# Research on sports event performance prediction based on multi-model fusion and feature engineering

Yutong Liu *, #, Wuyang Li #

Hohai University, Nanjing, China

* Corresponding Author Email: 18936012662@163.com

#These authors contributed equally.

**Abstract.** This study develops a multi-objective prediction model to solve complex prediction tasks in hierarchical data structures. The first is the random forest model, which improves the accuracy and stability of the model by constructing multiple decision trees and combining their predictions while solving the nonlinear dependency and convergence problems. The random forest model efficiently models complex relationships through global optimization of initial weights and biases. The second approach is an XGBoost model that utilizes advanced feature construction techniques focusing on improved feature tuning and regularization techniques to achieve a balance between accurate error correction and complex pattern capture. The framework emphasizes the importance of feature engineering, integrating objective and subjective feature weighting to improve the accuracy of multivariate datasets. By fusing machine learning methods with statistical paradigms, this integrated model improves predictive performance and provides actionable insights for complex and diverse use cases.

**Keywords:** Multi-model Fusion, Feature Engineering, Random Forest Model, XGBoost Model.

## 1. Introduction

Prediction of multivariate outcomes in complex systems is a challenging task due to nonlinear relationships in high-dimensional feature spaces, significant multicollinearity, and heterogeneity of data structures. Constructing accurate and robust prediction models requires a systematic analysis and solution of these problems. Traditional statistical methods, such as grey prediction models [1] and linear regression [2], have been widely used in fields such as economic indicator forecasting and time series analysis. Meanwhile, machine learning methods, such as feed-forward neural networks [3], provide effective alternatives for dealing with nonlinear relationships and time-dependent data. However, these methods usually show obvious limitations when facing high-dimensional features, complex data structures, and generalization capability requirements. Specifically, previous models often lack a systematic feature selection mechanism, are prone to fall into local optimal solutions, and exhibit poor generalizability when dealing with contexts with different data distributions and contextual heterogeneity.

To overcome the above challenges, this study proposes a multi-objective prediction model integrating Random Forest and XGBoost. The model significantly improves the prediction accuracy and robustness of complex multivariate systems by integrating advanced machine learning techniques and statistical optimization methods. Specifically, the model adopts a two-stage architecture that aims to provide a comprehensive solution to the multivariate prediction problem.

The first component of the framework is the Random Forest model, which is specifically designed to deal with the prediction of continuous-type target variables. Random Forest is able to effectively capture complex non-linear relationships in data by globally optimizing the initial weights and biases. In addition, Random Forest possesses excellent immunity to interference, and its properties based on Bagging integration and stochastic subspace methods enable it to perform well when dealing with high-dimensional data. Compared with a single decision tree, Random Forest introduces a random feature selection mechanism, which increases the diversity of the model, thus improving the accuracy and robustness of prediction and effectively reducing the risk of overfitting. By adopting the OOB

evaluation method, the model exhibits a low error rate and strong generalization ability in practical applications.

The second component is the XGBoost model, which further enhances the robustness and generalization of the model through fine-grained feature construction and regularization techniques, including L1 and L2 penalties. XGBoost employs a second-order gradient optimization algorithm and a tree-structured penalization mechanism, which in turn improves the predictive accuracy of the model while preventing overfitting [4]. In addition, the structural flexibility of XGBoost enables it to maintain high interpretability and consistent prediction performance when dealing with complex datasets with spatial and temporal heterogeneity, which further enhances the effectiveness of the overall model. In summary, through the design of this two-stage architecture, the multi-objective prediction model proposed in this study effectively combines the advantages of Random Forest and XGBoost, providing an efficient and robust solution for multivariate outcome prediction in complex systems.

## 2. Materials and Methods

### 2.1. Materials

#### 2.1.1. Data Extraction

In order to predict the winners of sports events, this study takes into account the following key factors: the type and number of programs, the winners of each country in previous events, and the participation of athletes from each country. Data acquisition utilized two principal open-access repositories: the crowd-sourced encyclopedia Wikipedia's comprehensive listing of Olympic sports (https://en.wikipedia.org/wiki/Olympic_sports) and https://www.comap.com/contests/mecm-icm.

