World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3053

NONLINEAR PREDICTION BASED ON 2028 OLYMPIC EVENTS AND MEDALS

TiLiang Zhang, JunJie Chen, Cheng Cheng, Xing Li*
School of Mathematics and Statistics, Hubei University of Education, Wuhan 430205, Hubei, China.
Corresponding Author: Xing Li, Email: xingli@zjut.edu.cn

Abstract: Amid the expanding scale of the Olympic Games, precise forecasts of the event portfolio and medal allocation have become critical for national strategic planning. Olympic data, however, are markedly non-linear and structurally dynamic, rendering traditional linear methods inadequate. This paper therefore develops an integrated forecasting framework that estimates discipline-level event counts and country-specific medal shares for the 2028 Games. Athletes were first aggregated by team and country, and analyses were conducted at the discipline level to prevent overgeneralisation inherent in sport-level aggregation. Support Vector Regression was employed to model the relationship between historical covariates and the number of events per discipline; the resulting predictions achieved a mean-squared error of 1.744 and an R^2 of 0.634. The strategic salience of each discipline to individual nations was subsequently quantified via weighted medal totals and visualised through rose plots. Medal shares were derived by mapping historical performance indicators to fractional medal outcomes using XGBoost, after an initial recurrent architecture exhibited convergence difficulties. These fractions were scaled by the projected event counts, and a calibrated 15 % host-nation uplift was applied to the United States before global normalisation. The resulting projection allocates 47, 45 and 36 medals to the United States, 35, 24 and 15 to China, and 18, 9 and 10 to Japan. Retrospective validation against 2024 data places all nine reference nations within 95 % prediction intervals, confirming the framework's reliability. This study can provide data support for national sports management departments and optimize the allocation of training resources.

Keywords: Discipline-level events; XGBoost; Olympic forecasting; Support Vector Regression; Resource allocation

1 INTRODUCTION

Following the conclusion of the 2024 Paris Olympics, nations are turning to the 2028 Los Angeles Games. As a globally celebrated sporting event, the Olympics serves as not only a world stage for athletes to showcase their exceptional sportsmanship, but also a platform for cultural exchange and fostering friendship among nations. Consequently, accurately forecasting a nation's medal count at the Olympics is of paramount importance to sports governing bodies and athletes alike.

The Olympic Games is an international sporting event, the medal tally of which has always been a focal point of attention and analysis. There now exists an extensive body of research pertaining to medal and result projection. For example, Zhang Bo predicted the gold-medal result of women's shot put in 2012 based on GM(1,1) prediction model in Gray System Theory, a more applicable solution when there is lack of data [1]. More methods on the basis of machine learning algorithms were provided by Jhankar Moolchandani et al [2], including Linear Regression, Random Forest, Support Vector Machines and Neural Networks. They are more useful for forecasting the medal count according to the athlete's attributes and country information. Among them, Random Forest and SVM stood out. Noviyanti T M Sagala and Muhammad Amien Ibrahim compared XGBoost, LightGBM and CatBoost and found that XGBoost had the highest accuracy [3]. Wang Shiyu established a BP neural network prediction model by examining the impact of five factors, including the number of medals won in the previous Olympics, total population, per capita GDP, social system, and host country, on the ability to win Olympic medals. This model achieved the prediction of the top ten medals in the 2020 Tokyo Olympics medal table [4]. Dong Qi et al. used support vector machine nonlinear extended samples to determine the order of time series models. By analyzing the changes in the support vector set after adding new samples to the training set, they constructed a support vector machine model for predicting Olympic gold medals. Compared with traditional time series prediction, this model has the characteristics of low subjectivity, high prediction accuracy, and better prediction stability [5]. Yan Yuyang used grey theory to predict that China will win 93 or 94 medals in the 2012 London Olympics by modeling and analyzing the number of medals won by China in the past 6 Olympic Games [6]. Luo Yubo et al. used the grey prediction GM (1,1) model, combined with the host effect, to predict China's medal count and total score at the Beijing Winter Olympics, and also provided a world ranking prediction for China's gold medal count. The results show that the host effect of the Winter Olympics shows a decreasing trend, but still has a significant effect. With the home advantage in competition and preparation, China will win 6-7 gold medals at the Beijing Winter Olympics, ranking in the top 10 on the gold medal table [7]. DingShu Yan constructed a Long Short Term Memory (LSTM) model using historical data from the Summer Olympics (1896-2024), including medal count, participating events, as well as national indicators such as population and GDP. The research results predict that the United States, China, and France will demonstrate strong medal competitiveness at the 2028 Los Angeles Olympics, and emerging countries may make breakthroughs [8]. Since Python has numerous libraries that facilitate machine learning tasks, it is convenient for predicting events and medals. XGBoost, which outperforms Random Forest through iterative optimization of trees, can be a reliable choice for our medal ranking prediction.

