

# Random Forest Plants Gold Medal Trees, Tobit Decodes Zero Breakthroughs

## Summary

The Olympic medal landscape evolves through athletic, coaching, and strategic factors, as nations pursue either sustained excellence or breakthrough performance.

For Task 1.1, we develop a **Random Forest model based on spatiotemporal characteristics of medals** to predict Olympic medals. We firstly analyze the **spatial universality** of medals and classify countries into three levels (high, medium, low) by using **K-means++**, then uncover a time-invariant law between medals and rankings. Subsequently, we identify **8** key features for medal prediction and evaluate using  **$R^2$ , NRMSE, and KL divergence** metrics. Finally, the **Random Forest's** results indicate significant improvements for India (+18), Mongolia (+18), and Argentina (+17), while notable declines are projected for Tajikistan (-35), Albania (-34), and Moldova (-31).

For task 1.2, we develop a **Joint Tobit-Random Forest Model based on latent competitiveness**. After introducing "**the zero-medal effect**", we use the **Tobit** model to quantify **latent competitiveness** and address zero-medal censoring. We then apply **Random Forest Classifier** to filtered data, identifying 4 potential first-time medal-winning countries (odds= 0.71, 95% CI: 3-5).

For task 1.3, we examine event-medal correlation through **Pearson** coefficient and **feature importance**. Analysis of countries at three levels reveals Athletics, Boxing and Rowing as most significant sports. Using **Medal Acquisition Rate (MAR)** and **t-tests** ( $p=0.003$ ), we confirm host countries' advantages in new events matching their strengths.

For task 2, we develop a **DID-based "great coach" effect capture and AHP-Grey relational coaching recruitment model**. After confirming DID (Difference-in-Differences) applicability through **parallel trends test**, we quantify the "great coach" effect using **DID** - Chinese Women's Volleyball Team improved by 0.5 in 1996 (3.75 in 2016). Through **AHP** and **Grey Relational Analysis**, we identify coaching needs: Great Britain in rowing/Sailing, Belgium in shooting/road cycling, and Australia in sailing, along with its respective impacts.

For task 3, we analyze **gender distribution** in Olympic medals, revealing the evolution from male dominance to growing **equality**, with distinct patterns across different performing nations. Meanwhile, we write a letter to National Olympic Committees to share our insights and suggestions. Finally, **Sensitivity analysis** on feature selection and retention rates for Olympic medal prediction confirms model robustness.

**Keywords:** Random Forest, Tobit, DID, Grey Relational Analysis

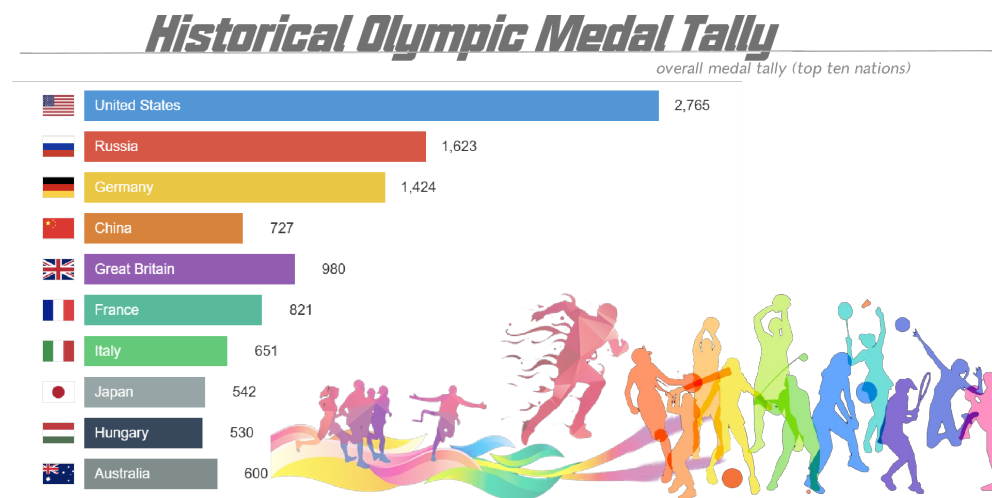
# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Background . . . . .	2
1.2	Problem Restatement and Analysis . . . . .	2
1.3	Our Work . . . . .	3
<b>2</b>	<b>Assumptions And Notations</b>	<b>3</b>
2.1	Assumptions . . . . .	3
2.2	Notations . . . . .	3
<b>3</b>	<b>Data Preprocessing</b>	<b>4</b>
<b>4</b>	<b>Predicting Olympic Podiums: When Trees Forecast Gold</b>	<b>5</b>
4.1	Universal Laws of Medal Distribution . . . . .	5
4.1.1	Spatial Cluster Effects of Medal Distribution . . . . .	5
4.1.2	Temporal Invariance in Medal Distribution . . . . .	6
4.2	Feature Engineering . . . . .	7
4.3	Olympic medal prediction based on Random Forest . . . . .	9
4.3.1	Random Forest Model . . . . .	9
4.3.2	Computing Confidence Intervals Using Bootstrap Method . . . . .	9
4.3.3	Selection and Implementation of Performance Metrics . . . . .	10
4.4	Medal Prediction Results and Cross-Country Analysis . . . . .	10
<b>5</b>	<b>Breaking the Zero: A Joint Tobit-Random Forest Model for Medal Prediction</b>	<b>12</b>
5.1	Zero-Medal Effect and Latent Competitiveness . . . . .	12
5.2	Construction of Prediction Models . . . . .	13
5.2.1	Quantifying Latent Competitiveness via Tobit Regression . . . . .	14
5.2.2	RF-based Medal Probability Estimation . . . . .	15
5.3	Predictions for First-Time Olympic Medalists in 2028 . . . . .	15
5.4	Event-Nation Medal Relationships: Patterns and Specializations . . . . .	15
<b>6</b>	<b>Model III:DID Analysis and AHP-Grey Model for Coach Effect</b>	<b>18</b>
6.1	“great coach” Effect . . . . .	18
6.2	Difference-in-Differences Model to Capture “great coach” Effect . . . . .	18
6.3	Coach Hiring: An AHP Analysis . . . . .	19
6.4	Grey Relational Analysis and Results . . . . .	20
<b>7</b>	<b>Other insights breaking down</b>	<b>21</b>
<b>8</b>	<b>Strengths And Weaknesses</b>	<b>22</b>
	<b>References</b>	<b>24</b>

# 1 Introduction

## 1.1 Background

"With three minutes left in the game, Brazil trails Argentina by just one medal in the total count..." The suspense of Olympic medal races captivates billions of viewers worldwide, transforming what was once a simple tally into a global drama of national pride and sporting excellence. Why do some nations suddenly break through while others plateau? Behind these numbers lies a complex interplay of talent and strategy, and great coaches can transform raw potential into Olympic success.



## 1.2 Problem Restatement and Analysis

We need to complete the following tasks based on the data given in the question:

- **Task 1: Develop a predictive model for Olympic medal counts:**

Develop a model to predict medal counts for countries, with uncertainty estimates and performance metrics.

- Project the medal counts for 2028 Los Angeles Olympics with prediction intervals, identifying countries likely to improve or decline.
- Model should predict which countries might win their first Olympic medal and provide probability estimates.
- Study the relationship between Olympic events and medal counts, analyzing country strengths and host nation advantages.

- **Task 2: Analyze the "great coach" effect.**

Examine how top coaches switching countries impacts medal counts, and recommend coaching investments for three countries.

- **Task 3: Synthesize insights for Olympic committees** Discover unique insights from the model and provide recommendations for national Olympic committees.

### 1.3 Our Work

For better arranging our process of problem solving, the flow diagram of our work is shown in Figure 1.

