World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3047

MEDAL PREDICTION FOR THE LOS ANGELES OLYMPIC GAMES BASED ON BAYESIAN OPTIMIZED BOOST REGRESSION

YiHan Gong
School of Statistics and Data Science, Qufu Normal University, Qufu 273165, Shandong, China.
Corresponding Email: yhgong1203@163.com

Abstract: To support resource allocation and strategic preparation for the 2028 Olympic Games, this study introduces a Bayesian-optimized Boost framework. First, key variables are extracted through feature engineering, including economic indicators such as GDP, population, host country effects, and recent rule changes in events. Next, Bayesian optimization is used to fine-tune the model's hyperparameters, and multiple metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) are employed to evaluate the model's performance. Then, time-series cross-validation based on 2024 data shows that the model achieves an R^2 of 0.91, outperforming baseline models such as decision trees and standard Boost. Predictions indicate that 23 delegations will gain more medals, while 46 countries may see a decline. Finally, k-means clustering is used to identify each country's dominant sports, quantify the impact of the host country effect on event selection, and provide data support for preparation.
Keywords: Olympic medal forecasting; Feature Engineering; Bayesian Optimization; k-means

1 INTRODUCTION

Olympic medal tallies mirror competitive strength; accurate forecasts aid strategic preparation. Multidimensional machine-learning models dominate current research yet still lack completeness and precision.

Early Olympic-medal forecasts centered on macroeconomics and demographics: Bernard & Busse [1] used population and GDP to show real GDP is a key medal predictor, and Zhang [2] confirmed higher GDP boosts medal odds via better facilities and coaches, yet both assumed linearity and ignored event specifics and host effects. Tian [3] emphasized event popularity, technical thresholds and rule changes as medal drivers, opening a new lens. Recent domestic work by Shi, Shi & Zhang [4] leveraged interpretable ML to show table tennis and badminton are highly predictable owing to top-team dominance, whereas football and shooting are not, but they still omitted event-host interactions. Luo & Cheng [5] demonstrated hosts can tilt outcomes by adjusting event schedules and environments, an interaction absent from traditional models. Machine-learning advances-e.g., Sadu et al.'s [6] Boost ensemble-integrate economic, demographic and historical data via non-linear, high-dimensional learning, outperforming linear regressions, yet Boost's efficacy hinges on hyperparameters; grid or random search is computationally heavy and prone to local optima, curtailing its full potential.

Addressing prior omissions, this study integrates economic, event-specific and host-region factors within a Bayesian-optimized Boost framework to forecast the 2028 Los Angeles medal distribution and deliver refined guidance for strategic preparation.

2 RESEARCH METHODS

2.1 Data Acquisition and Preprocessing

The data used in this study are drawn from the publicly available datasets released by the International Olympic Committee (IOC), covering the complete competition records from the 1896 Athens Games through the 2024 Paris Olympics.

2.1.1 Data cleaning

(1) Events that are for demonstration purposes are excluded from the model.

(2) Since 1924, the sports "Skating" and "ce Hockey" have been removed from the Summer Olympics. Out of respect for historical facts, these two sports are also excluded from the model.

(3) To ensure the rigor and scientific integrity of the model, this study remove the data from the 1906 Athens Olympics. Although this edition of the Games represented a noble pursuit of the Olympic spirit, it was not officially recognized as part of the Olympic series, and its results were not officially recorded. Thus, it has significant deficiencies in terms of data continuity and comparability [2].

2.2.2 Data merge

(1) For anomalies in the dataset ″summer Oly-athletes″ caused by changes in event names, such as ″Region1-1″ and ″Re gion1-2,″ this study unified them under the country code ″Region1.″

© By the Author(s) 2025, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

(2) During the integration of multiple datasets, this study found discrepancies between some athletes′ medal records and the national medal allocation. Based on the principle that each athlete can only win one medal per event, we recalculated the historical medal counts for each country to ensure data accuracy and consistency.
(3) The data were merged to count the number of athletes and the number of events for each country in each Olympics, without distinguishing between participating teams.