#### 2.1.2. Data Processing

Observing the existence of vacancies and outliers in the extracted data, before the model is built, this study carries out data preprocessing on the extracted data. Efficient data preprocessing can ensure the smooth operation of the model and improve the accuracy of the model prediction.

For the vacant values in the data: Considering that the awards and participation of most countries will not have large sudden changes in a short period of time, we adopt the nearest neighbor method for local interpolation. Skating and Ice Hockey, which are no longer part of the target events since 1924, were deleted to eliminate the systematic missing interference caused by the change of event classification.

Targeting outliers in the data: Removing non-conventional entry entities and historical records of participating countries that no longer exist. When outliers are detected, the average of highly correlated data values is interpolated to fix the outliers in specific data columns.

Data filtering: Considering that the effect of host effect on the number of national medals may interfere with the prediction of the historical trend of medals and the overall strength of the national athletes, and that with the change of the world pattern the correlation between the older data and the actual situation is less and less, we firstly smoothed the historical data with Exponential Moving Average (EMA). process, whose recursive formula is:

$$EMA_t = \alpha \cdot Xt + (1\text{-}\alpha) \cdot EMA_{t-1} \tag{1}$$

The algorithm can give higher weight to the recent data through the exponential decay mechanism, effectively suppressing the short-term abnormal fluctuations while retaining the long-term trend, so as to respond to the changes in the recent data more quickly. At the same time, the noise in the data is smoothed to exclude, to a certain extent, the influence of the host effect on the number of Olympic medals of the host country.

### 2.2. Methods

In order to comprehensively consider the number and types of events, the trend of national awards, and the impact of the host effect on the prediction of the number of medals, this study adopts a fusion

model to classify the data based on the countries that have won awards and those that have not won awards, and build the prediction models separately. For awarded countries, an ensemble model combining Random Forest, statistical regression, and XGBoost is developed. This model incorporates host country effect factors through dynamic weight allocation, enabling comprehensive prediction of medal counts by capturing non-linear relationships between historical performance, sports investment, and athlete population. For non-medal-winning countries, a support vector machine classification model with feature engineering is applied to identify emerging medal potentials such as Nepal and Papua New Guinea (prediction probability 65%-85%). Finally, fature importance analysis using XGBoost reveals swimming, diving, and track and field as the critical medal-contributing events for representative countries.

### 2.2.1. Random Forest (RF) Model

Random Forest is an ensemble method that aggregates predictions from multiple decision trees via bootstrap sampling and random feature selection, aggregating predictions to enhance generalization [5]. Its dual randomness mitigates overfitting while maintaining interpretability through feature importance metrics. Recent advancements focus on high-dimensional optimization, with efficient implementations enabling rapid training on large datasets [6].

For time series forecasting, lagged variables (e.g., past performance metrics) are incorporated as input features to transform the sequential data into a supervised learning problem. The algorithm's ability to inherently handle mixed-type variables (e.g., numerical performance indicators and categorical event attributes) makes it particularly suitable for modeling dynamic competition systems with heterogeneous data sources. Additionally, the feature importance analysis provided by RF helps identify critical temporal factors influencing tournament results, offering both predictive accuracy and interpretability for sports analytics applications.

### 2.2.2. XGBoost Model

XGBoost (eXtreme Gradient Boosting) Extreme Gradient Boosting, is an algorithm based on GBDT. XGBoost is an ensemble decision tree method where new trees are added to correct errors in the existing model until no further improvement is achieved. It enhances the gradient boosting algorithm by using classification and regression trees (CART) as base learners, which are generated sequentially. Initially, a prediction value is assigned to the base learner, and subsequent decision trees optimize this prediction by addressing the residuals of the prior tree [7]. The objective function of XGBoost consists of two parts: a loss function and a regularization term. Knowing the training dataset $\Gamma = \{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$, the loss function $l(y_i,\hat{y}_i)$, and the regularization term $\Omega(f_k)$, the overall objective function can be written as:

$$\mathcal{L}(\phi) = \sum l(y_i,\hat{y}_i) + \sum_k \Omega(f_k) \tag{2}$$

Where $\mathcal{L}(\phi)$ is the expression on linear space; i is the i-th sample, k is the kth tree; $\hat{y}_i$ is the predicted value of the i-th sample $x_i$.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

