2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Decoding Olympic Medal Success: A Multi-Factor Analysis and Predictive Framework

Yuqian Huang[1], Zishuo Liu[2], Lian Mei[1], and Suohai Fan[3*]

[1] Department of Computer Science, Jinan University, Guangzhou, 510632, China
[2] School of Cyberspace Security, Jinan University, Guangzhou, 510632, China
[3] Department of Mathematics, Jinan University, Guangzhou, 510632, China
huangyuqian@stu2021.jnu.edu.cn, tfsh@jnu.edu.cn

Abstract. A comprehensive and interpretable framework is proposed to forecast and analyze Olympic medal distributions by integrating ensemble machine learning techniques with statistical concentration diagnostics. Utilizing a structured dataset comprising both athlete-level and nation- level features—such as performance records, sport-specific metadata, host advantages, and historical patterns—the framework employs the XGBoost algorithm to predict medal counts for the 2028 Summer Olympics. The model achieves strong predictive performance, particularly for gold medal forecasting (RMSE = 2.42, accuracy = 93.66%), and is validated through rigorous cross-validation procedures. To explore structural disparities in medal allocation, Gini and Herfindahl–Hirschman indices are computed across multiple disciplines, revealing significant concentration in sports like swimming, gymnastics, and athletics, where a limited number of countries consistently dominate podium outcomes. Model interpretability is enhanced using SHAP (SHapley Additive exPlanations), which identifies the relative contributions of demographic, structural, and sport-specific variables to medal predictions. This integrative approach not only enables accurate and explainable Olympic forecasting but also provides actionable insights for evaluating competitive equity and informing national sports investment strategies.

Keywords: Olympic Medal Prediction, TOPSIS, XGBoost, Great Coach Effect, Gini Index, Herfindahl-Hirschman Index, Data Clustering.

1 Introduction

The Olympic medal table serves as a critical indicator of a nation's sporting prowess and overall strength. It draws significant attention from the global public, media, and governmental bodies. For instance, at the 2024 Paris Olympics, both the United States and China achieved a tie for first place in the gold medal tally, each

securing 40 golds. However, the United States surpassed all other nations in the overall medal count, accumulating a total of 126 medals. Meanwhile, France, as the host nation, demonstrated notable success by securing the fourth position in the total medal count. These outcomes underscore the substantial influence of the host country effect, historical performance trends, and the specific impact of different sports events on national rankings.

Addressing the challenge of predicting Olympic medal outcomes requires the integration of various complex factors, including national economic strength, population size, and the number of events in each Olympics. Additionally, the host country effect significantly shapes the performance dynamics of participating nations. In this study, we utilize several key datasets to build predictive models. These datasets encompass historical Olympic medal distributions, which provide medal counts across countries over multiple Olympic cycles; athlete data, detailing participation and performance metrics in individual events; information about host countries, including all summer Olympic host nations and corresponding years; and event data, categorizing the number and types of events in each Olympic Games.

The distribution of Olympic medals is governed by multifaceted and interdependent variables, such as economic and demographic factors, the com–position and number of sports disciplines, and the strategic influence of host nations. Therefore, to develop an accurate model for predicting future medal distributions, it is crucial to incorporate these factors into a sophisticated mathematical framework. In particular, this study places emphasis on identifying nations that are most likely to win their first medals and assessing the strategic event selection decisions made by host countries to maximize their medal outcomes [1] .

2 Related Work

2.1 Traditional statistical models in Olympic medal prediction

Prediction and analysis of competition results is a research process that integrates multiple indicators and data. Early studies mainly relied on econometrics and time series methods. Forrest et al.[2] pioneered the use of linear regression models, combined with macroeconomic variables, using GDP, population size, and the number of medals in previous games to construct a linear regression model to predict the Olympic performance and medal distribution of various countries. However, due to the limitations of linear assumptions, this method cannot capture the nonlinear interaction between athlete–level characteristics (such as participation frequency, medal probability) and national results. At the same time, due to the

rigidity of the parameter structure, it cannot effectively model changes in sports events (such as the reduction of skateboarding in Los Angeles in 2028). Baio G. et al.[3] proposed a Bayesian hierarchical model to predict the results of football matches, taking into account the offensive and defensive strength of the teams, home and away factors, and using the MCMC method to estimate the main effects. Although the prediction accuracy is 95%, the model only emphasizes teams with high goals or concedes, and introduces a hybrid model to reduce over–contraction, which increases the model complexity and calculation time, making it unsuitable for large–scale systems.

2.2 Machine Learning Advances and Interpretability Gaps

In order to overcome the limitations of traditional statistical models, research in recent years has turned to machine learning methods. Schlembach C et al.[4] applied two–stage random forests to a dataset of socioeconomic variables for prediction, demonstrating higher accuracy. However, this method is highly data–dependent and prone to overfitting. It also lacks dynamic adaptability and is difficult to predict sudden factors (such as the impact of COVID–19 on training). Igiri C P et al.[5] used knowledge discovery technology (KDD) and artificial neural networks (ANN) to build a more comprehensive system with higher prediction accuracy. However, the ANN model has poor interpretability and cannot clearly define the meaning of weights. KDD has the problem of feature selection bias at the data level. Therefore, even the most advanced machine learning models lack insights that can provide specific actionable suggestions and cannot optimize medal results. Our research fills this gap through interpretable multi–scale integration.