2.2 Method Introduction

Guided by medal prediction, this study follows a "clean–model–validate" pipeline: preprocess data, tune and optimize models per task, cross-validate, then forecast the 2028 Los Angeles Games and assess effects.
(1) Boost Regression Model Based on Bayesian Optimization
This model is the core method used in this study for predicting the number of Olympic medals. Its core idea is to capture the nonlinear relationships in the data through Boost regression and combine it with the Bayesian optimization algorithm to improve model performance. Boost regression iteratively constructs multiple decision trees as weak learners, and each tree fits the residuals of the current model to gradually optimize the prediction effect. Its goal is to minimize the MSE, thereby reducing the deviation between predicted values and true values. Bayesian optimization estimates the objective function of the hyperparameter space by constructing a surrogate model, efficiently searching for the optimal combination of hyperparameters. Compared with traditional grid search or random search, it can more accurately improve the generalization ability of the model. The advantages of this model are: first, it is suitable for prediction tasks with multiple features and complex nonlinear relationships, and can effectively integrate multi-dimensional features such as the host country effect, the number of participating events, and the number of athletes; second, it realizes automatic hyperparameter tuning through Bayesian optimization, reducing errors caused by manual intervention and improving prediction accuracy. Relevant studies have shown that similar gradient boosting models have high accuracy in Olympic medal prediction and can effectively capture the potential patterns in historical data[7].
(2) k-means Clustering Model
To explore the relationship between a country′s competitive strength and sports events, this study uses the k-means clustering algorithm to analyze medal data from the past three Olympic Games. This algorithm is an improved version of k-means clustering. It improves the stability and accuracy of clustering results by optimizing the selection method of initial clustering centers. In the clustering process, with countries, event categories, and medal results as key features, the data are divided into different clusters, thereby identifying different types of national groups such as ″focusing on specific events″, ″balanced development″, and ″strong in multiple events″. Its advantage is that it can intuitively reveal the distribution law of countries' advantages in sports events, providing a basis for analyzing the impact of the host country's event selection on medal distribution. Similar clustering methods have been proven effective in mining the laws of Olympic data[8].

2.3 Model Evaluation Metrics

To comprehensively evaluate the performance of the Olympic medal prediction model, MAE, MSE, RMSE, and R^2 are employed for quantitative analysis; detailed calculation formulas can be found in reference[9].

3 MODEL ESTABLISHMENT AND SOLUTION

3.1 Feature Engineering for Key Predictive Variable Extraction

Based on the results of exploratory data analysis (EDA), this study selected the following features as input variables for modeling.
(1) Host: Whether the country is the host nation The "host nation effect" was proven in exploratory analysis. Host countries usually enjoy a home advantage, which may lead to better performance in competitions. Additionally, the host country may be granted extra spots in certain events [9]. Therefore, this study set Host as a binary dummy feature, indicating whether the event is held in the athlete's home country. Host=1 represents the host country, and Host=0 represents a non-host country.
(2) Events: The number of events each country participates in. The number of events a country participates in reflects its level of activity and diversity in the Olympics. Coun- tries that participate in more events typically have stronger competition and more opportunities to win medals in various sports.
(3) Athletes: The number of athletes each country sends Although a total of 158,427 athletes participated in 292,942 events from 1896 to 2024, 78% of athletes never stood on the podium. 17% of athletes won only one medal during their Olympic careers, while the top 5% of athletes earned medals in multiple events. Previous Olympic prediction models did not use athletes' physical attributes, as these had low correlation with medal outcomes. Thus, we treat the total number of athletes in each country as a feature in our model, rather than focusing on individual athletes or teams. The number of athletes directly reflects a country's investment and talent pool for the Olympics. Countries with more athletes usually have more opportunities to win medals.

