation 3.

$$GI_m = \sum_{k=1}^{K} \hat{p}_{mk} (1 - \hat{p}_{mk}) \tag{3}$$

The importance of the variables in node m, meaning that the Gini index changes before and after the node m branch is given in equation 4

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r \tag{4}$$

GI_l、GI_r represents the Gini index of two new nodes split by node m.

The Gini importance of variables in random forests is defined in equation 5

$$VIM_j^{(Gini)} = \frac{1}{n} \sum_{i=1}^{n} V\,IM_{ij}^{(Gini)} \tag{5}$$

For information gain, assuming that we split dataset D on feature A, the information gain is defined in equation 6.

$$Gain(D, A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} Entropy(D_v) \tag{6}$$

Initially, regression models were trained using Random Forest and XGBoost, with the model evaluation parameters presented in Table 3.

Table 3. Training data for Random Forest and XGBoost

	MAE	MSE	RMSE	R^2
Random Forest	1.5410	28.1274	5.3035	0.8008
XGBoost	1.8186	32.5246	5.7030	0.7696

Next, the classification model was trained using Random Forest, Logistic Regression, and XGBoost, with the confusion matrix illustrated in Figure 4.

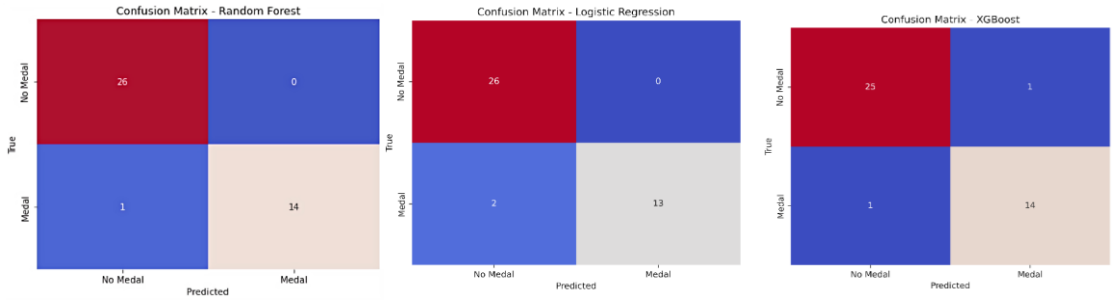

Figure 4. Confusion matrix of Random Forest, Logistic Regression, and XGBoost

In terms of performance in classification tasks, Random Forest, Logistic Regression, and XGBoost each have their distinct characteristics. Random Forest performs the best, achieving an accuracy of 0.9756 and an AUC of 0.9667. Both Logistic Regression and XGBoost have an accuracy of 0.9512; however, Logistic Regression has an AUC of 0.9333, while XGBoost has an AUC of 0.9474. Overall, Random Forest holds a leading position in the classification performance demonstrated by this set of data.

Our team observed that the R-squared value is not high when fitting with features from the last 3 sessions. Therefore, we plan to utilize data from the last 10 sessions for further fitting to achieve a higher degree of fit. This approach ensures a high goodness of fit while minimizing reliance on historical data as much as possible. The adjusted evaluation data is presented in Table 4.

Table 4. Training data for Random Forest and XGBoost over the last 10 sessions

	MAE	MSE	RMSE	R^2
Random Forest	0.4250	5.4407	2.3325	0.9757
XGBoost	0.4128	4.6658	2.1600	0.9792

It can be seen that after combining the data of the past decade, the model shows a high goodness of fit. Therefore, when we predict the medals of the 2028 Olympics, we will combine the data of two different models. This ensemble approach leverages the strengths of both models, potentially leading to a higher overall performance.

2.3. OLS statistical regression model

It often takes a certain amount of time for a coach to have an impact after their appointment. Therefore, we consider the influence generated by their tenure, for example, the impact of a coach appointed in 2000 begins to manifest in 2004. For each set of data, we establish an OLS regression model separately to examine the effect of the coach on medal acquisition. A partial data presentation is shown in Table 5.

Table 5. Chinese volleyball OLS regression results

Dep.Variable	Model	Method	R-squared	AIC	BIC	coef
y	OLS	Least Squares	0.260	49.54	30.85	7.1500

The data reveals that COEF is positive, with Lang Ping impact on China's volleyball medal effect being 7.1500, indicating that she has indeed played a positively contributive role in Chinese volleyball. A partial data presentation is shown in Table 6.

Table 6. American volleyball OLS regression results

Dep.Variable	Model	Method	R-squared	AIC	BIC	coef
y	OLS	Least Squares	0.115	54.95	55.75	7.8889

Likewise, a positive COEF value of 7.8889 indicates that Lang Ping's impact on the medal performance of the US volleyball not only positively influenced China but also had an even greater effect on the United States. A partial data presentation is shown in Table 7.