2.3 Competitive Imbalance Analysis and Causal Inference Shortfalls

Research on competitive imbalance is concentrated in the field of economics. Owen et al.[6] used the Herfindahl–Hirschman Index (HHI) to quantify the monopoly phenomenon in sports leagues, while Davidson[7] refined the calculation method of the Gini coefficient. However, these studies have significant flaws. The application of the HHI and the Gini coefficient fails to take into account the unique sports event clustering phenomenon and the dominance of athletes in the Olympics. Secondly, research on "great coaches" mostly remains at the qualitative analysis level and lacks differentiation analysis (DID) verification. However, Nachar's[8] U–test has never been applied to the quantitative analysis of coaching effects. Therefore, the disconnect between competition indicators and causal mechanisms seriously affects the effectiveness of policy making. Our study makes up for these shortcomings by

integrating these two fields, innovatively adopting the DID method to evaluate the impact of coaching, and enhancing monopoly visualization through t–SNE.

3 Data Description and Preprocessing

3.1 Data Sources

We collected the following data from the official website of the Olympic Games[9]:

- **Medal Table Dataset:** Records gold, silver, bronze, and total medal counts for all countries from 1988 to 2024.
- **Athlete Dataset:** Includes athlete–level participation details, such as coun– try, event, year, and medal result.
- **Event Dataset:** Provides year–specific sport and discipline information, with counts of all contested events.
- **Host Nation Dataset:** Lists the host country and city for each Olympic year.

These datasets collectively support both macro–level(country and medal) analyses and micro–level(athlete and event)modeling. All analyses in this study are strictly based on these datasets, in accordance with the modeling constraints.

3.2 Data Cleaning and Harmonization

To ensure consistency and reduce noise, we applied systematic data cleaning and integration operations across the four datasets. Country codes were first standardized using ISO–3166 conventions to align with IOC usage (e.g., "United States" → "USA"). Athlete names and event labels were normalized using Levenshtein distance to correct inconsistencies across different years.

Olympic events were categorized into 25 high–level sport types to address structural variations such as weight classes in combat sports. Incomplete or irrelevant columns—such as discipline names or sports federations—were removed. Missing values caused by historical anomalies or weather disruptions were filled with zeros to preserve matrix structure and dimensionality.

Athlete–level data was aggregated to produce more interpretable and model–relevant features, such as participation count, medal probabilities, and last active year. Additionally, contextual variables like host city and sport type were retained as proxy indicators of national advantage and sport–specific investment. Table 1 summarizes the derived features used for athlete evaluation.

3.3 Feature Extraction and Encoding

Table 1. Derived Features for Athlete–Level Analysis.

Feature	Description
Participation Count	Total number of Olympic appearances per athlete
Last Participation Year	Most recent year of participation
Medal Probability	Ratio of medal–winning events to total events
Total Medals	Number of medals won by the athlete
Gold Medals	Number of gold medals won
Nationality	Athlete's representing country
Sport Type	Sport category participated in
Host City	City of the Olympic Games participated in

To enhance downstream modeling performance, structured features were extracted from the cleaned datasets. These include both categorical and numerical fields relevant to predicting Olympic outcomes. Table 2 lists an overview of key features. These extracted variables serve as the foundation for both descriptive analysis and predictive modeling in subsequent sections.

Table 2. Key Features Extracted for Medal Modeling.

Feature	Example	Type
NOC (Country Code)	USA	String
Programs (Sport Code)	SWA (Swimming)	Categorical
Host City	Los Angeles, United States	String

4 Methodology

4.1 Motivations

To address the multifactorial challenge of Olympic medal prediction, this study adopts a unified modeling strategy that integrates supervised machine learning with structural pattern analysis. Rather than relying on a single predictive model, we propose a multi-perspective pipeline designed to capture both the quantitative determinants of national performance and the structural dynamics that influence competitive outcomes—such as host advantage and event re- structuring. At the core of our approach is an ensemble-based predictive model constructed using the XGBoost algorithm[10], which is trained on historical Olympic data including

medal distributions, event types, host cities, and athlete-level statistics. XGBoost is selected for its robustness to multicollinearity, ability to capture nonlinear feature interactions, and strong generalization performance across structured data. This predictive model estimates medal counts for each nation in the 2028 Los Angeles Olympics and also provides confidence intervals through bootstrapping, enabling probabilistic forecasts. To analyze the structural concentration of medal distribution across different sports, we compute the Gini coefficient [6] and Herfindahl–Hirschman Index (HHI) [7] . These statistical indicators measure the extent to which a few dominant countries monopolize specific sports disciplines. High values in these in- dices indicate reduced competitiveness and increased specialization, thereby helping to explain medal inequality in events such as swimming or weightlifting, which tend to be dominated by a handful of nations.

