Olympic Medal Prediction Based on Multi-Task Hybrid Modelling

Kaibing Yang[1,a,*,#], Jiahao Zhang[1,b,#], Yuefeng Chen[1,c,#]

[1]Future Technology Institute, Tianjin University, Tianjin, China

[a]kaibing_yang@tju.edu.cn, [b]jiahaozhang@tju.edu.cn, [c]2427165301@qq.com

*Corresponding author

[#]These Authors Contributed Equally to This Work

Keywords: Olympic Medal Prediction, Hybrid Model, Elastic Net Regression, XGBoost, Random Forest

Abstract: With increasing global investment in elite sports, accurate forecasting of Olympic medal outcomes has become a critical area of research. This study develops a comprehensive hybrid prediction framework for estimating medal counts at the 2028 Los Angeles Olympics. Leveraging a combination of statistical techniques and machine learning models, the framework addresses four key tasks: overall medal count prediction, identification of the first medal-winning country, assessment of event-specific contributions to medal totals, and evaluation of elite coaching impacts. For medal count forecasting, we construct a stacked ensemble model integrating Elastic Net Regression (ENR), XGBoost, LightGBM, and CatBoost, with clustering and multi-criteria decision analysis enhancing feature representation. The ensemble achieves a mean squared error of 1 and an R^2 of 0.963, projecting the U.S. to lead with 45 gold and 132 total medals. A two-stage random forest model is employed to predict the first medal-winning country, suggesting Luxembourg as a top contender. Gray relational analysis reveals strong positive correlations between the number of events, participating nations, and medal counts, while synthetic control methods confirm the significant impact of top-tier coaching on national performance. This integrated approach not only improves predictive accuracy but also offers actionable insights for national Olympic committees in optimizing resource allocation and strategic planning. The study underscores the importance of combining data-driven modeling with domain-specific knowledge for complex, high-stakes forecasting tasks.

1. Introduction

The Olympic Games have long symbolized international athletic excellence and national strength. In recent years, the growing global investment in elite sports programs has intensified competition not only in athletic performance but also in medal standings, which have become critical indicators of a country's comprehensive sports development [1]. Public attention and policymaking are increasingly influenced by projected medal outcomes, prompting the need for accurate and interpretable forecasting models [2].

Traditional Olympic medal prediction methods often rely on static variables such as historical medal counts or GDP per capita. However, these approaches frequently neglect dynamic and nonlinear factors like athlete mobility, event composition, host country advantages, and coaching influence. Moreover, predictions are typically released close to the Games when rosters are finalized, leaving limited room for strategic planning [3].

To address these limitations, this study proposes a comprehensive hybrid framework to predict Olympic medal outcomes for the upcoming 2028 Los Angeles Games. The framework integrates advanced machine learning techniques, data-driven clustering, multi-criteria evaluation, and scenario-based decision models to analyze and forecast medal distribution. Beyond overall medal counts, the model predicts the first medal-winning country, quantifies the relationship between events and medal outcomes, and assesses the influence of elite coaching on national performance.

This paper presents the following key contributions:

- Proposed a Hybrid Multi-Model Framework: We developed a comprehensive forecasting model

  DOI: 10.25236/iiicec.2025.012

that combines ENR, XGBoost, LightGBM, and CatBoost via stacking fusion to improve the accuracy of Olympic medal predictions.

- Introduced Feature Engineering and Athlete Evaluation Techniques: We applied K-means++ clustering for country classification and a Topsis+EWM multi-criteria decision method to evaluate athlete-level features, refining critical inputs for prediction.
- Designed a Two-Stage Random Forest for Early Medal Predictions: To forecast the first medal-winning countries, we constructed a classifier-regressor pipeline that filters high-potential nations and estimates their likelihood of early medal success.
- Integrated Event-Medal Correlation Modeling: Using gray correlation analysis (GRA), we identified which sports contribute most to medal gains across nations and how host countries can leverage event selection strategically.