(4) Athletes-per-Event: The average number of athletes per event for each country. A country's resource allocation to sports may vary. By analyzing the average number of athletes per event, we can indirectly understand the country's focus and resource distribution in sports events, reflecting its investment in sporting competitions.
(5) Region-coded dummy variables: Categorical region identifiers are converted into numerical form through dummy encoding. A binary column is created for each unique code in the "NOC" column. For instance, if the "NOC" column contains Region 1, Region 2, and Region 3, new columns-NOC-1, NOC-2, NOC-3-are generated, where 1 indicates membership of the corresponding region and 0 indicates non-membership.

3.2 The Establishment and Case Analysis of the Olympic Medal Table Prediction Model

3.2.1 Bayesian optimization boost regression model

(1) Boost Regression Model Boost regression is a regression method based on gradient boosting trees. Boost regression iteratively constructs a strong regression model consisting of multiple weak regression models. Each weak regression model is a decision tree, and it improves the model's predictive ability by fitting the residuals of the current model. Boost regression optimizes the model by minimizing MSE, thus minimizing the difference between the predicted and true values.
(2) Bayesian Optimization of Model Hyperparameters Bayesian optimization is a method used to optimize model hyperparameters. The key to Bayesian optimization is using a surrogate model to estimate the objective function of the hyperparameter space, so that in each iteration, the hyperparameter combination most likely to improve performance is selected. Compared to traditional grid search or random search, Bayesian optimization can more efficiently explore the hyperparameter space, thereby finding better hyperparameter combinations. The main principle and process of the Bayesian optimization Boost regression model we developed are shown in Figure 1.

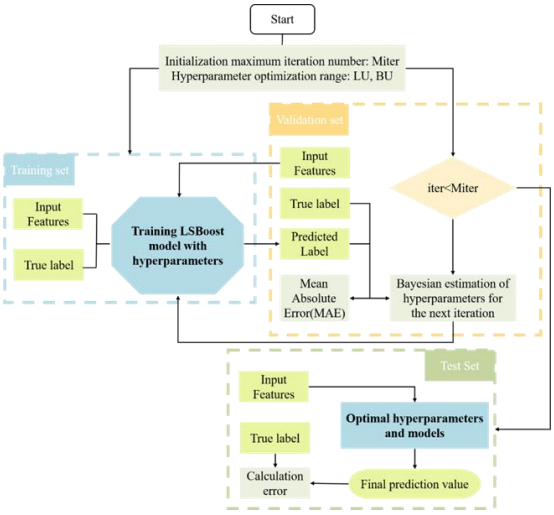

Figure 1 Bayesian Optimization Boost Model Flow Chart

(3) Model Evaluation We evaluate the model's performance using the test dataset, and we use the following performance metrics to assess the accuracy of the predictions: MAE, MAPE, MSE, RMSE, and R2.
(4) Probability Interval Prediction Metrics We use the following metrics to evaluate the accuracy of the probabilistic interval prediction model: PICP, PINAW, penalty coefficient CWC, MPICD, and AIS.

3.2.2 Model training and example analysis

This study decided to limit the training data to records from the 1992 Barcelona Olympics onwards, as there may have been significant differences before the dissolution of the
Soviet Union. This study do not need to scale the features because the sensitivity of our target variable will be incorporated into the feature coefficients.
This study aim for our Bayesian optimized hyperparameter Boost regression model to generalize to the 2028 Los Angeles Olympics. Therefore, we should divide the data records into training, validation, and test sets for model validation. This study use the 2024 Paris Olympics as our test set, which contains 206 records out of 1796 national participation records. The test set represents 11.4763% of the entire dataset.
Our target variable is the total number of medals each country wins at a single Olympics, but our dataset also contains the results for gold, silver, and bronze medals. Therefore, this study do not directly predict the total medal count; instead, this study predict each medal type separately three times. We can then sum the gold, silver, and bronze medals to obtain the total medal count. Finally, this study can use the trained model to predict medal counts for previously unseen data.
This study create training, validation, and test sets to balance bias and variance in the ma- chine learning model. We can test whether the model is overfitting or underfitting by comparing how well it explains the variance in the training and test sets. Similarly, this study expect the error levels based on prediction residuals to be comparable between the two datasets.