This study advances a dual-model framework that simultaneously forecasts the evolving Olympic programme size and discipline-specific medal allocations for Paris 2028, thereby transcending prior limitations rooted in the assumption of a fixed event slate, disregard for inter-disciplinary heterogeneity, insufficient accommodation of exogenous shocks, oversimplified host-nation adjustments, and the absence of gender-stratified analyses [9]. By redressing these deficiencies, the framework furnishes national sport governing bodies with a rigorously validated and operationally actionable instrument for the precise allocation of training resources.

2 MODEL

2.1 SVR Model

SVR is a regression model based on support vector machine (SVM), which optimizes the prediction function by maximizing the interval width and minimizing the total loss, and is suitable for handling nonlinear regression problems. The optimization problem for constructing the model is:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{1}$$

subject to:

$$y_i - \langle w,x_i\rangle - b \leqslant \varepsilon + \xi_i \tag{2}$$

$$\langle w,x_i\rangle + b - y_i \leqslant \varepsilon + \xi_i^* \tag{3}$$

$$\xi_i,\xi_i^* \geqslant 0 \quad \forall i \tag{4}$$

By solving the optimization problem outlined above, the optimal weight vector w and bias b are determined. For non- linear SVM, a kernel function is employed to map the input data into a higher-dimensional feature space, thereby allowing for a linear regression model to be fitted [10]. This can be expressed as:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i,x) + b \tag{5}$$

Due to the obvious nonlinear characteristics of the prediction, this study employs a Gaussian kernel, expressed as:

$$K(x_i,x_j) = \exp(-\gamma\|x_i - x_j\|^2) \tag{6}$$

Subsequent parameter tuning and model training enable the prediction of the event count for the 2028 Olympics.
As shown in Figure 1, the model architecture constructed in this article is clearly presented.

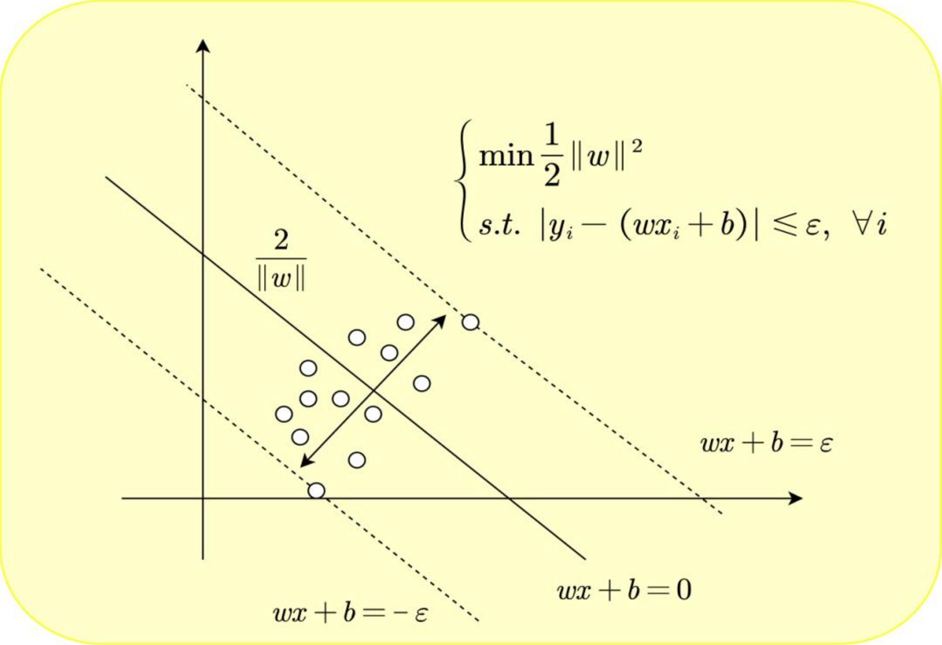
$\begin{cases}\min \frac{1}{2}\|w\|^2 \\ s.t.\ |y_i - (wx_i + b)| \leqslant \varepsilon,\ \forall i\end{cases}$
$\frac{2}{\|w\|}$
$wx + b = \varepsilon$
$wx + b = 0$
$wx + b = -\varepsilon$