Overall, the methodology aims to solve four core tasks: (1) predict total and class-specific medal counts per country in 2028; (2) estimate the likelihood of countries winning their first-ever Olympic medals; (3) evaluate how changes in event structure and host nation dynamics affect medal outcomes; and (4) assess competitive concentration and inequality using structural indicators. All modeling is conducted strictly within the constraints of the provided dataset, without incorporating external socioeconomic variables such as GDP or population.The primary notations used in this study are summarized in Table 3.

Table 3. Notations.

Symbol	Definition
M_{gold}	Predicted gold medals for a country.
M_{silver}	Predicted silver medals for a country.
M_{bronze}	Predicted bronze medals for a country.
M_{total}	Total predicted medals ($M_{gold} + M_{silver} + M_{bronze}$).
$S_{athlete}$	Athlete's performance score (TOPSIS).
P_{medal}	Probability of winning a medal.
β_3	Coach impact coefficient (DID model).
G	Gini Index for medal inequality.
HHI	Herfindahl–Hirschman Index for monopoly.
d_{ij}	Distance in t–SNE clustering.
$T_{program}$	Total events in a sport.
$W_{entropy}$	Weight (entropy method).
α	Sigmoid function control parameter.
β	Sigmoid function control parameter.
$C_{cluster}$	Cluster index in analysis.
x_{ij}	Value of jth feature for ith athlete/country.

4.2 Medal Prediction Using Sport-Intrinsic Features

To perform the entropy–weighted TOPSIS evaluation[11], we extracted a range of athlete–level features from the dataset, including **Participation Count** (e.g., 30), **Last Participation Year** (e.g., 2016), **Gold Medal Count** (e.g., 23), **Medal-Winning Probability** (e.g., 0.93), **Nationality** (e.g., USA), and **Sport Type** (e.g., Swimming). This multi–dimensional information forms the basis for computing entropy–based weights and determining each athlete's composite performance score.

Let m denote the number of athletes and n the number of features.The raw decision matrix $X \in \mathbb{R}^{m\times n}$ is defined as:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad (1)$$

where x_{ij} represents the value of the i–th athlete on the j–th feature.

To address the heterogeneity in feature scales and distributions, we employ the **Entropy Weight Method** to compute objective weights for each feature. These weights are then integrated into the **TOPSIS** framework to derive a comprehensive performance score for each athlete. We first normalize the decision matrix using vector normalization:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} \cdot x_i^2}}, 1 \le i \le m, 1 \le j \le n. \quad (2)$$

The normalized proportion p_{ij} is computed as $p_{ij} = \frac{r_{ij}}{\sum_{i=1}^{m} r_{ij}}$. The entropy e_j of each feature j quantifies its information diversity across athletes.The weight w_j derived from feature j is expressed as $w_j = \frac{1-e_j}{\sum_{k=1}^{n}(1-e_k)}$.

$$e_j = -\frac{1}{\ln m}\sum_{i=1}^{m} p_{ij} \ln p_{ij}, \ \ where\ 0 \le e_j \le 1. \quad (3)$$

Features with higher entropy (i.e., more uniform distribution across athletes) receive lower weights, while those with lower entropy (greater discriminatory power) are assigned higher weights. **Feature Classification for TOPSIS Evaluation** We classify each feature based on its relationship with athletic performance:

Benefit-type features (J_1): Higher values indicate better performance

- Participation Count: More Olympic appearances reflect experience
- Gold Medal Count: Direct measure of elite success
- Medal Probability: Success rate in medal-winning events

Cost-type features (J_2): Lower values indicate better performance

- Years Since Last Participation: Computed as (2024-Last Participation Year), where lower values indicate recent activity

This classification ensures proper determination of ideal solutions A^+ and A^-.

We define the ideal solution A^+ (best possible performance) and the negative ideal solution A^- (worst possible performance) as:

$$A^+ = (\max_i r_{ij} \,|j \in J_1, \min_i r_{ij} | \, j \in J_2), \tag{4}$$

$$A^- = (\min_i r_{ij} \,|j \in J_1, \max_i r_{ij} | \, j \in J_2), \tag{5}$$

where J_1 and J_2represent benefit–type and cost–type features, respectively (e.g., medal counts are benefit–type, while recency may be cost–type). For each athlete i, we compute the distances from the ideal and negative ideal solutions:

$$d_i^+ = \sqrt{\sum_{j=1}^{n} w_j(r_{ij} - r_j^+)^2}, \qquad d_i^- = \sqrt{\sum_{j=1}^{n} w_j(r_{ij} - r_j^-)^2} \tag{6}$$

where r_j^+ and r_j^- are the normalized values of the ideal and negative ideal solutions for feature j. The final TOPSIS score for athlete i is given by:

$$\text{TOPSIS\_Score}_i = \frac{d_i^-}{d_i^+ + d_i^-}. \tag{7}$$

This score ranges from 0 to 1, with higher values indicating better overall performance relative to other athletes. The TOPSIS evaluation obtained by processing dataset is as follows: **TOPSIS_Score:** The comprehensive performance score for each athlete, reflecting their relative strength across all features. **Total_Score:** Sum of TOP -SIS_ Scores for all athletes from a country. **Mean_Score:** Average TOP1SIS_ Score per athlete in the country. **Min_Score:** Minimum TOPSIS_Score among the country's athletes. **Max_Score:** Maximum TOPSIS_Score among the country's athletes. **Variance_Score:** Variance of TOPSIS_Scores, indicating consistency in performance. The Olympic Games evolve not only in athletic performance but also in pro– gram structure. Due to regional preferences and international policy trends, changes in the number and type of Olympic events are expected. For instance, the 2028 Olympics will be hosted in Los Angeles, and American sports such as cricket (CKT) are likely to gain prominence. In contrast, sports with higher risks like skateboarding (SKB) may experience a reduction. The simulated changes in program counts between 2024 and 2028 are shown in Figure 1. To ensure that our model effectively handles these dynamic inputs, categorical variables such as nationality, host city, and sport type are processed using one–hot encoding. For medal counts, we apply a $\log(x + 1)$ transformation to reduce data skew, smooth variability, and enhance model generalizability. At the country level, athlete performance is first aggregated using the TOP– SIS method[12] .From this, we

calculate national statistics including total score (**Total_Score**), mean score **(Mean_Score**), and extremal values (**Min_Score** and **Max_Score**). Using these aggregated features, we train an XGBoost model to predict gold, silver, and bronze medal counts. XGBoost, an efficient ensemble method based on gradient boosting of decision trees, is particularly suitable for structured tabular data with both numerical and categorical features.

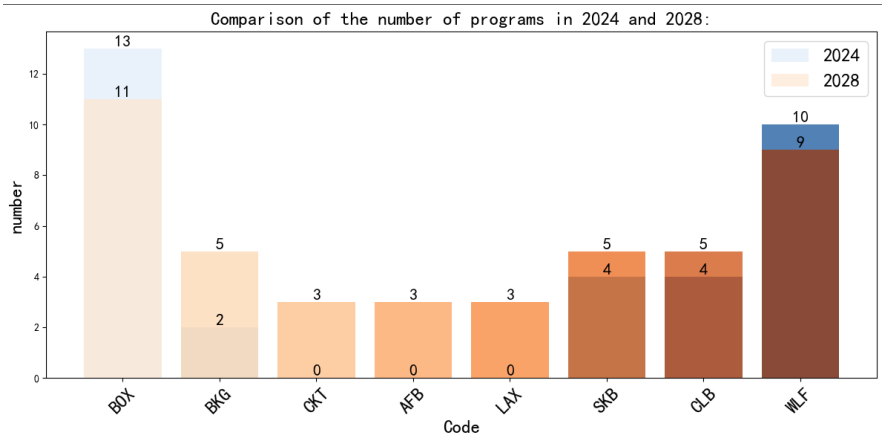


Fig. 1. Comparison of the number of programs in 2024 and 2028.

Algorithm 1 Difference–in–Differences Analysis

Input: Dataset containing:
    TOPSIS Mean Score: TOPSIS score for each athlete
    Medal Count: The number of medals in a certain event.For example, gold, silver and bronze in swimming.
    Is Coached: Binary treatment indicator (1=coached by coach, 0=otherwise)
Step 1: Prepare the data
    Construct interaction term $\text{group}_i \times \text{time}_t$
if Parallel trend assumption is satisfied **then**
        Plot pre–treatment trends of Medal Count by Is Coached group
end if
Step 2: Define the DID model

$$Y_{it} = \beta_0 + \beta_1 \cdot group_i + \beta_2 \cdot time_t + \beta_3 \cdot (group_i \times time_t)$$
$$+\beta_4 \cdot \text{trend}_t + \beta_5 \cdot (\text{trend}_t \times \text{group}_i) + \epsilon_{it}$$

Step 3: Estimate the model
    Apply OLS regression to obtain coefficient estimates
Step 4: Evaluate the results

16: **if** $\beta_3 > 0$ and $p < 0.1$ **then**
17: Conclude significant positive coaching effect on medals
18: **else**
19: Find no statistically significant coaching effect
20: **end if**
21: **Output:** Treatment effect β_3 and model diagnostics

4.3 Structural Inequities in Olympic Medal Distribution

The Olympic Games are often perceived as a celebration of global diversity and fair competition. However, our analysis reveals that beneath the surface lies a set of structural inequities that significantly affect the distribution of medals. In particular, we identify patterns of **monopoly of specialization** in certain sports, where a limited number of countries or individual athletes dominate the podium. To quantify these inequalities, we employed the **Gini index** and the **Herfindahl–Hirschman Index (HHI)**, both of which are widely used in economics to measure concentration and inequality. These metrics were computed for each sport based on historical medal distributions, providing a lens into competitive balance across disciplines. To evaluate the inequality of medal distribution across countries in each sport, we computed both the Gini index and the Herfindahl-Hirschman Index (H Hl). The Gini index is defined as $G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\overline{x}}$, where x_i represents the number of medals won by the *i*–th country, n is the number of participating countries, and $\overline{x}$ is the average medal count. A Gini index of 0 indicates perfect equality, while a value close to 1 implies severe concentration. Complementarily, the HHI is calculated as $\text{HHI} = \sum_{i=1}^{N} s_i^2$, where s_i denotes the share of medals earned by country i. The HHl ranges from $\frac{1}{N}$, representing a highly from competitive field, to 1, indicating a complete monopoly. The key interaction term β_3 in Equation 8 captures the causal effect of elite coaching:

$$Y_{it} = \beta_0 + \beta_1 \cdot group_i + \beta_2 \cdot time_t + \beta_3 \cdot (group_i \times time_t) + \beta_4 \cdot \text{trend}_t + \beta_5 \cdot (\text{trend}_t \times \text{group}_i) + \epsilon_{it} \quad (8)$$