This study evaluate the performance of the regression model by predicting the total number of medals for the 2024 Paris Olympics (our held-out test set) and comparing it with different regression models using evaluation metrics. This comparison illustrates the good performance of the Bayesian optimized Boost model, with predictions shown in Figure 2. Further visualization in Figure 3 presents the overall distribution of predicted versus true medal counts in the test set, intuitively reflecting the model's ability to capture the variation in medal outcomes across different countries. The close alignment between predicted values and true labels in the figure further validates the model's reliability for Olympic medal forecasting.

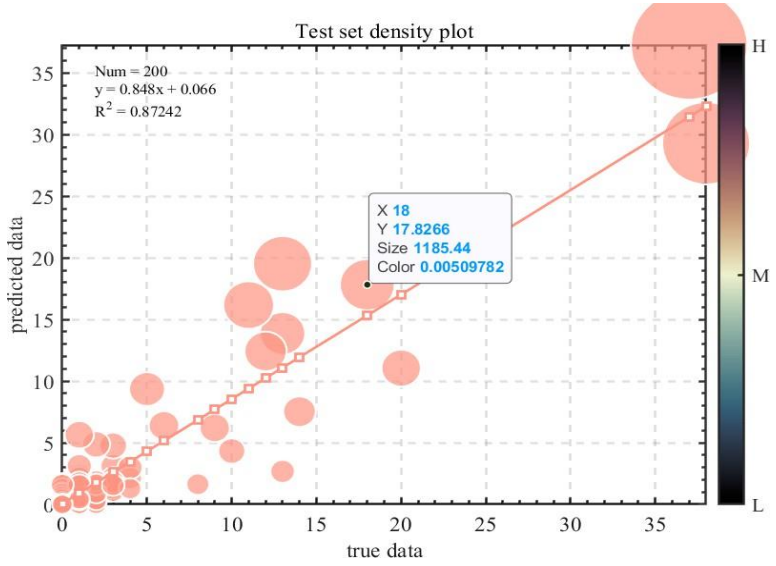

Figure 2 Test Set Density Plot

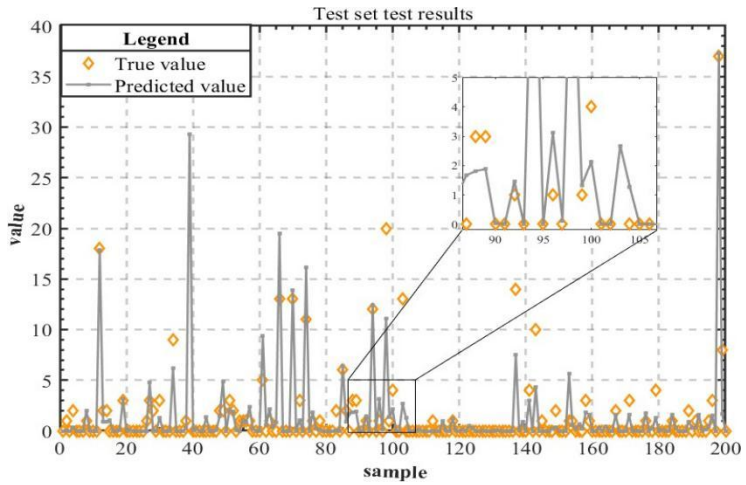

Figure 3 Test Set Effect Diagram

As shown in Table 1, our regression model's coefficient of determination R2 for the training set is 96%, and for the test set, it is 90%. Due to space limitations, we do not include the validation set results in the main text. The model error calculated using RMSE is 1.14 medals per country for the training set and 1.77 medals per country for the test set. Therefore, our regression model is balanced, capable of explaining the data, and has a small error.