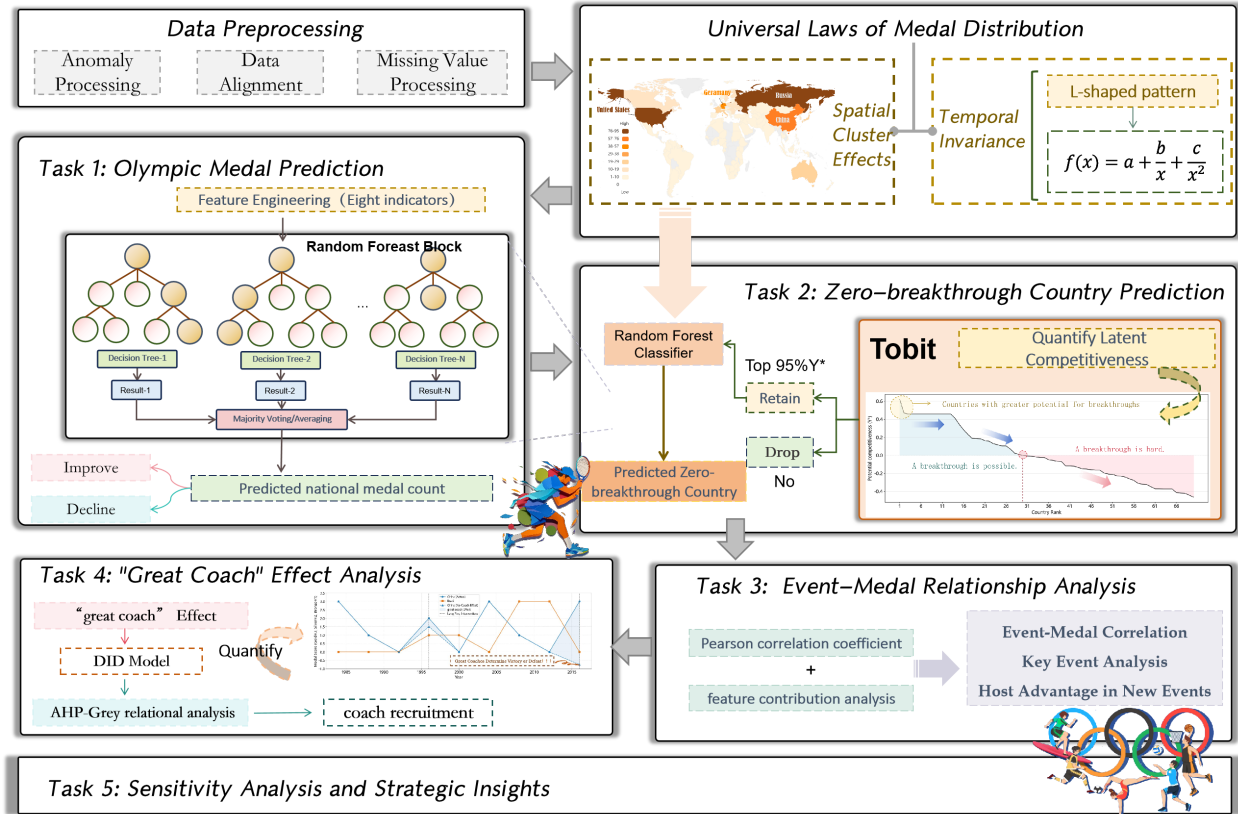


Figure 1: Our Work

## 2 Assumptions And Notations

### 2.1 Assumptions

We make several assumptions in our model. Using these assumptions can simplify our model and make it more applicable to complex real-world environments.

**Assumption 1 :** Olympic medals are weighted differentially (gold=3, silver=2, bronze=1) for performance comparison. **Justification:** This weighting system reflects standard Olympic evaluation practices and the hierarchical value of achievements.

**Assumption 2 :** Medal counts depend only on the given dataset variables. **Justification:** This simplifies analysis by isolating key factors, though real-world outcomes may involve additional influences.

### 2.2 Notations

The primary notations used in this paper are listed in Table 1.

Table 1: Notations

Symbol	Definition
$Y_i$	Observed medal count for country i
$Y_i^*$	Latent competitiveness of country i
$X_i$	Feature vector for country i
$\epsilon_i$	Error term in Tobit model
$\Phi(\cdot)$	Standard normal cumulative distribution function
$MAR$	Medal Acquisition Rate
$f(x)$	Medal distribution function

Note: Symbols not described in the table are based on the first appearance.

### 3 Data Preprocessing

Data pre-processing is a crucial step in ensuring data quality and analytical accuracy for Olympic Games research. In our study, we implement three primary steps, The schematic presentation is shown in Figure 2:

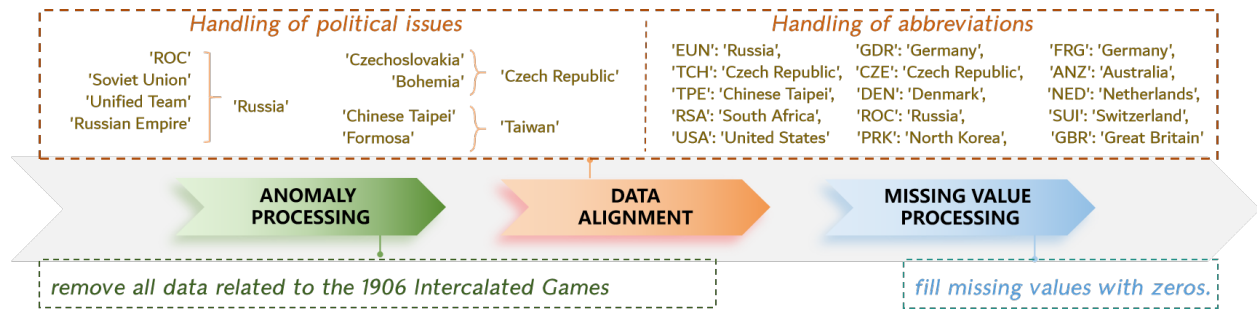


Figure 2: Data Preprocessing

#### (1). Anomaly Processing

We identify that the 1906 Games, held between the third and fourth International Olympic Games, are classified as Intercalated Games. These Games are not officially recognized by the IOC, and the medals awarded during this period are not officially acknowledged. To maintain data consistency and ensure analytical accuracy, we remove all data related to the 1906 Intercalated Games from our dataset.

#### (2). Data Alignment

Due to the complex political history of many participating nations, we implement a comprehensive standardization process for country names. This includes managing cases such as unifying "Soviet Union," "Unified Team," and "ROC" as "Russia,". Furthermore, we standardize country code abbreviations to their full country names to ensure dataset consistency.

### (3). Missing Value Processing

In the summerOly\_programs.csv dataset, we employ a logical imputation strategy. For empty cells in competition event quantities, we compare the total events numbers with the sum of non-missing values for each year. We determine these missing values represent non-participation in events and therefore fill them with zeros.

## 4 Predicting Olympic Podiums: When Trees Forecast Gold

### 4.1 Universal Laws of Medal Distribution

We examine if medal count distributions show invariance across time and space, and whether medal patterns follow universal distributions among countries.

#### 4.1.1 Spatial Cluster Effects of Medal Distribution

Despite having a large population, strong national power, and a significant economy, South Asia's India has consistently underperformed at the Olympics. This raises questions about the impact of factors like terrain, climate, and national traits on Olympic success, suggesting a possible spatial clustering effect in medal distribution.

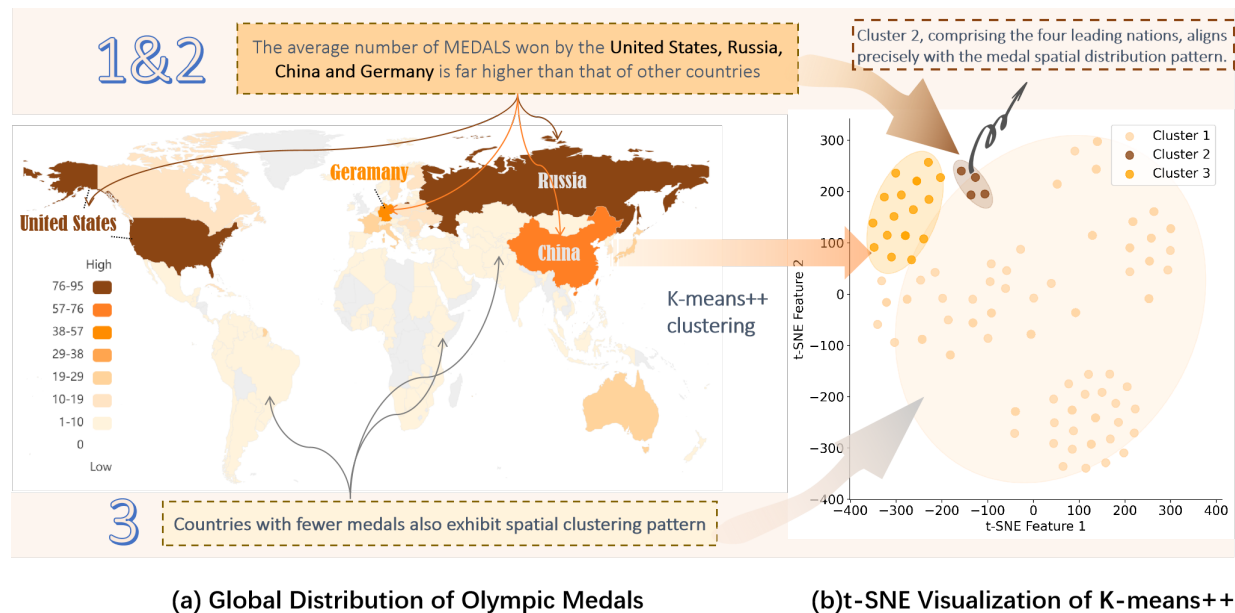


Figure 3: Spatial Analysis of Olympic Medal Distribution

To investigate the spatial patterns of Olympic medal distribution, we conducted a longitudinal analysis rather than focusing on individual Olympiads, which could be affected by temporal fluctuations (e.g., economic crises, COVID-19 pandemic). Figure 3(a) presents a spatial distribution map based on the arithmetic mean of medal counts across all Olympic Games, providing a representative view of national Olympic performance.

Figure 1(a) reveals strong spatial clustering in Olympic performance, with geographic neighbors showing similar medal counts. Major powers like the US, Russia, Germany and

China contrast with lower-performing regions in Africa, South America, and Southeast Asia.

The spatial autocorrelation in medal distribution reveals patterns requiring analysis. Using K-means++ clustering with 4 medal features (gold, silver, bronze, and total), we categorized countries into three distinct clusters, visualized through t-SNE in Figure 2(b).