While existing research attributes Olympic medal counts to macroeconomic indicators such as national GDP, this study identifies a significant yet over– looked factor: the impact of exceptional coaching. Although GDP remains a fundamental predictor of a nation's sporting success, we argue that coaching quality constitutes a critical yet underexplored determinant. To empirically validate this hypothesis, we employ the Difference–in–Differences (DID) method to isolate and measure the causal effect of coaching on medal performance. Our empirical analysis follows a

structured DID framework as outlined in Algorithm 1. The methodology is as shown above.

4.4 Uncovering the Non-Economic Drivers of First-Time Olympic Medals

The proposed scoring model employs an entropy–weighted fusion approach combined with logistic calibration to predict the probability of a nation winning its first Olympic medal. This probabilistic framework enables developing countries to assess the potential return on short–term Olympic investments. The model operates through three key computational steps. The model's working mechanism can be comprehensively described as follows: The algorithm first computes feature weights based on statistical significance. For each feature i, the weight w_i is determined by: $w_i = \frac{\frac{1}{p_i}}{\sum_{j=1}^{n}\frac{1}{p_j}}$, where the p value was obtained through the Mann–Whitney U test[8]. This entropy– weighting approach assigns higher weights to features with smaller p–values (p_i), reflecting greater statistical significance in predicting first–medal outcomes. The normalization ensures the weights sum to unity.The weighted features are combined into a composite score through linear aggregation: Score $= \sum_{i=1}^{n} w_i \cdot x_i'$, where x_i' represents the normalized feature values, and w_i are the computed weights from Step 1. This step produces a dimensionless score that synthesizes all predictive information while accounting for feature importance. The final probability of winning a first medal is obtained by mapping the composite score through a logistic function: $P(\text{Medal}) = \frac{1}{1+e^{-\alpha\cdot(Score-\beta)}}$. The parameters α (steepness) and β (inflection point) control the probability transformation. A threshold of 0.5 is typically used to classify nations into "likely" or "unlikely" to win their first medal. This three–stage pipeline effectively transforms raw input features into probabilistic predictions while maintaining interpretability through its transparent weighting and calibration mechanisms. The model's design particularly benefits developing countries by quantifying their Olympic medal potential based on measurable indicators, thereby informing strategic investment decisions.

5 Experiments and Results

5.1 Medal Prediction Using Sport-Intrinsic Features

To assess the predictive quality of our model, we define and calculate five evaluation metrics including Root Mean Square Error (RMSE), M1, which measures perfect prediction accuracy, M2 which evaluates the model's ability to identify all genuine

medalists, M3 which quantifies correct predictions for countries that realistically won't medal, M4 which reflects robustness against small prediction errors, showing correct classification within a practical 2 medal range. These metrics are summarized in Table 4. The performance results for each medal type are detailed in Table 5. It demonstrates strong predictive capability for gold and silver medals, achieving exact match rates exceeding 35% and confidence–interval accuracy above 93%. Predictions for bronze medals are slightly less accurate, likely due to higher dispersion in lower–tier podium finishes. Notably, the model achieves over 65% accuracy in predicting countries that will not win medals, indicating good sensitivity to sparse outcomes. The high M4 values across all medal types confirm the model's robustness within a practical error (±2).

Table 4. Evaluation Metrics for Medal Prediction.

Metric	Description
M_1	Percentage of countries for which the predicted medal count is exactly correct
M_2	Percentage of countries with correct non–zero medal predictions
M_3	Percentage of countries where the model predicts exactly the same and non–zero
M_4	Accuracy of predictions within a ± 2 medal margin (95% confidence)

Table 5. Evaluation Metrics for Medal Predictions.

Metric	Gold	Silver	Bronze
RMSE (medal)	2.42	2.09	2.52
M_1 (%)	37.32%	35.21%	22.89%
M_2 (%)	24.49%	25.25%	21.30%
M_3 (%)	65.91%	59.76%	27.94%
M_4 (%)	93.66%	93.31%	94.72%

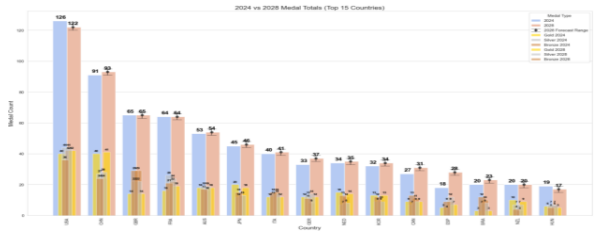


Fig. 2. Predicted Total Medals for Top 15 Countries in 2028 (Reference: 2024)

Figure 2 presents the projected total medal counts for the top 15 countries, comparing 2024 (Paris) with the expected outcomes in 2028 (Los Angeles), along with 95% confidence intervals. The United States maintains its dominant position but is expected to experience a slight decline in total medals, dropping from 126 in 2024 to 122 in 2028. China is projected to see a modest increase, rising from 91 to 93 medals. Other traditional powerhouses such as the United Kingdom and France show stable medal projections, suggesting consistent performance despite potential changes in event structures or athlete rosters. These outcomes reflect a strong historical momentum and stable athlete pipelines within these nations.