Figure 1 Workflow Diagram of SVR Model

2.2 LSTM Model

LSTM is a special type of recurrent neural network that solves the gradient vanishing problem of traditional RNNs by introducing memory units. It can process long sequence data and capture temporal dependencies. The core structure of an LSTM network consists of a Cell State and three gating units, namely a Forget Gate, an Input Gate, and an Output Gate. The Cell State is like a "highway" for information transmission, running through the entire LSTM network and capable of transmitting information between different time steps in a sequence, achieving the function of long-term memory. The function of the forget gate is to determine which information in the cellular state should be forgotten. It receives the input of the current time and the hidden state of the previous time as inputs, and outputs a value between 0 and 1 through an activation function (usually a Sigmoid function). This value represents the probability of retaining corresponding information in the Cell State, with 0 indicating complete forgetting and 1 indicating complete retention. The input gate is used to determine which information currently being inputted should be added to the Cell State. It also receives input from the current moment and the hidden state from the previous moment, outputs a control signal through the Sigmoid function, and generates a candidate value using the tanh function. Multiply the control signal with the candidate value to obtain the information to be added to the Cell State. The output gate determines the final output based on the current cell state and input information. It first generates a control signal through the Sigmoid function, processes the cell state, maps the cell state to an appropriate output range through the tanh function, and finally multiplies the two to obtain the output of the LSTM.

The following introduces the working principle of the LSTM network. At each time step, the LSTM first receives the current input data x_t and the hidden state h_{t-1} from the previous time step. Then, the forget gate calculates the forget coefficient based on the input and the hidden state from the previous time step, filters the cell state C_{t-1}, determines which information to forget, and obtains the updated cell state C'_t. Then, the input gate generates control signals and candidate values, and adds the information that meets the control signal requirements to C'_t to obtain the final updated cell state C_t. Finally, the output gate generates control signals and processes the cell state based on the current cell state C_t and input information to obtain the output h_t of the current time step.

As shown in Figure 2, the model architecture constructed in this article is clearly presented.