After the second-order Taylor expansion, remove the constant term and optimize the loss function term; regularization term expansion, remove the constant term and optimize the regularization term; merge the primary term coefficients and quadratic term coefficients to get the final objective function.

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}\left[ G_J w_j + \frac{1}{2}(H_J + \lambda) w_j^2 \right] + \gamma T \tag{4}$$

The variable is only the weight vector of the t-th tree.

### 2.2.3. SVM Model

In machine learning, support vector machine analyzes data and identification criteria for classification and regression analysis [8]. SVM is a supervised binary classification model. The SVM training algorithm is modeled and assigned to one of the classes, making it a nonlinear binary classification. It distinguishes between positive and negative samples by finding the maximum interval classification plane w x + b = 0 wx + b = 0. For the linearly indistinguishable case, the low-dimensional space is mapped to the high-dimensional space by the kernel technique to make it linearly distinguishable. Examples of the SVM model (such as points in space) are assigned to a sample that represents different classes of space in as clear a way as possible. In the support vector machine algorithm, the support vector is the key of the training set, which is closer to the decision point, and the graded surface is determined by the support vector in the sample. The classification decision function can be expressed as [9]:

$$f(x) = sgn(w^{*T}x + b^*) = sgn\left(\sum_{i=1}^{N} a_i^* d_i (x_i^T x) + b^*\right) \tag{5}$$

Platt Scaling, a parametric calibration method proposed by Platt (1999), addresses the problem of transforming discriminative model outputs into well-calibrated posterior probabilities. This technique employs a logistic regression (LR) framework with a sigmoid activation function to map the raw scores of a base classifier to the interval (0,1). Specifically, given the unresolved decision value (f(x)) of sample x produced by the base model (e.g., support vector machine), Platt scaling fits the parametric model:

$$(P(y=1|x) = \frac{1}{1+\exp(A \cdot f(x) + B)}) \tag{6}$$

Where A and B are hyperparameters estimated via maximum likelihood optimization on a validation dataset. Unlike non-parametric calibration methods, this approach maintains the computational efficiency and sparsity properties of the original support vector machine (SVM) while providing accurate probability estimates. The sigmoid fitting process preserves the discriminative power of the base model by leveraging its unthresholded outputs, which contain richer confidence information compared to binary predictions. Extensive empirical studies have demonstrated that Platt scaling consistently outperforms competing calibration techniques such as isotonic regression in scenarios requiring reliable probabilistic outputs for decision-making tasks [10].

## 3. Modeling and Solving

### 3.1. Prediction of Awarded Countries

### 3.1.1. Random Forest Time Series Forecasting Model

Aiming at the high-dimensional, nonlinear and time-series correlation characteristics of medal data, this study chooses to build a dual-engine prediction framework. In the first step, a random forest model was built for time series prediction using the filtered data. According to the time series division of the dataset, 80% of the dataset is the training set and 20% is the test set, this study selects the dynamic change of the number of gold, silver and bronze medals of each country over time as the feature to train the random forest model. In order to optimize the model structure and reduce the risk of overfitting, we set the number of tree models to 10, and at the same time take advantage of the characteristic of feature randomness to allow the model to be trained on different subsets of features, so as to enhance the generalization ability of the model. We used the Sliding Window Validation technique for cross-validation. The training window is set to be three consecutive sessions, i.e., the data of (t - 2, t - 1, t) are selected as a training window, and the prediction target is set to be the distribution of medals in t + 1 session. Input the above parameters into the built model to get the number of medals(D1) predicted by the random forest model. Finally, Fig.1 shows the predictions of

the number of gold, silver, and bronze medals in the Olympics, using the United States as an example, and visualizes these predictions.