- **High-medal countries:** Average medal count > 76 (e.g., United States, Russia, China and Germany), which precisely corresponds to the dominant nations shown in the darker shades on the distribution map.
- **Medium-medal countries:** Average medal count between 38-76 (e.g., Europe and the Asia-Pacific region), forming a concentration belt in the developed regions of the Northern Hemisphere.
- **Low-medal countries:** Average medal count < 38, predominantly developing nations distributed across Africa, South America, Southeast Asia, and the Middle East.

#### 4.1.2 Temporal Invariance in Medal Distribution

Following our spatial analysis, we investigated the temporal stability of medal distributions across Olympic cycles. We constructed rank-medal distribution curves for the 2020 and 2024 Olympics (Figure 4) to quantitatively characterize temporal patterns. This analysis aims to differentiate between inherent structural patterns and temporary performance fluctuations in the observed spatial clustering.

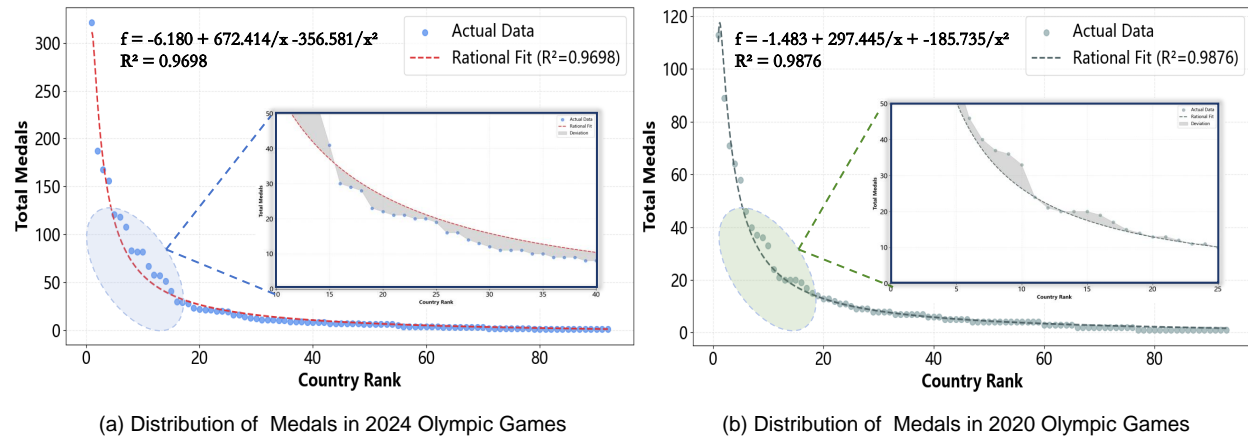


Figure 4: Distribution of Medals in Olympic Games

When analyzing data from Olympic Games across different periods, we observe two prevalent distribution patterns: first, a power-law distribution with stepwise decline evident in both medal rankings and medal quantities; second, an L-shaped distribution with multiple small peaks in the horizontal section when examining the number of countries against medal counts. As illustrated in Figure 4.



Medal distribution shows a right-skewed L-shape, with few dominant countries capturing most medals, followed by a stepwise decline and a long tail of low-performing nations. We fitted power functions to two Olympics' data for pattern analysis and prediction.

$$f_{2024}(x) = -6.180 + \frac{672.414}{x} - \frac{356.581}{x^2} \quad (R^2 = 0.9698) \quad (1)$$

$$f_{2020}(x) = -1.483 + \frac{297.445}{x} - \frac{185.735}{x^2} \quad (R^2 = 0.9876) \quad (2)$$

The high  $R^2$  value of 0.9698 and 0.9876 indicates that this distribution pattern is not random but follows a consistent mathematical relationship between country rankings and medal counts, demonstrating temporal invariance across different Olympic Games.

Therefore, these universal laws persist across Olympics, maintaining consistent medal distribution patterns within each country cluster and exhibiting invariance under both temporal and spatial transformations, conforming to the universal distribution function:

$$f(x) = a + \frac{b}{x} + \frac{c}{x^2} \quad (3)$$

Country ranking stability provides a reliable metric for validating predictions, while consistent distribution patterns offer additional verification standards.

## 4.2 Feature Engineering

Feature engineering transforms historical data into predictive features, optimizing the Random Forest model for Olympic indicators. Host Country Analysis includes three binary variables:

### (1). Host Country Analysis

#### i. Host Nation Indicator

Binary variable (0/1) indicating current host nation status. Host nations typically benefit from home crowd support, familiar venues, and increased infrastructure investment. For instance, China's performance at the 2008 Beijing Olympics demonstrated significant home advantage, securing 51 gold medals compared to 32 at the 2004 Athens Games.

#### ii. Previous Host Effect

Binary variable (0/1) reflecting the four-year post-hosting impact on athletic performance. This captures the sustained effects of sports infrastructure development and investment. The UK's continued strong performance in cycling post-2012 London Olympics exemplifies this legacy effect.

#### iii. Novel Event Indicator

Binary variable (0/1) identifying newly introduced Olympic events. Host nations often gain competitive advantages in new sports through early preparation and dedicated resources. Japan's success in skateboarding at the 2020 Tokyo Olympics illustrates this pattern, where they secured medals in this debut event.



## (2). Athletic Resource Indicators

### i. Historical Athlete Participation

The number of athletes indicates a nation's Olympic engagement and competitive investment, showing strong correlation with medal counts. We use the average participant numbers from the past two Olympics to balance data recency and competitive continuity, ensuring both timeliness and stability in our prediction model.

### ii. Gender Distribution Index

Through feature engineering, gender distribution becomes a crucial indicator as it reflects a nation's sports development strategy and resource allocation patterns, with studies showing significant differences in Olympic competition patterns between women's and men's events. Based on this trend, we use **the male athlete participation ratio from the previous two Olympics** as a predictive feature.

## (3). Historical Performance Metrics

### i. Medal Performance History( $M_t$ )

Historical medal count serves as a direct indicator of a nation's Olympic strength. By analyzing the total medal count from the previous two Olympics, we can establish a recent performance baseline that reflects both current competitive capacity and development momentum, calculated as follows:

$$M_t = \frac{M_{t-1} + M_{t-2}}{2} \quad (4)$$

where  $M_{t-1}$ ,  $M_{t-2}$  denote the total medals won in the previous two Olympics respectively.

### ii. Event Participation Breadth( $E_t$ )

Olympic event participation indicates a nation's sports development breadth. Event counts across two cycles reveal sustained investment while smoothing fluctuations. The formula is:

$$E_t = \frac{E_{t-1} + E_{t-2}}{2} \quad (5)$$

where  $E_{t-1}$ ,  $E_{t-2}$  indicate the number of events participated in the previous two Olympics.

### iii. National Athletic strength( $W_t$ )

Considering that different medal tiers represent distinct levels of athletic achievement, we assign differential weights to gold (3 points), silver (2 points), and bronze (1 point) medals, and calculate the weighted average from the previous two Olympic Games as follows:

$$W_t = \frac{\sum_{i=t-2}^{t-1} (3G_i + 2S_i + B_i)}{2} \quad (6)$$

where  $G_i$ ,  $S_i$ , and  $B_i$  represent the number of gold, silver, and bronze medals respectively for Olympics year  $i$ .

### 4.3 Olympic medal prediction based on Random Forest

#### 4.3.1 Random Forest Model

Given the requirements of collaborative prediction for multiple objectives (gold, silver, bronze medals and total) in Olympic medal prediction and the nonlinear correlation characteristics of features, this study adopts Random Forest algorithm to construct the prediction framework. As illustrated in Figure 5:

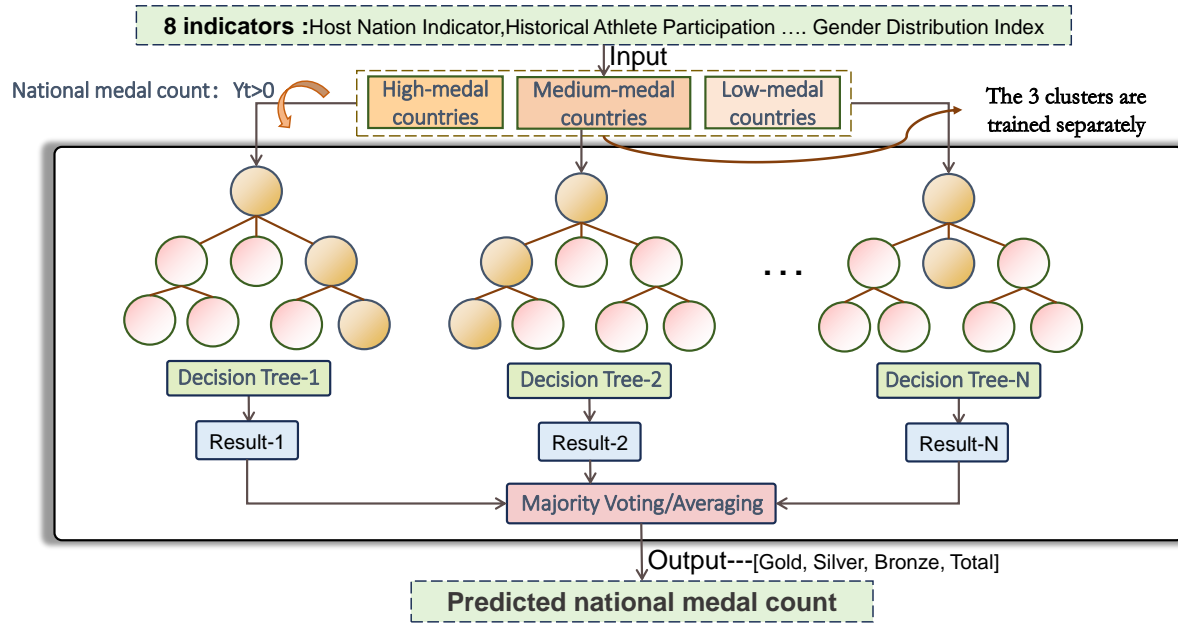


Figure 5: Random Forest frame diagram

Meanwhile, based on the spatial and temporal stability characteristics of medal distribution, three different types of countries - High-medal, Medium-medal, and Low-medal - were obtained through K-means++ clustering in the previous section. Therefore, this study conducts Random Forest training and prediction separately for each type.