2. Methodology

2.1. Medal count prediction model

Considering that the number of medals won by each country in each Olympic Games is not purely a linear or nonlinear change, in order to make the prediction result of the final model more reasonable and accurate, we adopt the hybrid model obtained by stacking and mixing the elasticity network regression, XGBoost, CatBoost and LightGBM models for the prediction of the number of medals [4]. The following will begin with a brief description of these models: Elastic Net Regression (ENR) is a linear regression model that combines Lasso regression and Ridge Regression to overcome the limitations of single regularization methods by using both L1 and L2 regularization terms. The core of the elastic regression network is its loss function, which combines the L1 and L2 regularization terms in the following form:

$$\text{minimize}\{\frac{1}{2n} \parallel y - Xw \parallel_2^2 + \alpha \cdot \rho \parallel w \parallel_1 + \frac{\alpha \cdot (1-\rho)}{2} \parallel w \parallel_2^2\} \tag{1}$$

Where Y is the target variable; X is the feature matrix; w is the model coefficients; α is the regularization strength, which controls the overall weight of the regularization term; and ρ is a mixing parameter with a value in the range of [0,1], which is used to control the relative weights of the L1 and L2 regularizations. When $\rho = 0$, the model degenerates to ridge regression. When $\rho = 1$, the model degenerates to Lasso regression. The core idea of XGBoost is to optimize the model step by step through an iterative process, where a new weak learner (usually a decision tree) is added at each iteration to correct the prediction error of the previous round of the model. Its main steps include:

1) Objective function

$$\mathcal{L} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{T} \Omega(f_k) \tag{2}$$

Where L is the loss function, $\Omega(\mathrm{f_k}) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{\mathrm{j}=1}^{\mathrm{T}} \mathrm{w_j^2}$ is the regularization term, T is the number of leaf nodes of the tree, w_j is the weight of the leaf nodes, and γ and λ are the regularization parameters.

2) Second-order Taylor expansion

In the tth iteration, the objective function can be approximated as:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[L(y_i, \hat{y}_i^{(t-1)}) + g_i h_t(x_i) + \frac{1}{2} h_t(x_i)^2 h_i \right] + \Omega(h_t) \tag{3}$$

Where $\mathrm{g_i} = \frac{\partial \mathrm{L}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t}-1)})}{\partial \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t}-1)}}$ and $\mathrm{h_i} = \frac{\partial^2 \mathrm{L}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t}-1)})}{\partial (\hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t}-1)})^2}$ are the first and second order derivatives of the loss function, respectively.

3) Leaf node weights

For each leaf node, the optimal weight w_j^* can be obtained by derivation:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{4}$$

Where I_j is the set of samples falling in the jth leaf node.

4) Optimized objective function

After substituting the optimal weights, the objective function can be simplified as:

$$\mathcal{L}^{(t)} = -\frac{1}{2}\sum_{j=1}^{T} \frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i+\lambda} + \gamma T \tag{5}$$

By maximizing this objective function, the optimal tree structure can be found. The following table shows the pseudo code list of the XGBoost algorithm.

CatBoost (Categorical Boosting) is a machine learning algorithm based on Gradient Boosting Decision Tree (GBDT) developed by Yandex, which is good at dealing with datasets containing categorical features.The core idea of CatBoost is to construct a model through the Gradient Boosting framework, adding a new decision tree each iteration to gradually optimize the objective function.

1) Loss Functions and Regularization

CatBoost's loss function contains two parts: the training error and the regularization term:

$$\mathcal{L}(F) = \sum_{i=1}^{n} L(y_i, F(x_i)) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

2) Target statistical code

$$\text{avg}_{\text{t}}\text{arget} = \frac{\text{countInClass}+\text{prior}}{\text{totalCount}+1} \tag{7}$$

Note that M_i is trained without using the example X_i. CatBoost implementation uses the following relaxation of this idea: all M_i share the same tree structures. The core idea of LightGBM is to construct the model by means of gradient boosted decision trees (GBDT), adding a new decision tree at each iteration to gradually optimize the objective function.

1) Objective fuction

The objective function of LightGBM also consists of a loss function and a regularization term:

$$\mathcal{L} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{T} \Omega(f_k) \tag{8}$$

2) Histogram Algorithms

In the histogram algorithm, the feature values are divided into k buckets and the gradient and Hessian values are calculated separately for each bucket:

$$G_k = \sum_{i\in \text{bin}_k} g_i, H_k = \sum_{i\in \text{bin}_k} h_i \tag{9}$$

Where g_i and h_i are the first and second order derivatives of the loss function, respectively.

3) Split Gain Calculation

For each split point of each feature, the split gain can be calculated by the following equation:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma \tag{10}$$

Where G_L and G_R are the gradient sums of the split left and right subtrees, respectively, H_L and H_R are the Hessian sums of the split left and right subtrees, respectively, and is the regularization parameter.

4) Leaf Node Weights

For each leaf node, the optimal weight w_j^* can be obtained by derivation:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i+\lambda} \tag{11}$$

Where I_j is the set of samples falling in the jth leaf node.