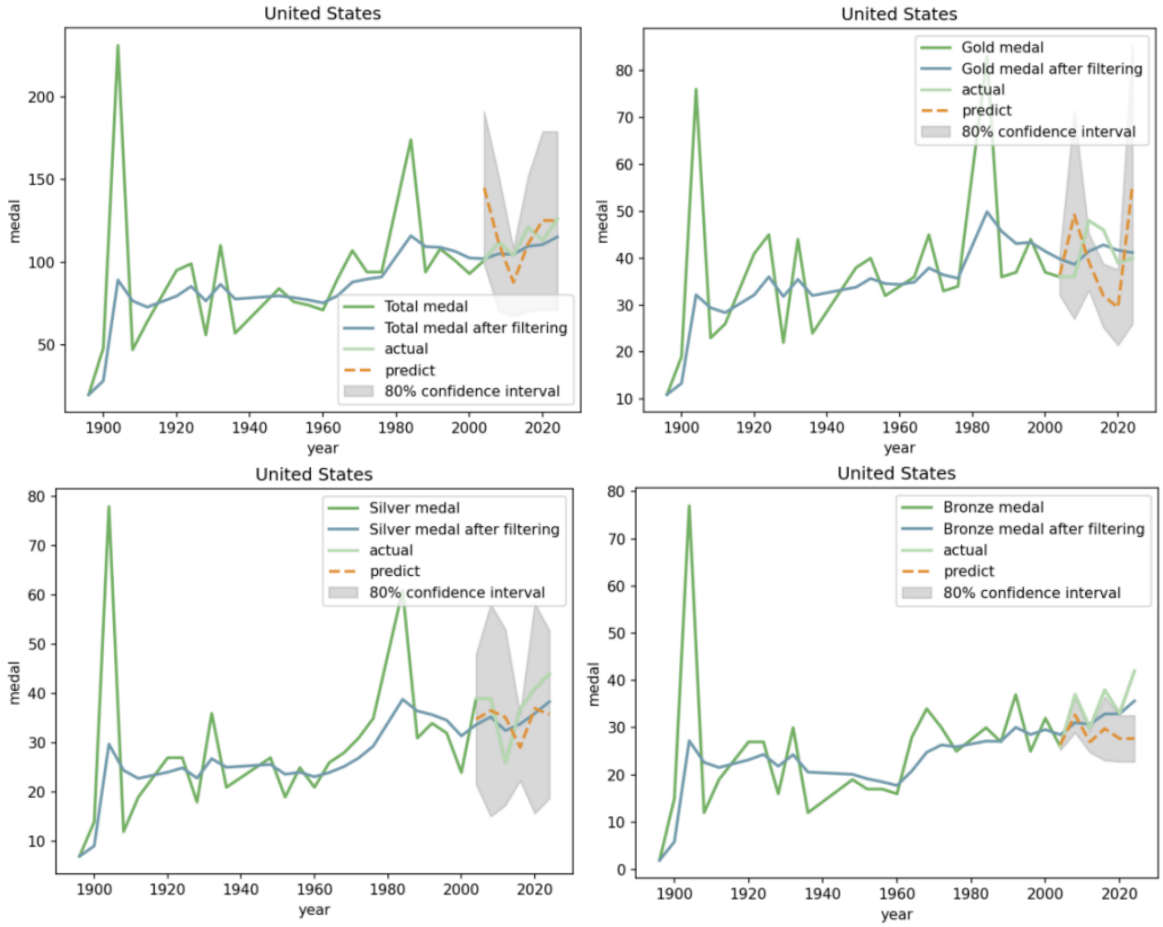


**Figure 1.** The results of Random Forest model.

The four pictures in Fig.1 show the time-series evolution of the number of U.S. Olympic medals (gold, silver, bronze and total medals) and the comparison of the prediction effect, in which the dark green solid line shows the characteristics of the original data before filtering, which shows significant non-smooth characteristics, especially in the host years such as 1984 (Los Angeles), 1996 (Atlanta) and other impulsive peaks, and the blue solid line is the filtered training set data, which can be The blue solid line is the filtered training set data, which can be clearly seen that the filtering effectively suppresses the abnormal fluctuations caused by the host effect, making the solid line smoother. The light green solid line is the actual data of the test set, and the orange dashed line is the predicted value of the test set. Grey areas indicate confidence intervals at 80% confidence level. From the figure, it can be seen that the trends and distances of the light green solid line and the orange dashed line are similar.