#### 4.3.2 Computing Confidence Intervals Using Bootstrap Method

This study presents a robust interval estimation approach for Olympic medal predictions, grounded in Bootstrap methodology. Given a prediction sequence  $Y = y_1, y_2, \dots, y_{1000}$  generated by random forest models, we implement the following procedure to construct confidence intervals:

**i. Group Averaging.** Divide the prediction sequence  $Y$  into  $n$  groups ( $n = 100$ ), with each group containing  $k$  consecutive predictions ( $k = 10$ ). For the  $i$ -th group, calculate its mean:

$$b_i = \frac{1}{k} \sum_{j=1}^k y_{(i-1)k+j}, \quad i = 1, 2, \dots, n \quad (7)$$

ii. **Sorting.** Generate a new sequence  $B = b_1, b_2, \dots, b_{100}$  and arrange it in ascending order to obtain:

$$B_{(1)} \leq B_{(2)} \leq \dots \leq B_{(n)} \quad (8)$$

iii. **Confidence Interval.** With  $\alpha = 0.1$ , construct the 90% confidence interval:

$$CI_{0.9} = [B_{(n\alpha/2)}, B_{(n(1-\alpha/2))}] = [B_{(6)}, B_{(95)}] \quad (9)$$

This grouped averaging method reduces outlier effects while providing robust interval estimates and reliable uncertainty quantification for medal predictions at the 2028 Los Angeles Olympics.

### 4.3.3 Selection and Implementation of Performance Metrics

We chose:  $R^2$  for overall fit, NRMSE for scaled error comparison, and KL divergence (Considering the Temporal Invariance in Medal) Distribution for distribution pattern validation."metrics:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

$$NRMSE = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}}{y_{max} - y_{min}} \quad (11)$$

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad (12)$$

where  $y_i$  represents the actual number of gold medals,  $\hat{y}_i$  denotes the predicted number,  $\bar{y}$  is the mean of actual values,  $n$  is the sample size,  $y_{max}$  and  $y_{min}$  are the maximum and minimum values of actual gold medals respectively, and  $P(x)$  and  $Q(x)$  represent the probability distributions of actual and predicted gold medals.

## 4.4 Medal Prediction Results and Cross-Country Analysis

### (1). Evaluation of Prediction Results

To validate our model's predictive accuracy, we compared the actual versus predicted medal rankings for 2024, resulting in the following table2.

Table 2: Model Performance Metrics for Different Medal Country Groups

Country Group	Gold		Silver		Bronze	
	$R^2$ Score	NRMSE	$R^2$ Score	NRMSE	$R^2$ Score	NRMSE
High-medal Country	0.8650	0.0796	0.8464	0.0705	0.8722	0.0579
Medium-medal countries Country	0.8763	0.0359	0.8786	0.0366	0.8747	0.0509
Low-medal Country	0.8667	0.0305	0.8619	0.0266	0.8810	0.0251

Note: KL Divergence between 2024 actual and predicted rankings = **0.891**

The KL divergence of **0.891** and high  $R^2$  scores (**0.8619-0.8810**) across all medal types indicate strong model performance. Low NRMSE values, especially for low-medal countries, further confirm the model's precision in forecasting Olympic medal distributions.

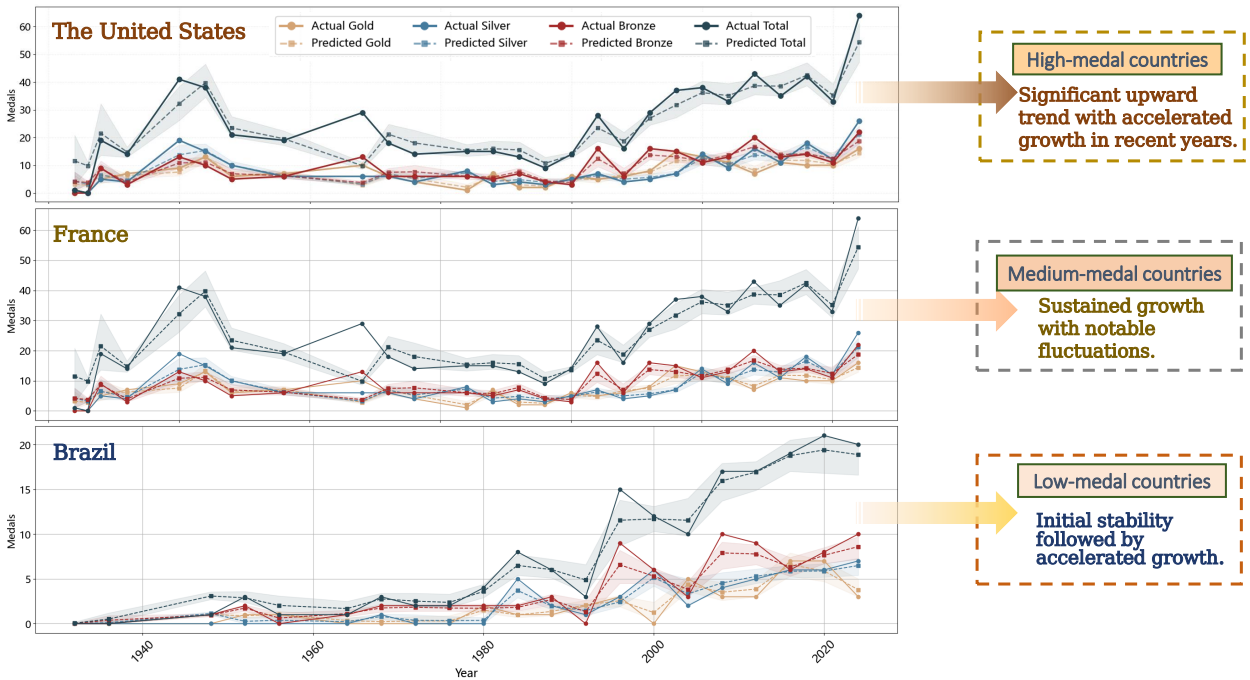


Figure 6: Three-Category National Representatives: Forecast Model Visualization

In terms of prediction accuracy, the model demonstrates precise capture of three categories of countries: accurately predicting the USA’s sustained dominance, France’s cyclical fluctuations (including the boost from hosting the 2024 Olympics), and Brazil’s transition from stability to breakthrough (particularly the peak during the 2016 Olympics). Trend-wise, high-medal countries maintain steady growth, medium-medal countries show stable progress with significant event-driven fluctuations, and low-medal countries have evolved from initial stability to accelerated growth. The model’s prediction intervals are appropriately set, reflecting both prediction confidence and increasing uncertainty in long-term forecasting.

(2). Analysis of Prediction Results

Using Random Forest modeling and Bootstrap confidence intervals, we predict medal outcomes for the 2028 Los Angeles Olympics. Our analysis presents medal forecasts for both leading nations (top 5) and trailing nations (bottom 5), including point estimates and 95% confidence intervals for all medal types.

Table 3: Top 5 Countries Predicted Medal Result for 2028 Los Angeles Olympics

Country	Gold	Silver	Bronze	Total	Total_clean	Rank
United States	50 [43-59]	39 [31-49]	34 [29-38]	123 [105-143]	123	1
China	35 [31-38]	28 [25-30]	26 [23-29]	89 [84-94]	89	2
Russia	19 [14-26]	18 [15-22]	16 [13-21]	53 [43-68]	53	3
Great Britain	16 [13-20]	17 [15-20]	19 [16-22]	52 [46-59]	52	4
Australia	18 [16-21]	15 [13-17]	16 [14-18]	49 [45-53]	49	5

Table 4: bottom 5 Countries Predicted Medal Result for 2028 Los Angeles Olympics

Country	Gold	Silver	Bronze	Total	Total_clean	Rank
Belarus	2 [1-3]	3 [2-4]	3 [2-4]	8 [6-10]	8	32
Philippines	1 [1-2]	2 [2-2]	2 [1-3]	5 [4-6]	5	47
Bermuda	0 [0-1]	1 [0-1]	1 [0-1]	2 [1-2]	2	92
Namibia	0 [0-1]	1 [0-1]	0 [0-1]	1 [1-2]	1	102
Uruguay	0 [0-0]	0 [0-0]	0 [0-0]	0 [0-1]	0	109

From the prediction results, there is a clear hierarchical pattern: The United States and China are significantly ahead, expected to win 123 and 89 medals respectively; Russia, Great Britain, and Australia form the second tier; other countries maintain below 50 medals. This hierarchical phenomenon is highly consistent with the spatial clustering effect we observed earlier.