After predicting the medal table with each model above, we will select the most accurate one as the meta-model of the hybrid model, and then stack the other models on top of it to get a more accurate hybrid model for prediction. The simple structure of our hybrid stacking model is shown as Figure 1.

Figure 1 Schematic diagram of the stacking modeling process.

Firstly, we need to define the prediction target of the model and the individual characteristics for the model to predict. The number of gold medals, silver medals, bronze medals and the total number of medals of each country are our prediction targets. As for the reference features, after analysing, selecting and pre-processing the data, we decided to use the following features as our reference features: the national sports intensity level, the total number of gold medals of the country in the past three years, the total number of medals of the country in the past three years, the number of registered sports of the country, the number of registered sports of the country, the number of registered athletes of the country and the number of athletes of each level, and lastly whether or not the country is a competing country [5].

Secondly, after determining each data, we first predicted each prediction model individually and identified the model with the highest accuracy as the meta-model for our hybrid stacked model.

Elastic Net Regression:

We return to equation X mentioned above:

$$\text{minimize}\{\frac{1}{2n} \parallel y - Xw \parallel_2^2 + \alpha \cdot \rho \parallel w \parallel_1 + \frac{\alpha \cdot (1-\rho)}{2} \parallel w \parallel_2^2\} \tag{12}$$

In order to accurately determine the regularisation strength α as well as the mixing parameter ρ, the Grid Search Cross-Validation (GSCV) technique is introduced in this study. This technique is a widely used method for hyperparameter optimisation to determine the optimal combination of hyperparameters by performing an exhaustive search on a predefined parameter grid and combining it with cross-validation to evaluate the model performance.

We set the range of α to 0 to 0.2, and the range of A to 0 to 1. The step size of both is set to 0.01, under which an exhaustive search is performed and for each set of hyper-parameter combinations, cross-validation is used to evaluate the performance of the model, and then the optimal hyper-parameter combinations are selected based on the average performance metrics of the cross-validation (e.g., accuracy, mean square error, etc.). Here we take the mean square error (MSE) as an example, which is calculated by the formula:

$$MSE(t) = \frac{1}{N_t} \sum_{i \in D_t} (y_i - \bar{y}_t)^2 \tag{13}$$

Xgboost:

Here we will not repeat the overall process of XGBoost directly see the final optimised objective function:

$$\mathcal{L}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{14}$$

Using iterative optimisation after bringing in the prepared reference features using Matlab, we

similarly obtained their corresponding prediction maps (see later).

2.2. Two-stage random forest algorithm

Two-Staged Random Forest is an improved random forest algorithm, mainly used to solve the deficiencies of traditional random forest in feature selection and model optimization [6].

Phase I:

In this phase, the raw features are first screened to remove redundant and unimportant features. This can be done by calculating the importance of the features (e.g., Gini index gain, mean square error, etc.). Then, a set of decision trees are trained based on the filtered subset of features. The goal of this stage is to reduce the feature dimensions and improve the training efficiency and generalization ability of the model.

We choose the Gini index gain to filter the features of the training set. For a feature A and division point s, the Gini index gain can be expressed as:

$$\Delta Gini = Gini(D) - \sum_{v\in\{left,right\}} \frac{|D_v|}{|D|} Gini(D_v) \tag{15}$$

Where D is the dataset of the current node, and D_v is the sub-dataset after division according to the division point s of feature A.

After filtering the features of the training set using this formula, we finally choose the number of national participating athletes, the number of sports involved, the number of sports, and the number of level 3 athletes as the features of the training set.

Then we select the above features from the dataset of those countries that have won only one medal and let the Random Forest classifier learn the features of the data of such breakthrough countries that have achieved 0 medals and predict the output accordingly, with an output of 0 or 1. 0 means that the country will not win a medal in the 2028 Olympics and 1 means that the country will win a medal.

Phase II:

In Phase I we determined which countries would achieve a medal 0 breakthrough, and in Phase II we needed to further refine our prediction model. In order to predict the specific number of medals for the countries that will win medals, we are still using Random Forest as a regression model to predict the number of medals [7]. In this phase we chose to use the same features as in the first phase of the algorithm, but only for those countries that were predicted to win medals in the first phase. The output is the number of medals that each country will win.

3. Results

For the prediction results (Figure 2) of the remaining two models (CatBoost, LightGBM), they are not repeated here due to space issues (see later for visualisation images).