Table 1 Comparison of Different Regression Model Evaluation Indicators

Dataset	Metrics	DT	Boost	Bayesian optimization Boost
Training Set	MAE	0.36	0.16	0.44
	MSE	0.94	0.08	1.29
	RMSE	0.97	0.28	1.14
	R2	0.98	0.99	0.97
	MAE	0.79	0.84	0.73
	MSE	3.70	4.01	3.12

Test Set	RMSE	1.92	2.00	1.77
	R2	0.89	0.90	0.91

3.3 Predictions and Analysis of the 2028 Los Angeles Olympic Medal Table

3.3.1 Construction of the 2028 dataset

To use Bayesian-optimized hyperparameter Boost prediction model to fore- cast the medal rankings for the 2028 Los Angeles Olympics, this study first need to update the parameters of the dataset. This study will set the United States as the host country and estimate the total number of athletes and events for each country based on the averages from the 2016 Rio Olympics, 2020 Tokyo Olympics, and 2024 Paris Olympics.

3.3.2 Prediction evaluation and results analysis

The prediction results indicate an overall increase in total medals for the 2028 Olympics, aligning with official announcements from the Los Angeles Olympic Committee regarding the addition of five new sports. This consistency validates the model's ability to incorporate real-world changes in event structures.

(1) Shifts in Medal Rankings

As visualized in Figure 4, significant upward momentum is expected for regions such as Region2, Region3, Region1, Region4, and Region5. These countries have implemented strategic investments in sports infrastructure over the past decade, upgraded athlete training systems with advanced technologies, and established youth development pipelines-efforts that are now poised to yield tangible results. Japan, for instance, has sustained its post-Tokyo Olympics focus on nurturing young talent, particularly in sports like gymnastics and swimming, making it a strong contender for increased medal hauls in 2028.

In contrast, traditional sports powerhouses including Region6, Region7, Region8, and Region9 may face a downturn in medal counts. The United States, while retaining its top position, is projected to see declines in sports like basketball and athletics due to rising global competition-particularly from European and Asian nations-and ongoing issues with unequal resource distribution across domestic sports federations. China, amid efforts to reform its sports system toward more market-driven development, may experience temporary dips in sports such as table tennis and badminton, where international rivals have narrowed the gap, but is expected to maintain a top-three ranking through consistent performance in diving, weightlifting, and shooting. Similarly, Australia and France may see reduced medals in their signature sports but remain competitive due to diversified participation across events.

(2) First Gold Medal Prediction

The 2028 predictions highlight a historic shift with Regions10 to16 set to secure their first Olympic gold medals. This breakthrough reflects years of targeted investment: these nations have focused on sports with lower global competition barriers and leveraged international coaching partnerships to accelerate athlete development. For example, Region12 has significantly improved its boxing program through collaboration with Cuban trainers, while Region15 has invested in high-altitude training facilities to boost its athletes' performance in long-distance running.

(3) Reliability of 80% Prediction Intervals

Table 2 presents the 80% prediction interval metrics for first-time gold medalists, with a PICP of 0.71 indicating that 71% of actual outcomes fall within the forecasted ranges, and a PINAW of 0.18 reflecting narrow interval widths relative to the data scale. These metrics, combined with a moderate CWC, demonstrate that the model's probabilistic forecasts strike a balance between precision and coverage-critical for guiding resource allocation in pre-Olympic preparation.