Table 7. Romanian gymnastics OLS regression results

Dep.Variable	Model	Method	R-squared	AIC	BIC	coef
y	OLS	Least Squares	0.118	65.34	65.31	5.611

Bela Karolyi's influence on Brazilian gymnastics has also demonstrated a positive impact, with a medal effect of 5.611.

It is evident that engaging the"Great Coach" can indeed effectively enhance the country's medal count. We employ the CUSUM method to identify countries whose medal sequences have been relatively stable over the past three years. The calculation formula found in equation 7-8.

$$C_t^+ = max\left(0, C_{t-1}^+ + (x_t - \mu)\right) \tag{7}$$

$$C_t^- = min\left(0, C_{t-1}^- + (x_t - \mu)\right) \tag{8}$$

Based on our analysis, we recommend that the U.S. men's gymnastics, Italian women's volleyball, and Russian badminton teams consider hiring exceptional coaches to achieve breakthroughs and improve their national standings, with the aim of securing more medals.

3. Results

3.1. Actual prediction results of the model

First, the classification model is used to predict whether a country will win the award. The probability of winning for some countries is shown in the Table 8.

Table 8. Predicted prob in several countries

NOC	Predicted_prob	NOC	Predicted_prob
TPE	0.98	UAE	0
TTO	0.01	UGA	0.36
TUN	0.92	UKR	1
TUR	1	URU	0
TUV	0	USA	1

For the prediction of countries likely to win awards, we employ a random forest regression model to forecast both the number of gold medals and the total number of medals separately, and employed quantile regression to provide a prediction interval. The ultimate partial prediction results are presented in Figure5.

Figure 5 Top ten predictions for the medal table

It is evident that the United States will remain at the top of the medal table, further widening the gap, thus continuing to solidify its position as a dominant sporting power.

We will evaluate the changes in the total number of gold medals and overall medals between 2024 and 2028 to gauge the progress and decline of each country. The countries with increases in both metrics are classified as progressive countries, while the rest are designated as regressive countries. Based on this criterion, we project 42 progressive countries and 29 regressive countries. Specific data are presented in Table 9.

Table 9. Detailed data on advancements and declines in some countries

NOC	Gold Change	Total Change	NOC	Gold Change	Total Change
USA	19.68	84.89	AUT	-9.44	-29.2
CHN	16.09	48.84	CIV	-21.26	-66.55
JPN	11.58	36.44	NOR	-8.04	-20.25
RSA	1.38	0.03	UZB	-18.96	-89.83

In the 2024 Paris Olympics, nations achieving their first-ever Olympic medals included Dominica, Saint Lucia, Albania, and Cape Verde. In the 2020 Tokyo Olympics, nations securing their inaugural Olympic medals were San Marino, Burkina Faso, Turkmenistan, Bermuda, Qatar, and the Philippines. The model predicts with 67.32% certainty that, by 2028, 4 countries will win their first-ever medals. Figure 6 depicts the number of countries that have won new medals over the past ten games.

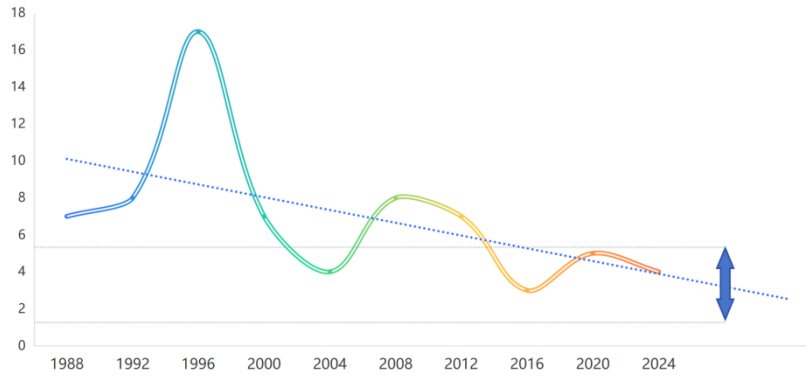

Figure 6. Number of countries that have won medals in the past ten games

To explore the relationship between different event categories and variations in numbers, we conducted a correlation analysis on the changes in the number of participants and the number of medals for each event across different years. The Pearson Correlation coefficient was employed to measure this correlation, with the relevant formula presented in equation 9.

$$\text{Gorrelation } (X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2} \quad (9)$$

The correlation between the variation in the number of events over the past three years and the number of medals is 0.03, and over the past decade, it's 0.09. This might be due to the introduction of new events in the Olympics (such as 3x3 basketball and skateboarding), which provides more countries the opportunity to win medals in these emerging disciplines, thereby enabling some nations to enhance their positions on the medal table and alter the existing distribution of medals.Following this, we constructed a correlation heatmap illustrating the relationship between projects and awards.