### 3.1.2. Construction of Olympic medal prediction based on XGBoost Model

Due to the non-linear relationship between the type and number of Olympic events and the number of medals, this study uses a machine learning approach to explore the relationship between the independent and dependent variables. XGBoost gradually adds new models in each iteration to expand its understanding of new data. Its formula is as follows:

$$M_b = M_{b-1} + \eta \cdot T_b \tag{7}$$

The goal of XGBoost training is to minimize the loss function, which includes both traditional loss functions and regularization terms that describe model complexity. Achieves a balance of accurate error correction and complex pattern capture. Based on the filtered historical Olympic data, a 3D

feature tensor $T \epsilon \mathbb{R}^{C \times Y \times E}$ is obtained where C: represents the total number of NOC, and Y represents the year span (1896-2024), E represents the project category.

Before building the prediction model, this study counted the number of events and their medals participated by each country in the medal table of previous Olympic Games, and calculated the medal share of each event. Then, this study input these data into the XGBoost model to evaluate the correlation between the winning of each sport and the number of medals. For model optimization design, this study used Bayesian optimization for hyperparameter search, and the objective function was a time-series cross-validated weighted MAE:

$$wMAE = \frac{1}{\sum w_y} \sum_{y=2000}^{2020} w_y \cdot \left| \hat{M}_y - M_y \right|$$

$$w_y = 1.2^{2024-y} \tag{8}$$

This study constructs a medal prediction framework based on the XGBoost model and introduces the SHapley Additive exPlanations (SHAP) method to enhance the interpretability of the model. The SHAP values quantify the contribution of specific item features to the final prediction results by decomposing the prediction results of each sample.

In conjunction with the program setup for the 2028 Olympic Games in Los Angeles, we have predicted the number of medals (D2) that each country will win at the next Olympic Games. Fig.2 shows the effect of using the XGBoost model to predict the number of gold and total medals for the United States in the next Olympics using the XGBoost model, with the dashed line showing the predicted results and the solid line indicating the true number of medals.