To better understand shifts in Olympic performance, we visualize net changes in total medal counts for the same set of top–performing countries (Figure 3). Spain shows the largest projected gain (+10 medals), possibly driven by improvements in team sports and niche disciplines. Germany, Canada, and Brazil each gain approximately 4 medals, highlighting progress in athlete development programs. Conversely, the United States is forecasted to lose four medals, reflecting increased competition from emerging countries and possible saturation in dominant sports. Other nations with anticipated medal losses include Romania, Kyrgyzstan, and Sweden, each facing declines due to lower event‐level competitiveness or structural changes in qualifying rules.

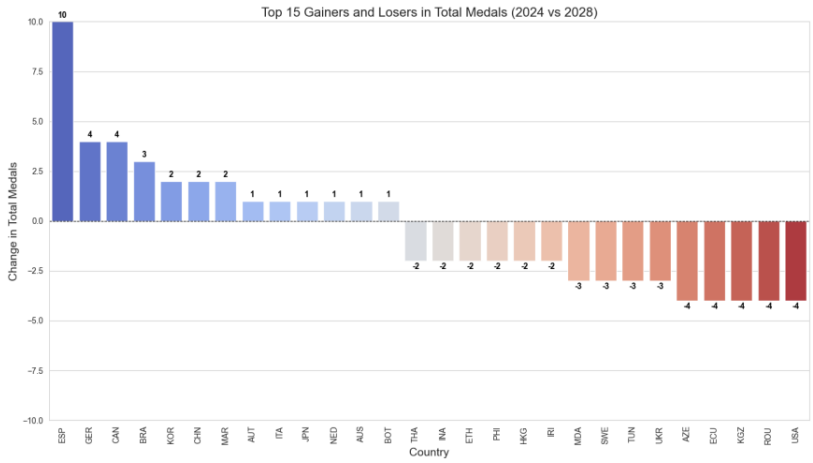
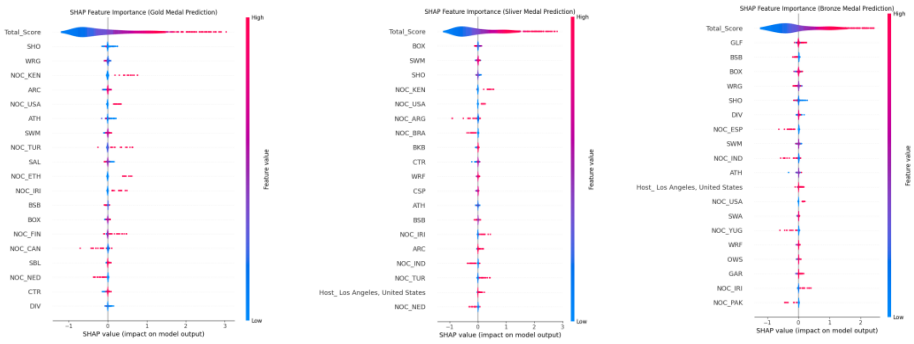


Fig. 3. Net Change in Medal Counts (2024 vs. 2028): Top 15 Nations.

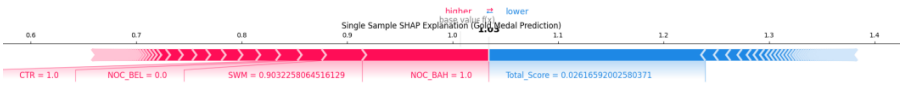
To enhance interpretability, we also applied SHAP (SHapley Additive exPlanations) to quantify each feature's contribution. The Total_Score was the most important predictor across all medal types.In gold medal predictions, sport– specific

features such as SWM (swimming) and control variables like CTR showed positive influence. For silver and bronze medals, nationality–based factors like NOC_BAH and performance metrics like SHO (shooting) and GLF (golf) played key roles. The global SHAP feature importance values are visualized in Figure 4, showing the relative influence of features for each medal type. To further elucidate the model's reasoning at the individual prediction level, we present SHAP force plots for single sample predictions in Figure 5. These plots highlight how different features contribute positively or negatively to the final prediction score.The global SHAP feature importance values are visualized in Figure 4, highlighting relative contributions across all features. To further interpret the model's decision–making at the instance level, we present SHAP force plots in Figure 5, which illustrate how individual variables drive specific national predictions upward or downward. These tools collectively enhance transparency and offer practical guidance for national sports strategies.