Table 5: Countries with Most Significant Ranking Changes in Medal Predictions

Improve		Decline	
NOC	Ranking Change	NOC	Ranking Change
India	18	Tajikistan	35
Mongolia	18	Albania	34
Argentina	17	Moldova	31
Slovakia	17	Ethiopia	30
Jordan	16	Grenada	27
Serbia	12	Kyrgyzstan	22
Indonesia	12	Pakistan	21
Czech Republic	10	Kazakhstan	20
Dominican Republic	9	Bahrain	20
Morocco	9	Dominica	20

We found that due to increased investment in sports infrastructure and talent development in recent years, the rankings of India and Mongolia are projected to rise by 18 places. Conversely, due to reduced resources or competitive pressures, Tajikistan is expected to drop by 35 places. Our predictions effectively capture these current trends.

## 5 Breaking the Zero: A Joint Tobit-Random Forest Model for Medal Prediction

### 5.1 Zero-Medal Effect and Latent Competitiveness

#### (1). Zero-medal Effect

The "Zero-Medal Effect" describes the phenomenon where countries with certain sports resources and potential have never won Olympic medals. For example, Nepal, despite its natural advantages in outdoor sports like mountaineering, has not secured any medals since its first Olympic participation in 1964.

The existence of the Zero-Medal Effect suggests that while many countries may underperform in the Olympics, their latent sports competitiveness might not yet be fully realized. Therefore, when constructing predictive models, it is essential to account for the unique characteristics of zero-medal countries. By pre-screening these nations, we can avoid overlooking their potential to win medals in future Games.

## (2). Latent Competitiveness

In contrast to the Zero-Medal Effect is the concept of "Latent Competitiveness," which refers to a nation's potential medal competitiveness in the Olympics, even if this potential has not yet translated into actual medals, as illustrated in Figure 7.

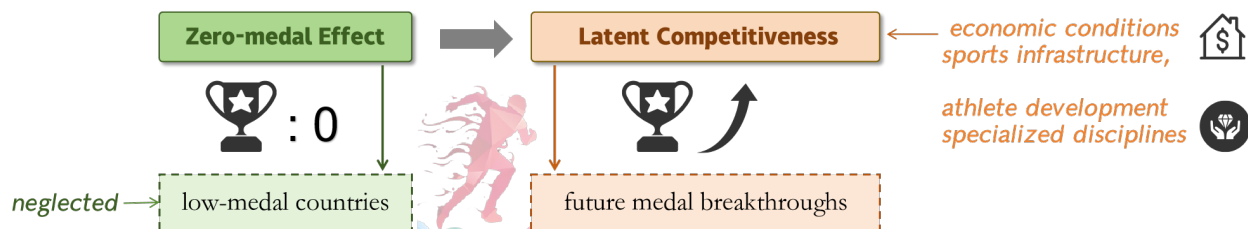


Figure 7: from Zero-Medal to Latent Competitiveness

For example, while some African countries have historically underperformed in the Olympics, their potential in track and field events has been widely recognized. By quantifying the latent competitiveness of these nations, it can be used as a feature to predict and achieve future medal breakthroughs.

## 5.2 Construction of Prediction Models

To identify the latent competitiveness of low-medal countries (defined as those with few or no medals) in Model 1, this study employs the following analytical framework. As illustrated in Figure 8:

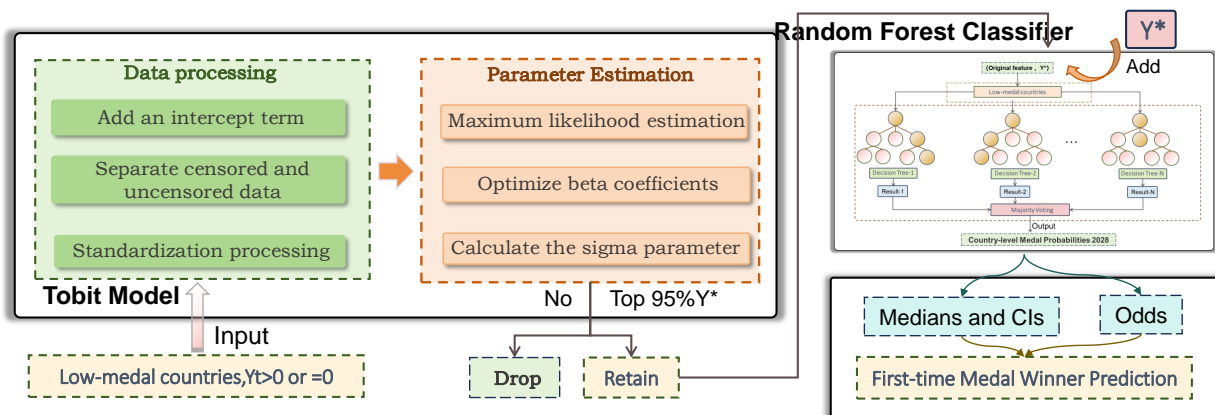


Figure 8: Model Framework

A Tobit model calculates nations' latent competitiveness ( $Y^*$ ), with top 95th percentile countries selected. This  $Y^*$  metric combines with original features in a random forest

classifier to predict 2028 medal probabilities. Bootstrap resampling of the model estimates first-time medalists and their odds ratios.

### 5.2.1 Quantifying Latent Competitiveness via Tobit Regression

To address the data censoring issue of numerous zero-medal countries, we employ a Tobit model with a latent variable  $Y^*$  representing inherent competitiveness. This framework allows for negative competitiveness values and only shows positive medal counts when latent competitiveness exceeds a threshold, enabling unbiased parameter estimation using both zero and positive observations.

#### i. Define latent competitiveness

The latent Olympic competitiveness is modeled as:

$$Y_i^* = \mathbf{X}_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \quad (13)$$

where  $Y_i^*$  is country  $i$ 's latent competitiveness,  $\mathbf{X}_i$  contains eight explanatory variables, and  $\epsilon_i$  is normally distributed with zero mean and variance  $\sigma^2$ .

#### ii. Define the Observation Mechanism

The observation mechanism characterizes the relationship between observed Olympic medal counts and latent competitiveness. When a country's latent competitiveness is positive, its medal count directly reflects its potential capability; when the latent competitiveness is zero or negative, the country obtains no medals.

$$Y_i = \begin{cases} Y_i^* & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \quad (14)$$

where  $Y_i$  denotes the observed medal count and  $Y_i^*$  represents the latent competitiveness.

#### iii. Derivation of Latent Competitiveness

The likelihood function consists of two components describing zero and positive medal distributions, using standard normal cumulative distribution and normal density functions respectively. These components are combined into a log-likelihood function:

$$\ln L = \sum_{Y_i=0} \ln[\Phi(-\frac{X_i\beta}{\sigma})] + \sum_{Y_i>0} \ln[\frac{1}{\sigma}\phi(\frac{Y_i - X_i\beta}{\sigma})] \quad (15)$$

Through the calculations above, we obtained the latent competitiveness of each country:  $Y_i^*$

#### iv. Country Selection Based on Latent Competitiveness

The Olympic medal distribution shows a temporally invariant L-shaped long-tail pattern. Accordingly, we exclude the bottom 5% of countries by latent competitiveness to optimize the random forest model and reduce noise.

We examined the latent competitiveness distribution among zero-medal nations (Figure 9). The x-axis shows country rankings by latent competitiveness, while the y-axis displays



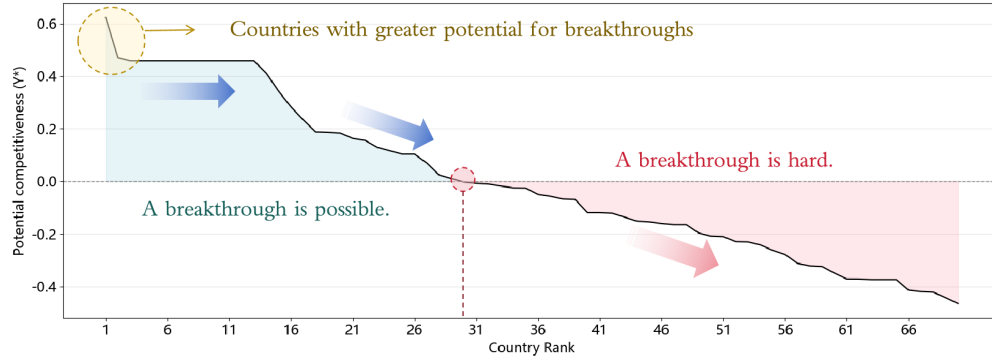


Figure 9: Distribution of Countries with Potential for Olympic Breakthroughs

the deviation from mean competitiveness. Analysis reveals approximately 30 countries demonstrate breakthrough potential. This preliminary exploration provides insights into untapped Olympic potential among these nations.

### 5.2.2 RF-based Medal Probability Estimation

**i. Random Forest classifier.** After Tobit screening, we construct the following feature set for random forest binary classification:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_8, Y^*]^\top \quad (16)$$

where  $\mathbf{X}$  represents the feature set comprising eight original variables ( $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_8$ ) and the estimated potential competitiveness ( $Y^*$ ), incorporating both explicit and latent competitive factors.