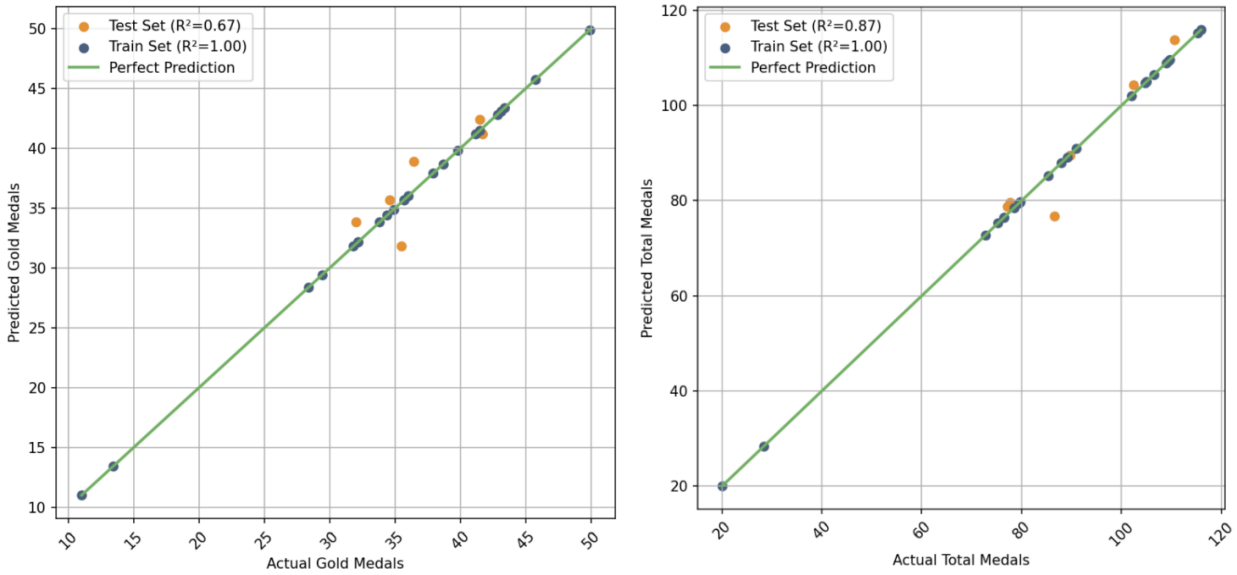


**Figure 2.** The results of XGBoost model.

Specifically, for each predicted sample, the model produces a predicted value, and the SHAP value is the value assigned to each feature in that sample. Suppose the i sample is $x_i$, the j feature of the i-th sample is $x_{i,j}$, the model's predicted value for the i sample is $y_i$, and the baseline for the entire model (usually the mean of the target variable over all samples) is $y_{base}$, then the SHAP value obeys the following equation.

$$y_i = y_{base} + f(\mathrm{x}_{i,1}) + f(\mathrm{x}_{i,2}) + \cdots + f(\mathrm{x}_{i,k}) \tag{9}$$

Where $f(\mathrm{x}_{i,1})$ is the SHAP value of $x_{i,j}$. Intuitively, $f(\mathrm{x}_{i,1})$ is the contribution value of the 1st feature in the i sample to the final predicted value $y_i$, when $f(\mathrm{x}_{i,1}) > 0$, it means that the feature

improves the predicted value, and it also has positive effect; on the contrary, it means that the feature makes the predicted value lower, and it has negative effect. The result of visualizing the SHAP values obtained by running the XGBoost model is shown in Fig.3, which demonstrates that the the impact of different events on the total number of medals by the United States(top9).

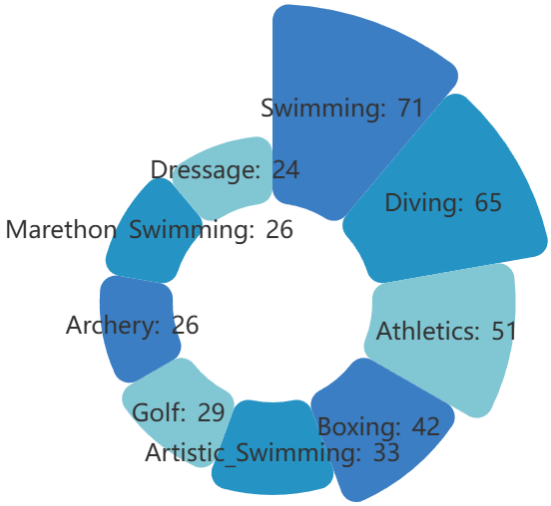


**Figure 3.** SHAP value for each event.

This paper analyzes the predicted data with the actual situation and found that the correlation calculated by XGBoost is small for low volatility sports, so we introduced the winning rate to assist the prediction and improve the realism of the prediction. For the United States, basketball is the sport with the highest medal winning rate, but it is less important for the XGBoost model, probably caused by its small variation in the number of evens. We calculated the winning rates of gold, silver, bronze and total medals for each country in each event based on how the events will be run in 2028, and finally calculated the number of medals ($D_3$) scenario. For example, the formula for calculating the gold medal winning rate for a country c in a single event is as follows:

$$r_{c,e} = \frac{M_{cumulative}}{\sum_{t=1}^{T} P_{c,e,t}} \tag{10}$$

Where $r_{c,e}$ represents country c's winning rate in item e. T denotes the total number of sessions (e.g., 33 Summer Olympics from 1896 to 2024). $P_{c,e,t}$ denotes the number of participations of country c in the t Olympic Games in item e (if no award is won, the number of participations is still counted in the denominator).