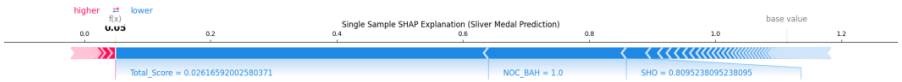


(a) Gold (b) Silver (c) Bronze

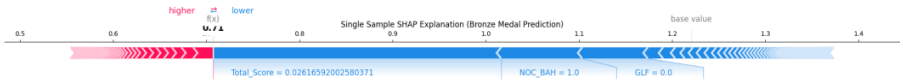
Fig. 4. SHAP feature importance for gold, silver, and bronze medal predictions.



(a) Gold Medal Prediction



(b) Silver Medal Prediction



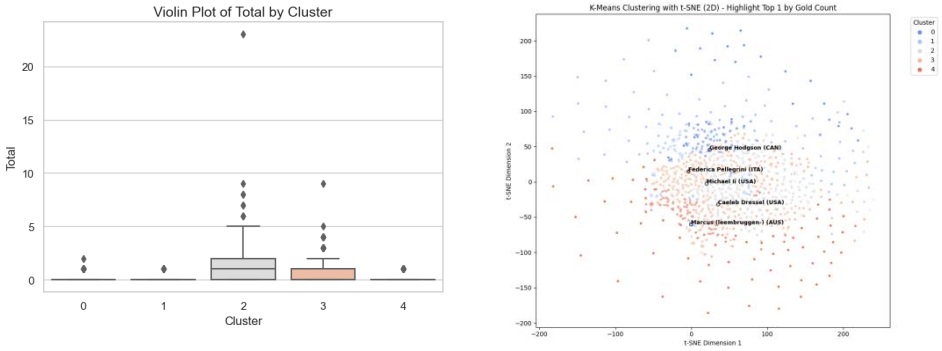
(c) Bronze Medal Prediction

Fig. 5. Local SHAP Explanation for Individual Country Predictions.

5.2 Structural Inequities in Olympic Medal Distribution

Olympic medal distributions reveal persistent structural inequities across sports disciplines and nations. Our findings indicate that events such as **swimming**, **gymnastics**, and **track and field** consistently exhibit high values of the Gini index and Herfindahl-Hirschman Index (HHI), metrics that capture concentration and inequality. For instance, in swimming, over 60% of the gold medals from 2000 to 2024 were claimed by only three countries, resulting in an HHI above 0.35 and a Gini index consistently exceeding 0.6. These figures point to entrenched advantages enjoyed by a select group of nations, often due to historical investments in coaching pipelines, sports infrastructure, and youth training systems. To gain deeper insights into the manifestation of such concentration at the individual level, we performed a detailed case study on swimming, a discipline with high medal density and frequent repeat participation. The structure of swimming events — spanning freestyle, butterfly, medley, and relays—enables a small number of exceptional athletes to accumulate medals at a much higher rate than others. We clustered Olympic swimmers using a modified Canopy‑KMeans algorithm based on performance statistics and metadata[13], and employed t‑distributed stochastic neighbor embedding (t‑SNE) for dimensionality reduction and visualization. As shown in Figure 6, the left subfigure presents a box plot of gold medal counts per swimmer, highlighting a highly skewed distribution. While most athletes win only one or two medals, a few outliers dominate the record books. Notably, Michael Phelps (USA) stands out with an unprecedented 23 gold medals, followed by Caeleb Dressel (USA) with 9 golds, forming distinct high‑density clusters in the embedding space. In contrast, swimmers such as George Hodgson (CAN), Marcus Leembruggen (AUS), and Federica Pellegrini (ITA) each secured only one or two gold medals, clustering in the low‑dominance region. This distribution underscores the monopolistic nature of elite performance in swimming, where a small cohort of athletes—often from a few dominant nations—account for a disproportionate share of Olympic success. Notably, **Michael Phelps** of the USA demonstrates an unparalleled level of

domin-ance, securing 23 Olympic gold medals—far exceeding his closest peers. The t–SNE projection in the right panel shows well–separated clusters, with elite performers such as Phelps and Dressel forming isolated nodes away from the dense core of average competitors. These athlete–level patterns reinforce the macro–level findings of structural concentration and resource disparity. In terms of actionable insights, national Olympic committees (NOCs) should consider fostering cooperative programs that promote shared access to elite coaching, data infrastructure, and cross–border training camps. Furthermore, the International Olympic Committee (IOC) may explore mechanisms to maintain competitive diversity, such as introducing athlete participation caps per event, expanding event quotas for underrepresented countries, or rotating event formats to disrupt entrenched advantages. By addressing both the structural and individual dimensions of medal monopolization, the Olympic movement can move closer to its ideals of fairness, inclusivity, and global representation.



(a) Box plot of the gold distribution

(b) t–SNE Visualization of Swimming Clusters with Gold Leaders

Fig. 6. Comparison of Gold Distribution and t–SNE Visualization in Swimming.