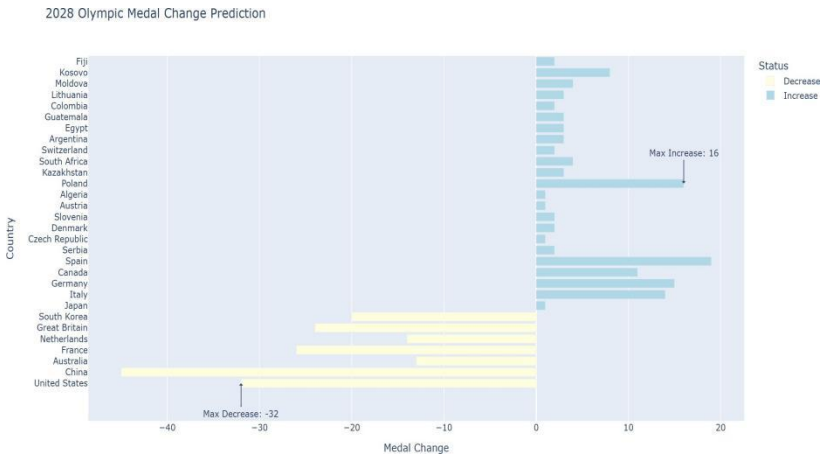

Figure 4 2028 Los Angeles Olympics Predictions

Table 2 0.8 Prediction Interval Evaluation Indicator

Metrics	Numeric
PICP	0.71

PINAW	0.18
CWC	1.35
MPICD	3.73
AIS	3.70

3.4 Medal and Event Relationship Model Based on k-means++ Clustering

Building on the 2028 medal predictions which highlighted shifts in national rankings and emerging breakthroughs, this section further explores the underlying dynamics of Olympic performance by examining how countries' medal hauls correlate with their specialization in specific events. To uncover these patterns, we employed the k-means++ clustering algorithm, leveraging medal data from the 2016 to 2024 Olympic Games.

3.4.1 Data selection and clustering methodology

The analysis focused on three core variables: unique country codes, event categories such as athletics, swimming, gymnastics, and medal counts (gold, silver, bronze) across events. Prior to clustering, data were standardized using Z-score normalization to ensure consistent scaling of variables, and the algorithm was set to partition countries into 3 distinct clusters based on their event-specific medal distributions.

3.4.2 Clustering results: national strength patterns in Olympic events

The k-means++ analysis identified three distinct clusters, each reflecting a unique pattern of national excellence in Olympic sports.

(1) Cluster 1: Specialized Dominance in Specific Events

Represented by countries such as South Korea, Switzerland, Germany, Japan, and France, this cluster is characterized by concentrated medal hauls in one or two flagship events. See Table 3 for the advantageous events and medal counts of these countries. For example, South Korea excels in archery, Switzerland in mountain bike cycling, and Japan in skateboarding. This pattern highlights targeted investment in niche sports, where these nations have developed specialized training systems and competitive advantages.

Table 3 Category 1 Countries and Advantageous Projects

Country	Cluster	Most Skilled Project	Project Award Count
KOR	1	Archery	13
SUI	1	Cycling Mountain Bike	12
JPN	1	Skateboarding	11
GER	1	Equestrianism	10
FRA	1	Handball	10

(2) Cluster 2: Focused Excellence in Single High-Impact Events

The Netherlands typifies this cluster, with exceptional performance in a single event-cycling road-where it earned an average of 9.51 medals over the past three Olympics. Unlike Cluster 1, which may span two sports, Cluster 2 nations demonstrate deep expertise in a single discipline, often with consistent podium finishes that contribute significantly to their total medal tally.

(3) Cluster 3: Multidimensional Strength Across Multiple Events

Countries like the United States, China, and the United Kingdom fall into this cluster, exhibiting balanced medal distributions across diverse events. See Table 4 for details. The U.S. dominates athletics, China leads in diving, and the UK excels in cycling. This breadth reflects comprehensive sports development strategies, with robust investment across multiple disciplines and strong talent pipelines.

Table 4 Category 2 Countries and Advantageous Projects

Country	Cluster	Most Skilled Project	Project Award Count
CHN	3	Diving	13
USA	3	Athletics	12
GBR	3	Cycling	11

3.4.3 Host country's event selection: impact on medal distribution

The clustering results further align with the observation that host countries strategically shape medal outcomes through event selection. By introducing new sports or revising rules to align with their strengths, hosts can enhance their medal potential. For instance, the 2016 Rio Olympics added surfing and golf-sports with growing participation in Brazil-while the 2020 Tokyo Olympics prioritized karate and skateboarding, where Japanese athletes had established competitive edges. Such decisions not only boost domestic interest in these sports but also directly elevate the host's medal performance.