**ii. Application of Bootstrap methodology for predicting national medal-winning probabilities and associated confidence intervals in the 2028 Olympic Games.** An ensemble of 1000 random forest classifiers, trained on 63.2% Bootstrap samples, was implemented to generate robust probabilistic predictions for national medal outcomes in the 2028 Olympics.

## 5.3 Predictions for First-Time Olympic Medalists in 2028

Using a Tobit model and Random Forest classifier, we quantified zero-medal countries' latent competitiveness and medal-winning probabilities. Results in Table 6 show 4 countries (odds=0.71, 95% CI: 3-5) are predicted to win their first Olympic medals in 2028.

Netherlands Antilles shows the highest probability for first-time Olympic medals in 2028, followed by Oman and Guinea. Wide confidence intervals support prediction robustness. Model performance metrics are detailed in Table 7.

The excellent performance metrics, with accuracy of 0.971 and ROC AUC of 0.993, demonstrate our joint Tobit-Random Forest model's robust prediction capabilities.

## 5.4 Event-Nation Medal Relationships: Patterns and Specializations

### (1). The Relationship Between The Events And Medals

Table 6: Top 10 Countries with Highest Probability of First Olympic Medal in 2028

NOC	Mean Probability	95% CI
Netherlands Antilles	0.755383	[0.574, 0.970]
Oman	0.572343	[0.300, 0.760]
Guinea	0.56219	[0.377, 0.666]
Cambodia	0.511248	[0.300, 0.950]
Equatorial Guinea	0.485858	[0.300, 0.950]
Federated States of Micronesia	0.485858	[0.300, 0.950]
Nicaragua	0.475240	[0.335, 0.630]
Kiribati	0.458753	[0.300, 0.742]
Rwanda	0.456845	[0.300, 0.705]
Angola	0.454550	[0.344, 0.718]

Table 7: Random Forest Classification Metrics

Metric	Accuracy	Precision	Recall	F1 Score	ROC AUC
Value	0.971	0.993	0.924	0.957	0.993

In Olympic competition, national strength is evaluated through both individual events and overall event distribution patterns.

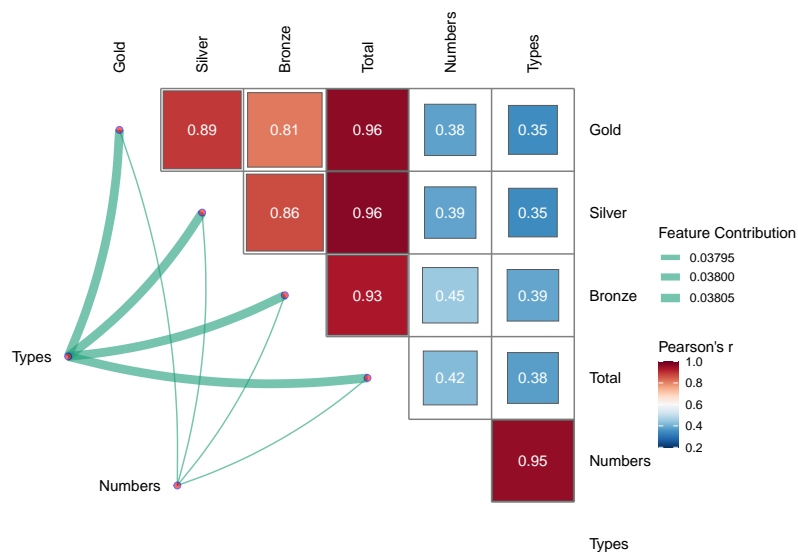


Figure 10: A Heatmap Analysis of Event Distribution and Medal Performance

To analyze event-medal relationships, we used **Pearson correlation coefficient** and feature contribution analysis, examining both event breadth and diversity. We focus on **number of events** and **types of events** as key indicators for medal correlation analysis, as shown in Figure 10.

Event numbers and types show modest positive correlation ( $r \approx 0.35$ ) with medal counts, both contributing similarly to medal success.

## (2).Key Sports Analysis Across Different Medal-Achieving Nations

We analyzed medal correlations across sports using representative countries from three K-means++ clusters, visualized in Figure 11's bubble plot.

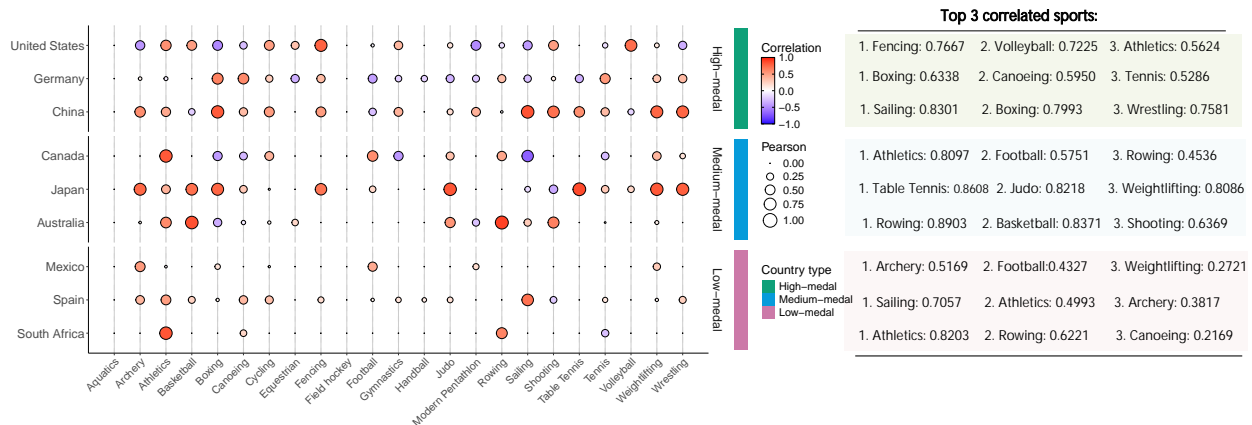


Figure 11: Olympic Sports Correlation Bubble Plot And Top 3 Correlated Sports

Results show **Athletics** as the most common top sport, due to its global accessibility. **Boxing** and **Rowing** also frequently rank among top contributors to national medal counts. High-medal nations show broad success across multiple sports, indicating well-rounded sports development systems. Mid-level nations excel in specific sports clusters (like Japan's strength in Judo), while low-medal nations focus resources on select events where they maintain competitive advantages.

## (3).Impact of Novel Events on Host Nation Performance

Generally, host nations demonstrate a significant advantage in medal acquisition across Olympic events. To specifically examine whether this advantage extends to newly added events, we developed a comparative analysis framework examining medal acquisition rates (MAR) between host and non-host nations.

We define the medal acquisition rate as a metric to measure a nation's performance efficiency in newly added events, calculated by dividing the number of medals won by the number of participating athletes.

To test whether host nations have a significant advantage in newly added events, we conducted an independent samples **t-test** with the following hypotheses:

$H_0$ : There is no significant difference in the medal acquisition rates between host nations and non-host nations in newly added events

$H_1$ : Host nations have a higher medal acquisition rate than non-host nations in newly added events

Analysis shows significant performance differences between host and non-host nations (p

= 0.003 < 0.05), demonstrating hosts' competitive advantage in new events. Host nations enhance performance through expertise, targeted training, and strategic event selection aligned with their sporting culture.

## 6 Model III:DID Analysis and AHP-Grey Model for Coach Effect

### 6.1 "great coach" Effect

The "Great Coach Effect" shows how mobile elite coaches improve national sports performance. Evaluation challenges include: isolating coaching impact from other factors, allowing time for system implementation, and assessing post-departure sustainability.

### 6.2 Difference-in-Differences Model to Capture "great coach" Effect

This study uses Difference-in-Differences (DID) to measure the Great Coach Effect by comparing treatment (elite-coached teams) and control groups across different coaching periods, controlling for confounders while quantifying coaching impact.

#### (1). Testing the Validity of DID Model

To test the parallel trend assumption of DID model, we specify the hypotheses:

$H_0 : \delta_{pre} = 0$  (No differential pre-trends between treatment and control groups)

$H_1 : \delta_{pre} \neq 0$  (Existence of differential pre-trends)

When  $p > 0.05$  for  $\delta_{pre}$ , we fail to reject  $H_0$ , validating the parallel trend assumption for DID estimation.

#### (2). Specification of DID Model

$$Score_{i,t} = \beta_0 + \beta_1 \cdot Treatment_i + \sum_{t \in T} \beta_t^p \cdot Post_t + \sum_{t \in T} \beta_t^d \cdot (Treatment_i \cdot Post_t) + \epsilon_{i,t} \quad (17)$$

where  $Score_{i,t}$  represents Olympic performance,  $Treatment_i$  and  $Post_t$  are treatment and post-intervention dummies,  $\beta_t^d$  is the treatment effect, and  $\epsilon_{i,t}$  is the error term.

#### (3). Results of DID Model

Evaluating Coach Effects through Difference-in-Differences: The Case of Lang Ping and Chinese Volleyball.

##### i. Test Results.

The parallel trends test yields a t-statistic of -0.3430 ( $p=0.7489$ ), confirming no significant pre-treatment differences between Chinese and Brazilian teams and validating the DID model's reliability.