The right of Olympic host countries to propose new sports or emphasize specific disciplines introduces a geopolitical dimension to medal distribution patterns. Host countries typically exploit this institutional advantage through two synergistic mechanisms: 1) strategic event selection favoring sports in which the country has historically dominated; 2) targeted allocation of resources for athlete development programs and world-class training facilities. Thus, the host country's probability of winning a prize can be significantly increased through judicious selection of events.

### 3.1.3. Prediction Results of Ensemble Prediction Model

To comprehensively integrate temporal dynamics, event category variations, host nation advantages, and national performance efficiency, we developed an ensemble prediction model with the following formulation:

$$\hat{M} = \underbrace{(\lambda_1 D_1 + \lambda_2 D_2 + \lambda_3 D_3)}_{Model\ Ensemble} \times \underbrace{exp(\beta H)}_{Host\ Effect} \tag{11}$$

Where $\hat{M}$ is the predicted number of medals, $\lambda_1$, $\lambda_2$, $\lambda_3$ are the weighting coefficients of the predicted value of the Random Forest model, the predicted value of the XGBoost model, and the predicted value of the award rate, respectively, and exp ( $\beta H$ ) is the impact on the number of medals caused by the host effect. Since $\lambda_1+\lambda_2+\lambda_3=1$, for simplicity of computation we assume here that all three have the same weighting. ( $\lambda_1=\lambda_2=\lambda_3=\frac{1}{3}$ ).

Exp ( $\beta H$ ) is the coefficient of the host effect. Using statistical modeling methods, screening out the previous Olympic Games on the prediction of representative medal data, through regression analysis model calculated in the case of the United States as the host country of the value of 1.40, for other non-host countries in the calculation of taking 1.

### 3.1.4. Model Evaluation

**Table 1.** Model Evaluation

| Evaluation Tndicators | RF | XGBoost |
|---|---|---|
| RMSE | 11.4 | 4.42 |
| MAE | 14.0 | 3.11 |
| $R^2$ | 0.71 | 0.87 |
| MBE | 4.7 | -0.23 |

Analyzing Table 1 reveals that the RF model takes into account fewer influencing factors and has a higher error, whereas the XGBoost model predicts relatively better results but also has a certain amount of error

## 3.2. Prediction of First-Time Winning Countries

### 3.2.1. The Establishment of SVM Model

An analysis of the characteristics of countries that have won medals for the first time at the past three Olympic Games reveals that the winning countries are generally characterized by a wealth of experience and by a variety and concentration of participation in various sports. Specifically, the accumulation of Olympic experience and the concentration of participation in certain events increase the likelihood of the country winning a medal. Therefore, the following three characteristics were selected as model inputs:

The number of Olympic Games in which the country has participated ( $O_i$ ): the number of times the i country has participated in the Olympic Games is summed up to reflect the country's Olympic experience.

The number of events in which the country has participated ( $E_i$ ): the number of all Events in which the i country has participated in all previous Olympics is summed up to reflect the country's level of participation in Olympic competitions. Also define the number of times the country participates in the sport as $e_i$ .

Herfindahl-Hirschman Index ( $HHI_i$ ): HHI is a commonly used measure of concentration. It is usually used in the economy to measure market concentration, but is also suitable for measuring the degree of concentration of countries in sports, reflecting the country's focus and strength in certain sports.

$$HHI=\sum_{i=1}^{n}p_i^2 p_i=\frac{e_i}{E_i} \tag{12}$$

Where HHI values closer to 1 indicate that the country is more concentrated on a particular type of sport; HHI values closer to 0 indicate that the country's participation in the sport is more decentralized.

The target variable ($M_i$) is whether or not the country won the award, with a value of 1 if it won the award and 0 if it did not. We divided the dataset into training and testing set, with 70% of the training set. For all features, the following normalization formula is used for processing:

$$x_{scaled} = \frac{x - \mu}{\sigma} \tag{13}$$

Where X is the original data, μ is the mean of the feature, and σ is the standard deviation of the feature.