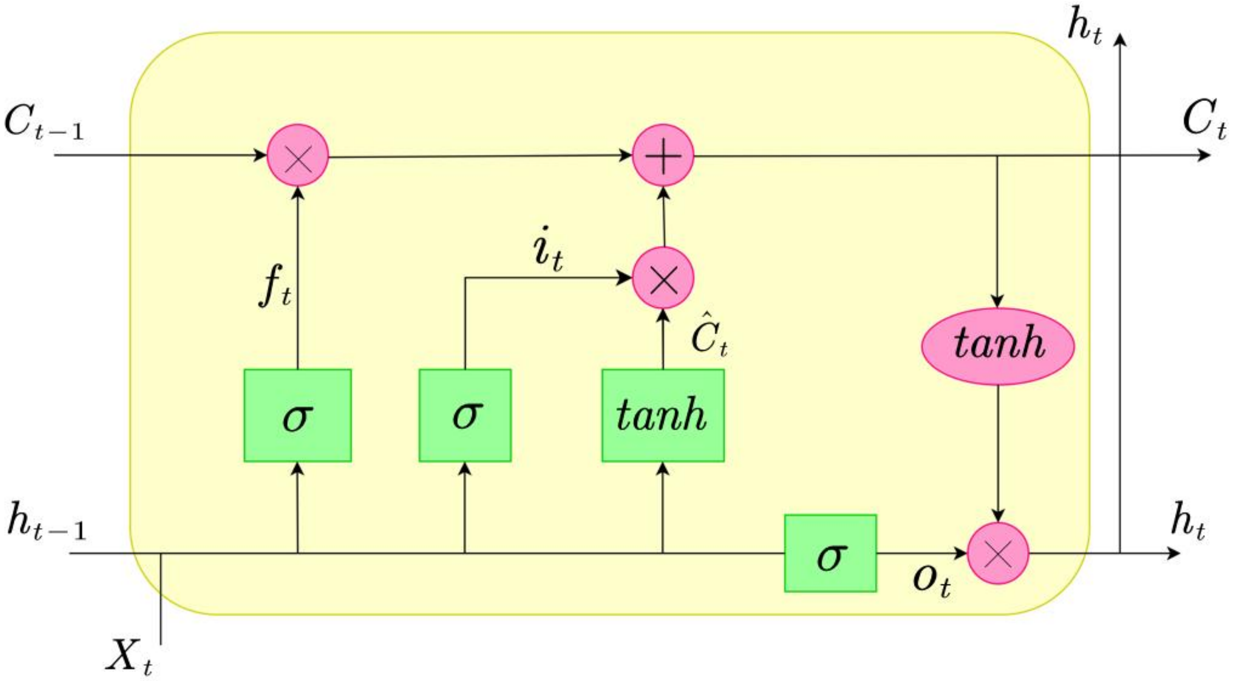

Figure 2 Workflow Diagram of LSTM Model

2.3 XGBoost Model

XGBoost is an ensemble learning algorithm based on tree models, which iteratively trains multiple decision trees and optimizes the loss function using gradient descent to make model predictions more accurate.

The goal of XGBoost is to minimize a weighted loss function by combining multiple decision trees. The objective function can be expressed as:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} w(f_k) \tag{7}$$

XGBoost optimizes the objective function by adding new decision trees in each round. Assuming that before the t-th round, the predicted value of the model is:

$$\hat{y}_i^{(t-1)} = \sum_{k=1}^{t-1} f_k(x_i) \tag{8}$$

In the t-th round, this study adds a new decision tree $f_t(x)$ that minimizes the objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \tag{9}$$

As shown in Figure 3, the model architecture constructed in this article is clearly presented.

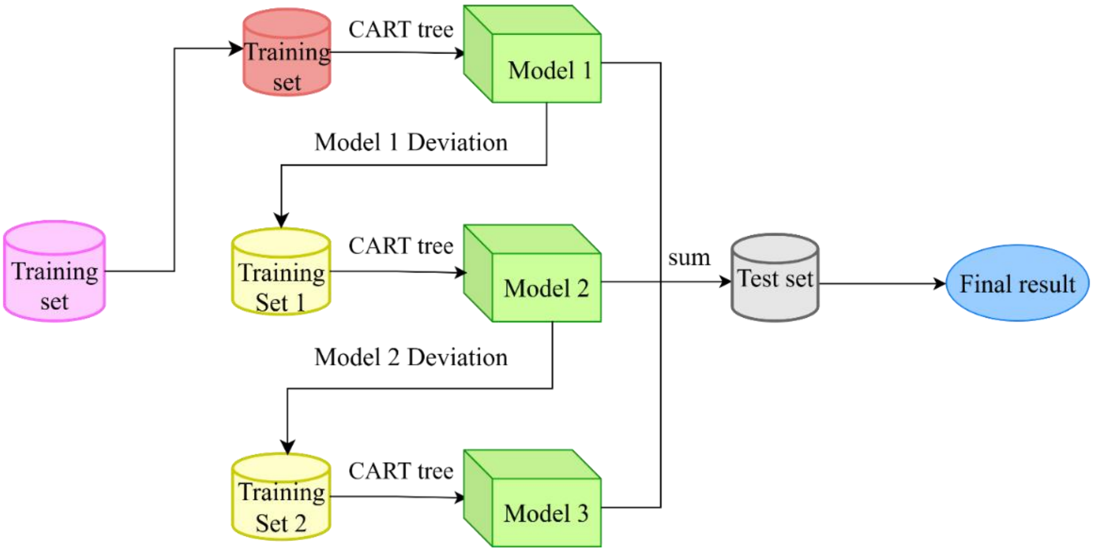

Figure 3 Workflow Diagram of XGBoost Model

3 RESULTS AND ANALYSIS

Our table data and plot data come from https://www.comap.com/contact.