##### ii. DID Results.

DID analysis shows varying coaching impacts: a minor effect in 1996 (0.50) versus a stronger influence in 2016 (3.75), with the higher coefficient reflecting improved perfor-

mance under Lang Ping's leadership. We compared Chinese and Brazilian women's volleyball Olympic performances to illustrate the Great Coach Effect's impact., as shown in the figure12:

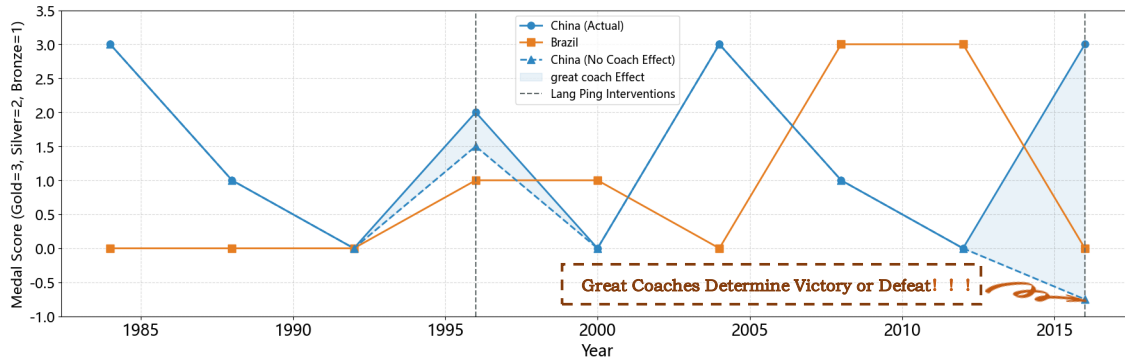


Figure 12: Performance Comparison of Chinese and Brazilian Women's Volleyball Teams

The graph compares Chinese and Brazilian women's volleyball Olympic performances, showing China's actual results exceeding predictions in 1996 and 2016 under strong coaching leadership. Brazil's decline after 2008-2012 peaks, contrasted with China's 2016 resurgence, demonstrates how great coaches' tactical and psychological guidance can determine Olympic success.

### 6.3 Coach Hiring: An AHP Analysis

Following validation of the "great coach" effect through DID modeling, we developed an AHP and Grey Relational Analysis framework to evaluate optimal coaching investment opportunities for countries in the 2028 LA Olympics. This comprehensive approach helps quantify potential elite coaching impact across various sports.

#### (1). Evaluation Framework Development

To systematically evaluate the potential impact of elite coaches across different sports, we developed four key indicators:

**i. Medal Volatility:** Using five-Olympics medal count standard deviation to measure performance inconsistency, indicating coaching guidance needs.

**ii. Historical Performance Pattern:** Evaluating potential through weighted historical medal analysis - distinguishing between zero-medal patterns and sporadic successes.

$$P = \frac{1}{5} \sum_{t=1}^5 (3G_t + 2S_t + B_t) \quad (18)$$

where  $G_t$ ,  $S_t$ , and  $B_t$  represent gold, silver, and bronze medals in Olympics  $t$ .

#### iii. Competition Intensity (HHI)

The Herfindahl-Hirschman Index reflects competitive balance within each sport. Lower HHI values indicate more dispersed medal distributions, suggesting greater opportuni-

ties for coaching impact:

$$HHI = \sum_{i=1}^n s_i^2 \times 10000 \quad (19)$$

where,  $s_i$  represents country  $i$ 's medal share. HHI values indicate competition levels: <1500 shows high competition, 1500-2500 moderate concentration, and >2500 high concentration in medal distribution.

#### iv. Team Characteristics

Team sports often demonstrate greater potential for coaching impact due to the complexity of coordination and tactical systems. The metric is calculated by the number of athletes required in each sport event.

#### (2). AHP Weight Determination

Using AHP, we established a judgment matrix and calculated indicator weights via eigenvector method. With  $CR = 0.0022 < 0.1$ , the weight results are reliable.

$$A = \begin{bmatrix} 1 & 3 & 2 & 1.5 \\ 1/3 & 1 & 1/1.5 & 1/2 \\ 1/2 & 1.5 & 1 & 1/1.5 \\ 1/1.5 & 2 & 1.5 & 1 \end{bmatrix} \quad (20)$$

The weights of the four evaluation indicators are: medal volatility (38.6%), historical performance (27.7%), team characteristics (19.6%), and competition intensity (14.1%).

### 6.4 Grey Relational Analysis and Results

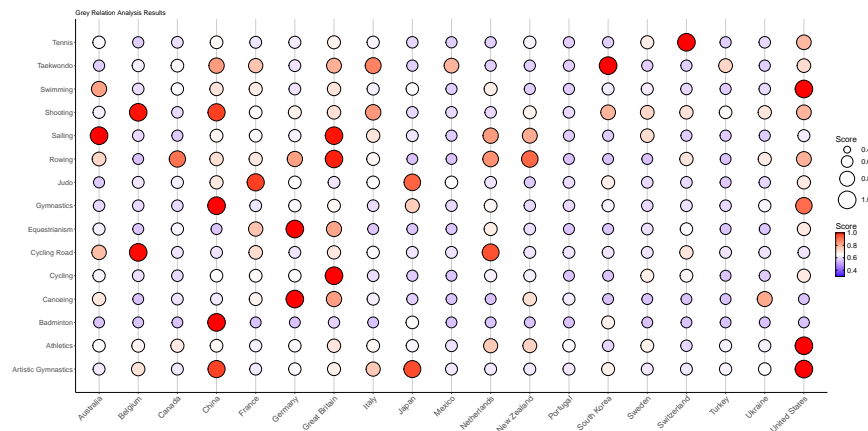


Figure 13: Olympic Coaching Priority Matrix: Cross-National Sport Analysis

Based on the scoring results, Great Britain needs the greatest coaching in sailing and rowing, Belgium in shooting and road cycling, while Australia requires the most coaching assistance in sailing. For the sports mentioned above, in Olympic history, notable "great coach" effects include New Zealand's Dick Tonks (rowing, multiple golds 1994-2016),

China's Wang Yifu (shooting, 21 golds 1989-2017), Britain's Dave Brailsford (transformed cycling through "marginal gains"), and Australia's Victor Kovalenko (sailing, 6 golds since 1997). Using the DID model to estimate their "great coach" effects, the improvement effects for Great Britain (Sailing and Rowing), Belgium (shooting and road cycling), and Australia (sailing) are estimated at 7.3, 17.2, 16.6, and 7.3 respectively.

## 7 Other insights breaking down

### (1) Temporal Trends in Olympic Medal Gender Distribution

Literature review reveals that prediction models for Olympic performance typically achieve higher accuracy in women's events compared to men's events. To further investigate the impact of gender differences on Olympic competition, Figure 14 illustrates the evolution of Olympic medal gender distribution from 1900 to 2020.

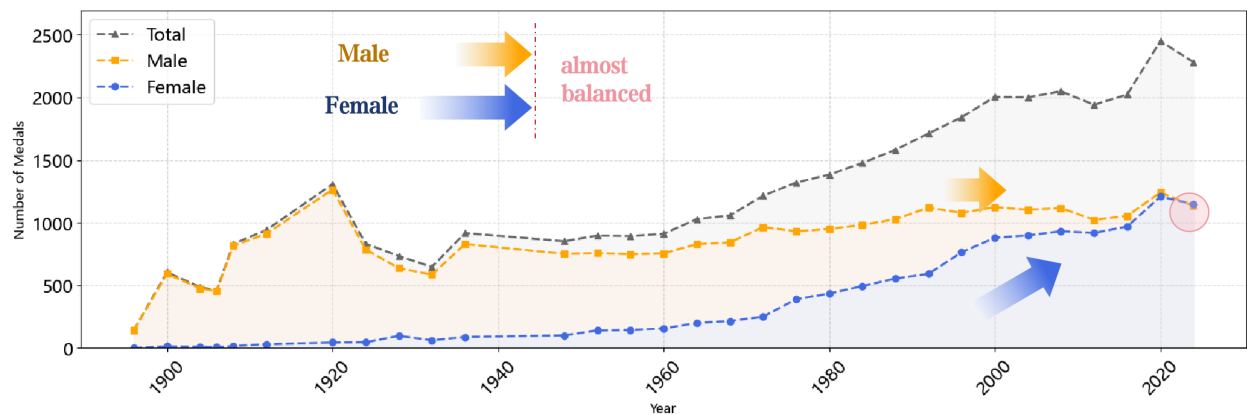


Figure 14: Temporal Trends in Olympic Medal Gender Distribution (1900-2020)

Early 1900s saw male dominance in Olympics. Female participation increased steadily over time. By 2020, medal distribution between genders reached **near-equality**. The growing opportunities for women in sports suggest greater potential for medal wins.