After the comparison of the individual linear regressions (Figure 3), the Elastic Net Regression model with the better MSE and $R2$ was selected to be used as the meta model for the hybrid model.

We put the predictions of other models except the meta-model into the meta-model Elastic Net Regression. Using the principle of stacking, we let the meta-model Elastic Net Regression learn the prediction results of other models and combine them reasonably, and finally we get the hybrid stacked model we need, which combines the advantages of all the models and avoids the disadvantages, so that the final results are more reasonable and more accurate.

By comparing the individual experimental results in Figure 2, it can be found that the stacking model significantly improves the experimental accuracy.

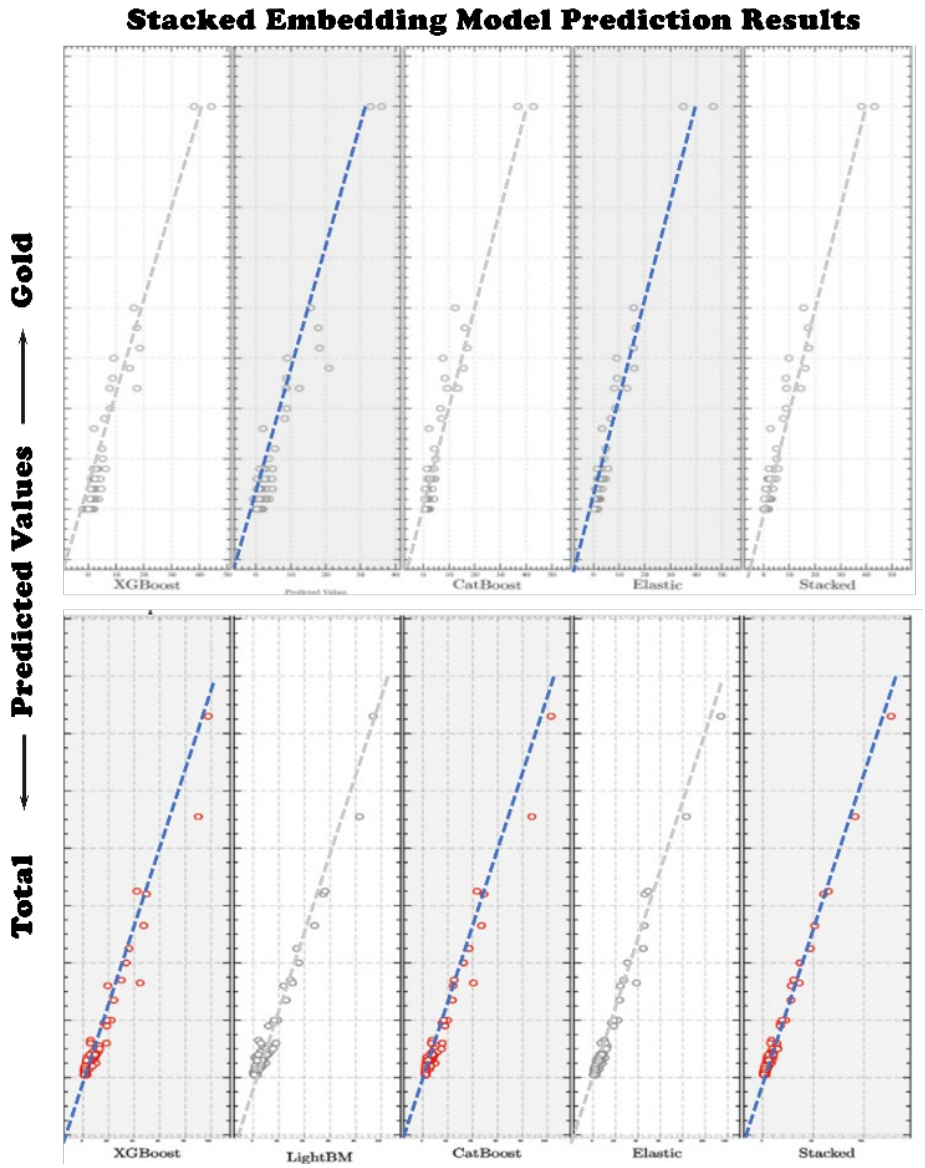
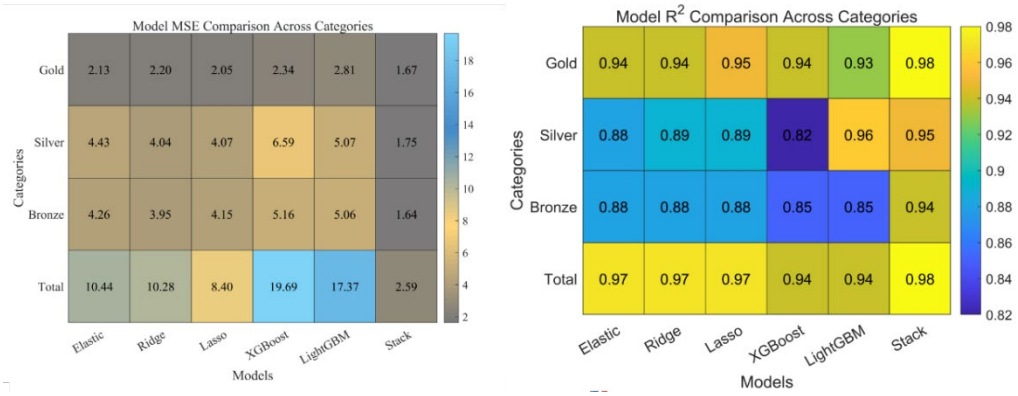