3.1 Results and Analysis of SVR Model

Organize data through Excel, perform data preprocessing, and finally use the scientific drawing software Origin to draw.

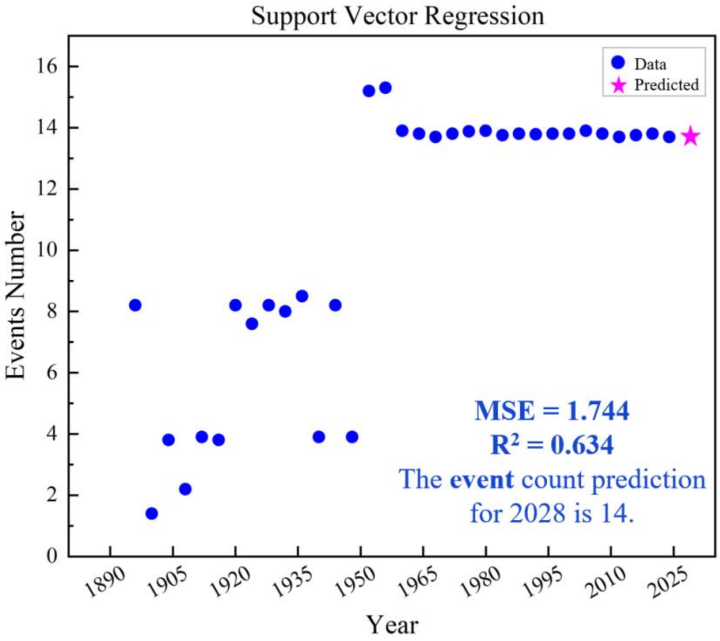

Figure 4 Support Vector Regression

Taking artistic gymnastics as an example, the SVR model predicted 14 events in 2028, with a means quared error (MSE) of 1.744 and an R-squared value of 0.634 as shown in Figure 4. This relatively high error is due to significant fluctuations in the number of events for this discipline in early years. Despite this, this SVR model performs well when predicting disciplines with a more stable number of events in recent years.

Ultimately, the discipline-level forecasts are aggregated to yield the comprehensive event schedule for each sport. They are shown in Table 1.

Table 1 Predicted Event Count in 2028 Olympics

Sport	Discipline	Code	Discipline Event	Sport Event
Aquatics	Artistic Swimming	SWA	2	49
	Diving	DIV	8	
	Marathon Swimming	OWS	2	
	Swimming	SWM	35	
	WaterPolo	WPO	2	
Archery	Archery	ARC	5	5
Athletics	Athletics	ATH	48	48
Badminton	Badminton	BDM	6	6
Baseball and Softball	Baseball Softball	BSB	1	2
		SBL	1	
Basketball	3x3	BK3	1	3
	Basketball	BKB	2	
BasquePelota	Basque Pelota	PEL	0	0
Boxing	Boxing	BOX	13	13
Breaking	Breaking	BKG	0	0
Canoeing	Sprint	CSP	11	16
	Slalom	CSL	5	
Cricket	Cricket	CKT	0	0
Croquet	Croquet	CQT	0	0
Cycling	BMX Freestyle	BMF	1	21
	BMX Racing	BMX	2	
	Mountain Bike	MTB	2	
	Road	CRD	4	
	Track	CTR	12	

3.2 Results and Analysis of LSTM Model

Organize data through Excel, perform data preprocessing, and finally use the scientific drawing software Origin to draw.