Additionally, the structure of events-particularly the distinction between individual and team sports-significantly influences medal distribution. Variations in team sizes, competition formats, and weight categories lead to stark

differences in medal counts per event. At the Tokyo Olympics, 128 athletes participated in taekwondo, and 32 medals were awarded. Since taekwondo is an individual sport, each event awards two bronze medals. In contrast, hockey, as a team sport with nearly 400 athletes, only had six medals available. This disparity means countries excelling in high-yield individual sports are more likely to accumulate total medals-a pattern echoed in the specialization of Cluster 1 nations.

The consistency between these clustering insights and the predictive performance of our model further validates the robustness of our analytical framework. As shown in Table 1, the Bayesian-optimized Boost model outperforms decision trees and standard Boost models across key metrics in the test set, with an R^2 of 0.91-1.1% higher than the standard Boost model and 2% higher than decision trees. This superior predictive power confirms that the event-specific patterns identified by k-means++ clustering are not only statistically meaningful but also contribute to more accurate medal forecasts. Such alignment between clustering results and predictive performance strengthens the practical value of our findings for host countries' event selection strategies.

4 CONCLUSIONS

This study focuses on Olympic medal prediction and related effect analysis, adopting the technical route of "data preprocessing - model construction - validation and application". It built a solid data foundation through data cleaning and integration, then used a Bayesian-optimized Boost regression model for prediction. The model, which mines nonlinear relationships via Boost regression and achieves automatic hyperparameter tuning with Bayesian optimization, showed a high R^2 of 0.91 in 2024 data validation, outperforming benchmark models like decision trees and standard Boost. Meanwhile, the k-means clustering model analyzed medal data from the past three Olympics, identifying countries' dominant events and quantifying the host country effect on event selection. It successfully predicted the 2028 Los Angeles Olympics medal distribution, indicating increased medals for multiple regions and first golds for six regions. Future improvements may include incorporating political and cultural factors, as well as athletes' individual characteristics like age and injuries, to enhance prediction accuracy. The findings can provide a scientific basis for Olympic committees to formulate strategies and optimize resource allocation.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Bernard B A, Busse R M. Who Wins the Olympic Games: Economic Resources and Medal Totals. The Review of Economics and Statistics, 2004, 86(1): 413-417.

[2] Zhang C. A Study on the Competitive Strength and Promotion Strategies of the Chinese Delegation at the 23rd to 32nd Summer Olympic Games. Dissertation, Wuhan Sports University, 2023.

[3] Tian M. Expansion of the event-group training theory to the event-group theory. Chinese Sports Coaches, 2019, 27(01): 3-7.

[4] Shi H, Zhang D, Zhang Y. Can Olympic Medals Be Predicted? — From the Perspective of Interpretable Machine Learning. Journal of Shanghai University of Sport, 2024, 48(04): 26-36.

[5] Luo Y, Cheng Y, Li M, et al. Prediction of China's Medal Count and Overall Strength at the Beijing Winter Olympics: Based on the Host Country Effect and Grey Prediction Model. Contemporary Sports Technology, 2022, 12(21): 183-186.

[6] Sadu B V, Bagam S, Naved M, et al. Optimizing the early diagnosis of neurological disorders through the application of machine learning for predictive analytics in medical imaging. Scientific Reports, 2025, 15(1): 22488 -22488.

[7] Peng J. Research on hyperparameter optimization of Bayesian optimization algorithm based on GP. China New Technologies and Products, 2025(11): 38-40.

[8] Chen Z, Feng J, Yang D, et al. Hierarchical clustering algorithm for complex structure datasets based on hybrid neighborhood graph. Journal of Intelligent Systems, 2025(8): 1-11.

[9] Shao J. Research on Time-Series Forecasting Based on Deep Neural Networks. Dissertation, Jilin Institute of Chemical Technology, 2024.