Figure 2 Detailed results of the stack model.

Since whether a country has a medal or not is a non-zero or one thing, it has only two cases and we need a binary classifier, so the two-stage random forest algorithm is the most appropriate algorithm to accomplish the prediction. Specific evaluation indicators of the stack model are in Figure 3.



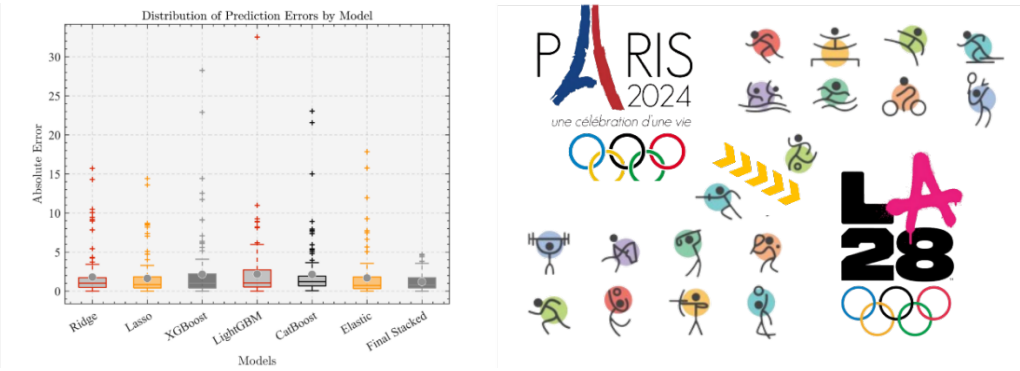


Figure 3 Specific evaluation indicators of the stack model.

After the two-stage random forest prediction, the results can be obtained as shown in Figure 4(c), while Fig. 4 demonstrates the training process and SHAP correlation degree map.

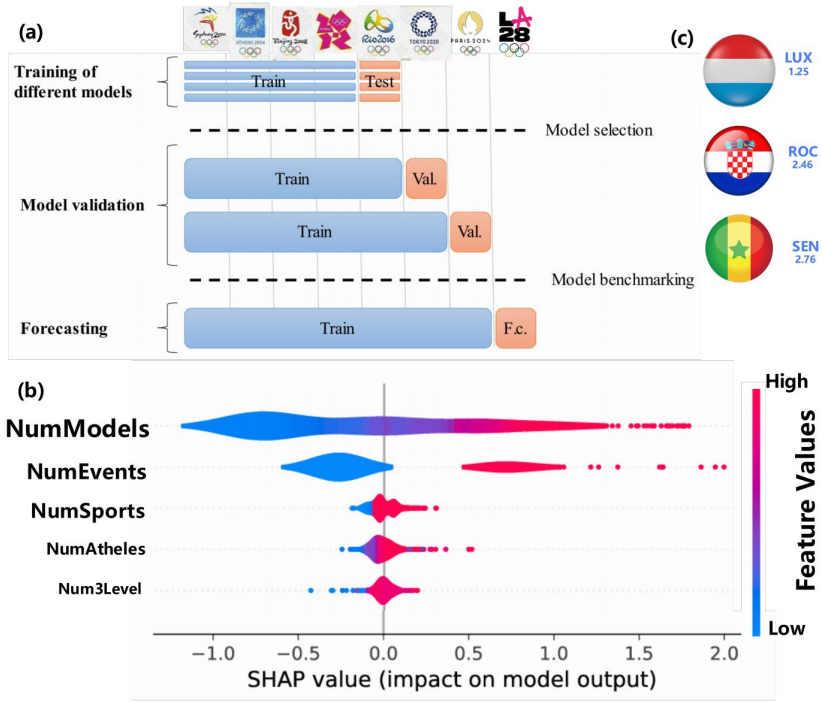


Figure 4 Two-random forest.