### (2) Cross-national Analysis of Gender Advantage in Olympic Performance

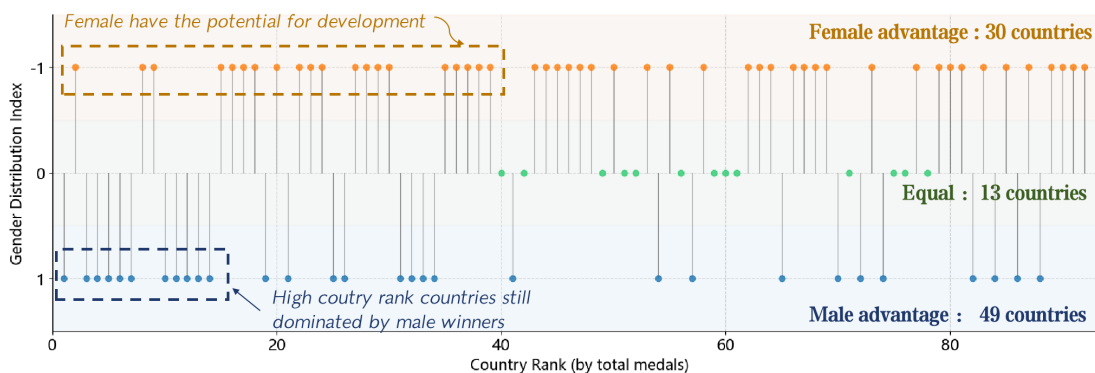


Figure 15: Cross-national Analysis of Gender Advantage in Olympic Performance



First, we calculate the average number of medals for each country across all Olympic Games and determine country rankings. When male medals exceed female medals, it's marked as 1; when equal, marked as 0; and when fewer, marked as -1, resulting in the following figure15.

Even in 2024, top Olympic nations show male-dominant medal wins, though female athletes show strong growth potential across all ranking levels.

Based on this trend, we use the **male athlete participation ratio from the previous two Olympics** as a predictive feature.

### (3) Sensitivity Analysis

#### i. Retention rate of countries in the Tonit model

The retention rate selection is crucial for countries with lower latent competitiveness, effectively mitigating their long-tail effect. This model uses a 5% retention rate, and we conduct sensitivity analysis to observe its impact on the number of countries winning their first medals.

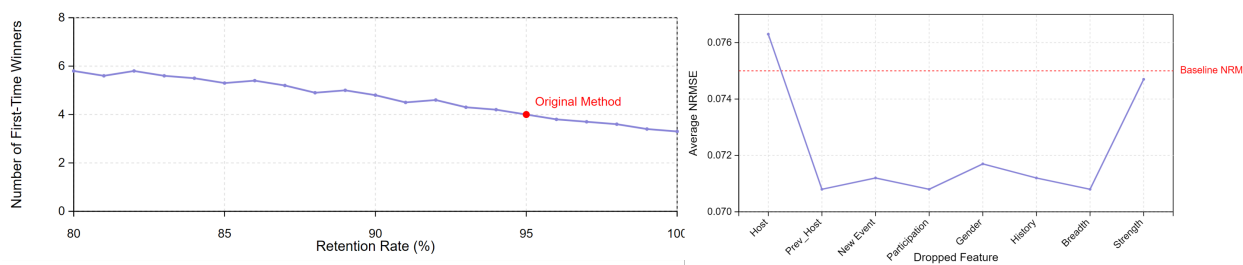


Figure 16: Sensitivity Analysis of Countries' Retention Rate

The sensitivity analysis shows that higher retention rates (80%-100%) lead to fewer first-time winners (from 6 to 3.5). The original method uses 95% retention rate, resulting in 4 first-time winners.

#### ii. Sensitivity Analysis of Medal Prediction Features in Random Forest Model

This analysis reveals key factors affecting medal predictions and improves accuracy.

Sensitivity analysis shows Host and Strength are the most significant predictors, while Previous Host has minimal impact. Other features show moderate influence, and overall model performance remains strong.

## 8 Strengths And Weaknesses

**Strengths:**1. The Tobit-Random Forest joint model effectively addresses the censoring issue of zero threshold data.

2. The DID (Difference-in-Differences) model effectively identifies and quantifies the "great coach" effect.

**Weaknesses:**The AHP-GRA model has high subjectivity, especially in determining weights.

# *New insight:* *Gender Distribution Analysis in Olympic Medals*

To: Olympic Committees  
From: Team#2510704  
Date: January 28, 2025



Dear Distinguished Committee Members,



Ours Olympic gender analysis reveals major equality progress, with near-equal medal distribution by 2020 versus male dominance in the 1900s. While top countries still show male advantage in medals, female athletes display growth potential across all levels. Data shows 49 countries with male advantage, 30 with female advantage, and 13 with gender parity.

These insights offer clear guidance for National Olympic Committees. Supporting female athletes should be a priority. Committees should invest in training programs, resources, and infrastructure to expand opportunities for women in sports. Focusing on disciplines where female potential remains untapped and promoting gender balance in resource allocation can significantly enhance medal prospects.

Moreover, highlighting successful female athletes as role models can inspire greater participation among women. By analyzing gender-specific performance trends, committees can identify areas of strength and refine strategic decisions to maximize medal outcomes. For high-ranking countries, targeted investments in female athlete development, such as recruiting experienced coaches or implementing mentorship programs, can provide a critical competitive edge.

The progress toward gender equality in sports is transformative. By leveraging these insights, National Olympic Committees can accelerate this progress while optimizing performance. We encourage you to incorporate these strategies into your planning to achieve both equity and excellence.

## References

- [1] Sanchez-Fernandez, P., & Vaamonde-Liste, A. (2016). Olympic Medals: Success Predictions for Rio-2016. *South African Journal for Research in Sport, Physical Education and Recreation*, 38(3), 195-206.
- [2] Schlembach, C., Schmidt, S.L., Schreyer, D., & Wunderlich, L. (2022). Forecasting the Olympic medal distribution – A socioeconomic machine learning model. *Technological Forecasting and Social Change*, 175(1), 121314-121328.
- [3] Shi, H., Zhang, D., & Zhang, Y. (2024). Can Olympic Medals Be Predicted? *Journal of Shanghai University of Sport*, 48(4), 26-36.
- [4] Nagpal, P., Gupta, K., Verma, Y., & Kirar, J.S. (2023). Paris Olympic (2024) Medal Tally Prediction. In: Sharma, N., et al. (eds.) *Data Management, Analytics and Innovation. Lecture Notes in Networks and Systems*, 662, 249-264.
- [5] Badoni, P., Choudhary, P., Rudesh, C.P., & Singh, N.T. (2023). Predicting Medal Counts in Olympics Using Machine Learning Algorithms: A Comparative Analysis. *International Conference on Advanced Computing & Communication Technologies (ICACCTech)*, 8(3), 116-121.
- [6] Moolchandani, J., Chole, V., Sahu, S., Kumar, R., Shukla, A., & Kumar, A. (2024). Predictive Analytics in Sports: Using Machine Learning to Forecast Outcomes and Medal Tally Trends at the 2024 Summer Olympics. *International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 4(2), 1987-1992.
- [7] Blais-Morisset, P., Boucher, V., & Fortin, B. (2017). The impact of public investment in sports on the Olympic medals. *Revue Economique*, 68(4), 623-642.
- [8] Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, 35(2), 741-755.
- [9] Vagenas, G., & Palaiothodorou, D. (2019). Climatic Origin Is Unrelated to National Olympic Success and Specialization: An Analysis of Six Successive Games (1996-2016) Using 12 Dissimilar Sports Categories. *Sport in Society: Cultures, Commerce, Media, Politics*, 22(12), 1961-1974.
- [10] Liu, N., Chen, Z., & Bao, G. (2021). Role of media coverage in mitigating COVID-19 transmission: Evidence from China. *Technological Forecasting and Social Change*, 163, 120435.

# Report on Use of AI

## 1.OpenAI GPT 4O

Query1: < Is there a significant clustering effect in the spatial distribution of Olympic medals?>

Output:< Yes, there is a significant clustering effect in the spatial distribution of Olympic medals. Wealthier nations, those with larger populations, and countries with stronger sports infrastructure tend to win more medals, creating clusters of success often concentrated in economically developed regions. Historical, geopolitical, and cultural factors also contribute to this clustering.>

## 2.OpenAI GPT 4O

Query1: <How to select appropriate machine learning algorithms to handle nonlinear features and complex relationships>

Output:< To select appropriate machine learning algorithms for handling nonlinear features and complex relationships, consider the following: tree-based models (e.g., Random Forest, XGBoost, LightGBM, CatBoost) are excellent choices as they naturally handle feature interactions and nonlinearity; Support Vector Machines (SVM) with nonlinear kernels (e.g., RBF or polynomial) are effective for capturing complex decision boundaries; neural networks are well-suited for large datasets and highly complex relationships; k-Nearest Neighbors (k-NN) can handle nonlinearity but may struggle with high-dimensional or large datasets. Additionally, feature engineering techniques like polynomial features or kernel methods can enhance simpler models. Finally, use cross-validation to compare model performance and select the best model based on interpretability, computational cost, and scalability.>

## 3.OpenAI GPT 4O

Query1: <To what extent are new Olympic event selections influenced by host country advantages? How significant is the impact of these selections on the host country's medal performance?>

Output:< New Olympic event selections are often influenced by host country advantages to some extent, as host nations may advocate for the inclusion of sports in which they have historical strengths or a strong competitive edge, aiming to boost their chances of winning medals. This influence is particularly evident when the proposed events are popular domestically but less globally competitive, giving the host country a strategic advantage. The impact of these selections on the host country's medal performance can be significant, as they may secure a greater share of medals in the newly added events, potentially improving their overall ranking. However, the extent of this impact varies depending on the number of new events, the competitiveness of other nations in those sports, and the host country's ability to capitalize on the opportunity through preparation and investment.>