5.3 Predictive Analysis of Nations Likely to Secure Their First Olympic Medal

A crucial aspect of Olympic forecasting lies in identifying nations poised to earn their first–ever medal—an indicator of expanding global competitiveness. Using a binary XGBoost classifier trained on sport–intrinsic features (e.g., athlete count, qualification breadth) while excluding socioeconomic factors, we estimate each non–medalist country's likelihood of medal success. Two standout candidates emerge: AIN (Athletes under a Neutral Flag) with a high probability of 0.81 and odds of 1.24, driven by consistent cross–discipline performance and near–podium finishes; and EOR (Refugee Olympic Team) with a probability of 0.53 and odds of

1.89, reflecting growing parity in qualification and support systems. As shown in Figure 7, both cases highlight the rising competitiveness of previously underrepresented entities. These results bear important policy implications. By directing mentorship, funding, and global cooperation toward such emerging contenders, the IOC and NOCs can reinforce values of inclusion and equity. More broadly, this analysis offers a strategic tool for monitoring competitive diffusion in Olympic sport and guiding targeted development interventions.

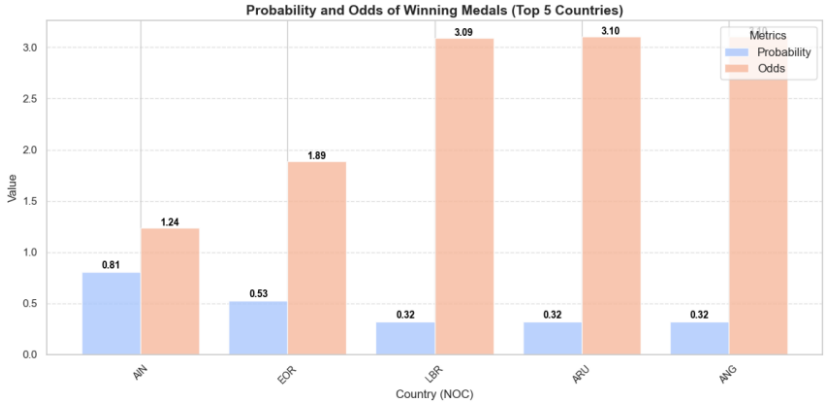


Fig. 7. Predicted Probabilities and Odds for First–Time Medal–Winning Nations.

6 Conclusion

This study proposes a robust and interpretable framework for predicting Olympic medal distributions by integrating machine learning, causal inference, and structural analysis. The XGBoost model, informed by carefully engineered features, demonstrates high predictive accuracy, while SHAP enhances interpretability of key drivers across medal types. Moreover, concentration metrics and athlete clustering reveal persistent structural inequalities, with a small group of nations and individuals dominating high–multiplicity events. These findings highlight the interplay between talent, institutional investment, and leadership, offering actionable insights for national sports policy and future research in sports analytics.

Acknowledgments. This work was supported by the National Undergraduate Innovation and Entrepreneurship Training Program of China (Project No. 202510559076) at Jinan University, a nationwide initiative administered by the Ministry of Education. We gratefully acknowledge the funding and project supervision provided through Jinan University during the full course of this research.

References

1. G. Csurilla and I. Fert˝o. How to win the first Olympic medal? And the second? Social Science Quarterly, vol. 105, no. 5, pp. 1544-1564 (2024)
2. D. Forrest, I. Sanz and J. D. Tena. Forecasting national team medal totals at the Summer Olympic Games. International Journal of Forecasting, vol. 26, no. 3, pp. 576-588 (2010)
3. G. Baio and M. Blangiardo. Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics, vol. 37, no. 2, pp. 253-264 (2010)
4. C. Schlembach, S. L. Schmidt, D. Schreyer and L. Wunderlich. Forecasting the Olympic medal distribution-A socioeconomic machine learning model. Technological Forecasting and Social Change, vol. 175, p. 121314 (2022)
5. C. P. Igiri and E. O. Nwachukwu. An improved prediction system for football match result. IOSR Journal of Engineering, vol. 4, no. 12, pp. 12-20 (2014)
6. P. D. Owen, M. Ryan and C. R. Weatherston. Measuring competitive balance in professional team sports using the Herfindahl–Hirschman index. Review of Industrial Organization, vol. 31, pp. 289-302 (2007)
7. R. Davidson. Reliable inference for the Gini index. Journal of Econometrics, vol. 150, no. 1, pp. 30-40 (2009)
8. N. Nachar. TheMann–Whitney U: A test for assessing whether two independent samples come from the same distribution. Tutorials in Quantitative Methods for Psychology, vol. 4, no. 1, pp. 13-20 (2008)
9. Olympics.com. Olympics.Com l Olympic Games, Medals, Results & Latest News. [Online]. Available: https://www.olympics.com/en/.
10. T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794 (2016)
11. Z. Liang, J. Nie, Q. Ou, et al.: A Dual–Weighting TOPSIS Approach to Evaluating Olympic Sports. Advances in Engineering Technology Research, vol. 13, no. 1, pp. 1359-1359 (2025)
12. P. Cerda and G. Varoquaux. Encoding high–cardinality string categorical variables. IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 3, pp. 1164-1176 (2020)
13. A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija and J. Heming. K–means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, vol. 622, pp. 178-210 (2023)