Based on market odds experience, the following odds calculation formula can be obtained:

$$odds = \frac{p}{1-p} \tag{16}$$

Thus the relevant results can be obtained as shown in Table 1.

Table 1 Odds by Country Code.

	Country		
	LUX	ROC	SEN
Odds	1.25	2.46	2.76

4. Conclusions and Future Work

This study develops a comprehensive and data-driven framework for predicting Olympic medal distributions, with a particular focus on the 2028 Los Angeles Olympic Games. By integrating linear regression models (Elastic Net Regression) and nonlinear ensemble learning algorithms (XGBoost, LightGBM, CatBoost) through a stacked modeling strategy, the proposed system achieves high accuracy and strong generalization ability [8]. The hybrid framework not only captures complex

patterns in historical Olympic data but also leverages domain-specific insights through techniques, and the synthetic control method for evaluating coaching influence [9].

The results indicate that medal outcomes are strongly influenced by factors such as athlete number, event participation scope, coaching expertise, and host-country advantages. Particularly, countries that strategically allocate coaching resources and prioritize high-yield events tend to gain a competitive edge. The two-stage random forest model further enhances the interpretability and accuracy in forecasting early medal distribution, while the sensitivity analysis confirms the robustness of key variables across different scenarios.

Looking ahead, there remains significant potential to enhance the predictive power and applicability of the model. Incorporating real-time dynamic data such as athlete injuries or qualification updates could make predictions more responsive to external changes. Additionally, integrating temporal learning frameworks like recurrent neural networks or transformer-based models may better capture inter-Olympic trends. Expanding the framework to other multi-sport events, such as the Asian Games or Commonwealth Games, would help validate its generalization capability. Finally, combining algorithmic modeling with expert feedback from coaches or national committees could open new avenues for decision-making support in elite sports management [10].

Overall, this work not only provides a practical medal forecasting tool but also contributes to a deeper understanding of the multifactorial dynamics behind Olympic success, offering valuable implications for national sports strategy and Olympic planning.

References

[1] Leroy A. Multi-task learning models for functional data and application to the prediction of sports performances[D]. Université Paris Cité, 2020.

[2] Zhao S, Cao J, Steve J. Research on Olympic medal prediction based on GA-BP and logistic regression model[J]. F1000Research, 2025, 14: 245.

[3] Schlembach C, Schmidt S L, Schreyer D, et al. Forecasting the Olympic medal distribution–a socioeconomic machine learning model[J]. Technological Forecasting and Social Change, 2022, 175: 121314.

[4] Raja M, Sharmila P, Vijaya P, et al. Olympic Games Analysis and Visualization for Medal Prediction[C]//2025 International Conference on Artificial Intelligence and Data Engineering (AIDE).IEEE, 2025: 822-827.

[5] Sayeed R, Hassan M T, Rahman M N, et al. Machine Learning Models for Predicting Olympic Medal Outcomes[C]//2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI). IEEE, 2025, 3: 1-6.

[6] Bian X. Predicting Olympic medal counts: The effects of economic development on Olympic performance[J]. The park place economist, 2005, 13(1): 37-44.

[7] Schlembach C, Schmidt S L, Schreyer D, et al. Forecasting the Olympic medal distribution during a pandemic: a socio-economic machine learning model[J]. arXiv preprint arXiv:2012.04378, 2020.

[8] Wang Y, Wang J, Huang T Y, et al. STGCN-LSTM for Olympic Medal Prediction: Dynamic Power Modeling and Causal Policy Optimization[J]. arXiv preprint arXiv:2501.17711, 2025.

[9] Thirumalai C, Monica S, Vijayalakshmi A. Heuristics prediction of olympic medals using machine learning[C]//2017 International conference of Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2017, 2: 594-597.

[10] Yang Y, Feng W. Research on Olympic Medal Prediction Based on the LSTM-GWO-DeepForest Model[C]//2025 International Conference on Digital Analysis and Processing, Intelligent Computation (DAPIC). IEEE, 2025: 844-849.