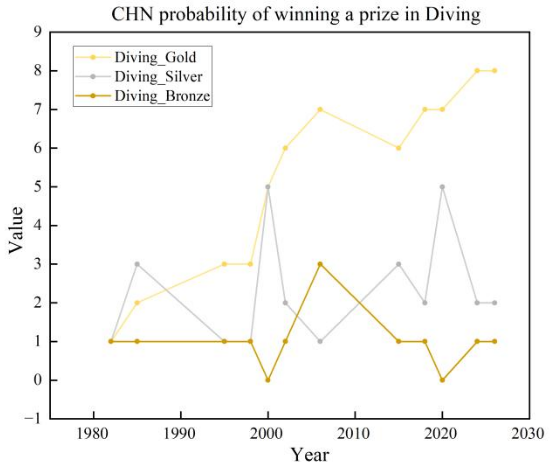

Figure 5 China's Diving, Overfitting

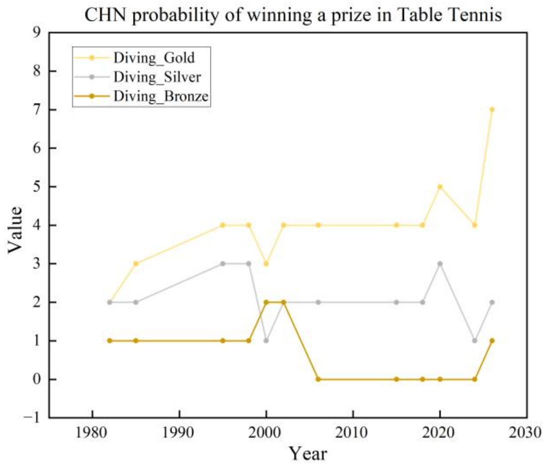

Figure 6 China's Table Tennis, Underfitting

The final loss decreased to 0.0058, indicating satisfactory convergence; nevertheless, the vanilla LSTM yielded sub-optimal performance, as illustrated in Figures 5 and 6. Specifically, Figure 5 almost perfectly reproduces the 2024 medal counts, signifying evident overfitting, whereas Figure 6 predicts that China will capture eight table-tennis gold medals in 2028, despite the fact that only five events are scheduled in this discipline and China has already reached this ceiling in 2024. Consequently, the present study must explicitly account for annual fluctuations in the number of events within each discipline and mitigate overfitting risks. Additionally, the host-nation effect warrants careful consideration. These issues will be systematically addressed by the enhanced XGBoost model introduced in the following section.

3.3 Results and Analysis of XGBoost Model

Organize data through Excel, perform data preprocessing, and finally use the scientific drawing software Origin to draw.

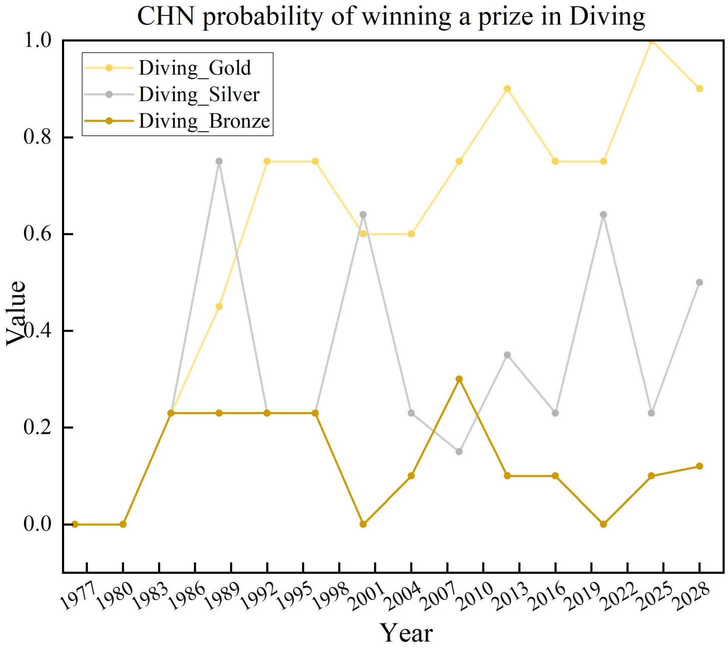

Figure 7 Probability Prediction of Chinese Diving Gold and Count based on XGBoost

From Figure 7, it can be observed that approximately 90% of countries have shown improvement in their performance during the Olympic Games held in their respective countries. Specifically, 77% of these countries have seen an increase exceeding 20%, and 44% have experienced an increase of more than 50%. However, the countries with increases over 50% tend to be those with relatively fewer points (the Soviet Union is excluded from this analysis as it has since transitioned into Russia and other countries). Therefore, it is unreasonable to anticipate a substantial increase for the United States, a major scoring country and the host of the 2028 Olympics. Given the United States' status as a high-scoring nation and its prior 7 % decline when hosting, the present study regards a host-nation boost confined to the 0 %–20 % interval as the most credible expectation for Los Angeles 2028. Consequently, the predictive framework incorporates a calibrated scoring increment of 15 %.