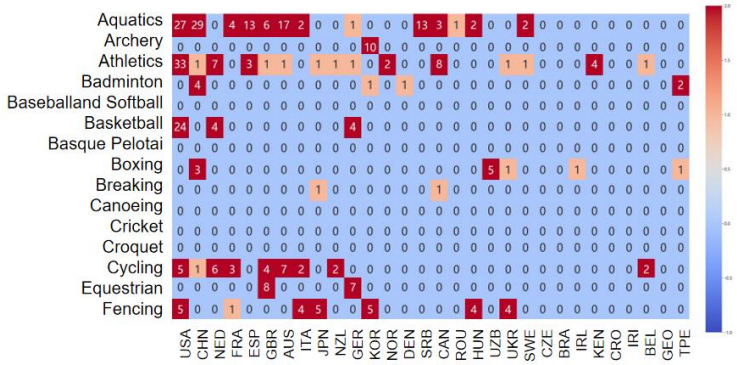

Figure 7. Project correlation heat map

As indicated by the thermodynamic chart in Figure 7, significant disparities exist in the number of medals awarded across various competition categories among different nations. This study observed that certain nations display superior medal performances in specific disciplines, closely aligning with their sporting traditions and specialties. For instance, the United States excels in numerous fields, particularly in track and field, aquatics, and basketball. These sports boast long-standing traditions and superior facilities in the country, enabling it to amass substantial medals in these categories.

Next, we will study whether the host country will win more medals. The average medals of host countries and non-host countries are shown in Table10.

Table 10. Average medals by host country

	host country	non-host country
Average	49.6667	13.6716

It is quite evident that the host country has exhibited a significant advantage, possibly due to its leverage in shaping the Olympic program for that year, allowing it to propose, within certain limits, the inclusion of events that reflect its unique cultural traits or sporting strengths.

We estimate that there will be progress in 42 countries and regression in 29 countries for the 2028 Olympics. Additionally, we estimate with a probability of 67.32% that four new teams will win their country's first Olympic medal in 2028.

The number of events does impact the tally of national medals, though to a limited extent; what correlates highly with medal counts, however, is the variety of events. Additionally, host countries often enjoy an added advantage in event selection, allowing them to secure more medals in their home Olympics.

4. Conclusions

The model constructed in this study comprehensively integrates historical data, the quantity and diversity of events, host country effects, and other influencing factors while maintaining broad applicability across various competitions and regions. By incorporating a large volume of historical data, the model is capable of continuous learning and self-improvement. Additionally, the model incorporates machine learning algorithms such as regression and classification models to enhance the accuracy and comprehensiveness of predictions. Furthermore, the model's flexibility has been enhanced, enabling it not only to predict medal counts for traditional Olympic powerhouses but also to assess the likelihood of other nations securing their inaugural medals.More importantly, this model is not specifically designed for predicting Olympic medals. Its high flexibility makes it widely applicable to various future sports events, including but not limited to the World Cup and the Asian Games. By adjusting features and data sources, it can provide strategic recommendations for national sports investment decisions.

References

[1] Olympic Games - Summer, Winter Olympics, YOG & Paralympics, https://www.olympics.com/en/olympic-games

[2] Olympic Sports, https://la28.org/en/games-plan/olympics.html#la28-main-footer

[3] IOlympic Results, Gold Medalists and Official Records, https://www.olympics.com/en/olympic-games/olympic-results

[4] Cao Ziyang, Wu Zhifeng, Kuang Yaoqiu, et al. Calibration and Application of DMSP/OLS Nighttime Light Imagery in the China Region [J]. Journal of Geo-Information Science, 2015, 17(09): 1092-1102.Olympic Results, Gold Medalists and Official Records, https://www.olympics.com/en/olympic-games/olympic-results

[5] Liu Huajun, Du Guangjie. Regional Disparities in China's Economic Development and Test of Stochastic Convergence: Based on DMSP/OLS Nighttime Light Data from 2000 to 2013 [J]. The Journal of Quantitative & Technical Economics, 2017, 34(10): 43-59. You Ning, Han Libo, Li Shiqiang, et al. Per Capita Energy Consumption Carbon Emissions in the Urban Agglomeration of Central Yunnan Based on DMSP/OLS-NPP/VIIRS Nighttime Light Data [J]. Journal of Lanzhou University (Natural Sciences), 2024, 60(06): 764-772+781.Chen Yongtao, Guo Xiaoying, Tao Huijie. A Brief Discussion on the Current Research Status and Development Trends of Machine Learning [J]. China New Telecommunications, 2018, 20(08): 173.

[6] Zou Huadong. On-line Optical Inspection of PCB Hole Information Based on Machine Learning[J]. Journal of Liaoning Technical University (Natural Science Edition), 2012, 31(01): 93-97.

[7] Wang Qi, Tan Juan. Optical ultra-precision detection technology based on artificial intelligence technology [J]. Laser Magazine, 2021, 42(02):156-160BERNARD A B, BUSSE M R. Who wins the Olympic Games: Economic resources and medal totals[J]. Review of Economics and Statistics, 2004,86(1): 413-417