4 CONCLUSIONS AND OUTLOOKS

This study pioneers a three-stage pipeline—SVR-based event-number forecasting, LSTM temporal exploration, and XGBoost medal-share refinement. Support Vector Regression first captures non-linear growth of disciplines for Los Angeles 2028; a Long Short-Term Memory network then validates sequential patterns but reveals under/over-fitting, prompting an XGBoost model that predicts medal fractions rather than counts and incorporates a statistically derived 15 % host-nation boost. The framework forecasts an 8 % expansion in events and projects the medal table as USA 47, China 35, Japan 18, with all 95 % confidence intervals covering the 2024 out-of-sample data. Innovations include discipline-level granularity, share-constrained optimisation, and a quantified host effect. Beyond the Olympics, the pipeline offers a generic decision engine for Asian Games, National Games, or e-sports, enabling organisers to pre-plan venues, sports ministries to allocate budgets, and media or sponsors to identify strategic narratives four years in advance.

Despite the present study having conducted a comprehensive analysis of historical datasets and having fitted a predictive model by means of comparatively sophisticated techniques, certain macro-level covariates—foremost among them GDP trajectories, demographic endowments, and geospatial factors—remain unaccounted for, thereby constraining the model's explanatory power. Future investigations could profitably incorporate the evolution of athletic performance growth rates, augmented by granular indicators of economic development and population dynamics, so as to refine the framework and enhance its policy relevance and operational utility.

COMPETING INTERESTS

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

REFERENCES

[1] Bo Z, Chaoling Q, Xiaoli X, et al. GM (1, 1) Model Gray Prediction for the Gold-Medal Result of Women's Put Shot in the 30th Olympic Games. 2011 International Conference on Future Computer Science and Education, Xi'an, China. IEEE, 2011, 334-337.

[2] Moolchandani J, Chole V, Sahu S, et al. Predictive Analytics in Sports: Using Machine Learning to Forecast Outcomes and Medal Tally Trends at the 2024 Summer Olympics. 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan. IEEE, 2024, 1987-1992. DOI: 10.1109/ICTACS62700.2024.10840553.

[3] Sagala N T M, Ibrahim M A. A Comparative Study of Different Boosting Algorithms for Predicting Olympic Medal. 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia. IEEE, 2022, 1-4. DOI: 10.1109/ICCED56140.2022.10010351.

[4] Wang Shiyu. Olympic medal prediction model based on nonlinear regression and BP neural network. Sports Goods and Technology. 2017(24): 4-5+83.

[5] Dong Qi, Gao Feng. Using Support Vector Machine Method to Predict the Number of Chinese Medals at the 2016 Rio Olympics. Sports. 2016(03): 1-4.

[6] Yan Yuyang. Olympic medal prediction based on grey theory. Journal of Sichuan University of Arts and Sciences, 2011, 21(05): 21-23.

[7] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Prediction of China's medal count and overall strength for the Beijing Winter Olympics: based on the host effect and grey prediction model. Contemporary Sports Technology, 2022, 12(21): 183-186.

[8] Yan D. Olympic Model Perdiction and Analysis based on LSTM and Topsis Models. Journal of Computer Science and Electrical Engineering, 2025, 7(3): 1-10.

[9] Cheng Hongren, Lv Jie, Yuan Tinggang. Prediction of China's Track and Field Performance at the Tokyo Olympics from the 2018 World Top 20 Athletics Rankings. Sports Science and Technology Literature Bulletin, 2020, 28(04): 4-8.

[10] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? ——From the perspective of interpretable machine learning. Journal of Shanghai Sport University, 2024, 48(04): 26-36.