Support Vector Machine (SVM) is a generalized linear classifier for binary or multiclassification problems. SVM achieves classification by finding the maximum margin hyperplane, and its decision boundaries can effectively differentiate between samples of different classes. The SVM method has superior robustness, and can show strong generalization ability in the case of small amount of data, and shows better performance compared with other traditional machine learning algorithms. performance compared to other traditional machine learning algorithms.

### 3.2.2. Projected Results

Five countries are predicted to win the 2028 U.S. Olympic Games in Los Angeles for the first time, with the following probabilities of winning.

**Table 2.** First-time winners of the 2028 Olympics

| Team | NOC | Probability |
|---|---|---|
| Nepal | NEP | 0.8568819523297446 |
| Guam | GUM | 0.8180144493443432 |
| Papua New Guinea | PNG | 0.712949500885763 |
| Mali | MLI | 0.708166571466395 |
| Sierra Leone | SLE | 0.6841524700131688 |

Looking at the data in the Table 2, Nepal has the highest projected probability of winning at 0.8569, indicating that the country has a high chance of winning a medal at the 2028 Los Angeles Olympics. This is closely followed by Guam (Guam) with a probability of winning at 0.8180, also showing a high degree of competitiveness. Other countries, such as Papua New Guinea (Papua New Guinea), Mali (Mali) and Sierra Leone (Sierra Leone), have slightly lower probabilities of winning, but still show some potential, especially Papua New Guinea (0.7129) and Mali (0.7082), suggesting that their probability of winning a medal for the first time in the Olympics should not be underestimated.

### 3.2.3. Model Performance Evaluation

**Table 3.** Model evaluation results

| | Accuracy | Recall | Precision | F1 | AUC |
|---|---|---|---|---|---|
| Training set | 0.914 | 0.914 | 0.835 | 0.873 | 0.984 |
| Test set | 0.923 | 0.923 | 0.852 | 0.886 | 0.962 |

From Table 3, the model shows high accuracy and good generalization ability on both the training and test sets, especially the high AUC values and good F1 scores indicate that the model is very capable of differentiating in classification tasks. The slight improvement on the test set also indicates that the model has a better ability to adapt to new data that may be encountered in practical applications. In conclusion, the model is able to predict better which countries are likely to win medals for the first time.

### 3.3. The Ensemble Prediction Model Results

We used an integrated prediction model to predict the gold, silver, bronze and overall medal rankings for the 2028 Olympics, as shown in Table 4.

**Table 4.** 2028 medal table PREDICTED (top 10)

| Rank | NOC | Gold | Silver | Bronze | Total |
|---|---|---|---|---|---|
| 1 | USA | 62 | 57 | 45 | 164 |
| 2 | CHN | 33 | 19 | 18 | 70 |
| 3 | GBR | 16 | 19 | 20 | 55 |
| 4 | GER | 13 | 16 | 16 | 45 |
| 5 | AUS | 14 | 11 | 18 | 43 |
| 6 | JPN | 16 | 12 | 13 | 41 |
| 7 | FRA | 10 | 15 | 15 | 40 |
| 8 | ITA | 10 | 10 | 15 | 35 |
| 9 | KOR | 11 | 8 | 12 | 31 |
| 10 | NED | 8 | 10 | 12 | 30 |

## 4. Conculstion

In this study, we propose a multimodal integrated learning framework that constructs a dual-channel prediction architecture for complex systems by synergizing Random Forest (RF), Extreme Gradient Boosting (XGBoost), Statistical Modeling (SM), and Support Vector Machines (SVMs), aiming at high-precision prediction in complex situations. The fusion model integrates the correlations between multiple factors, such as different categories, trends and effects, and is applied to the analysis of different datasets separately. Evaluation of the model's prediction results shows that the fusion model's prediction accuracy is significantly better than that of any single model and is able to consider the dimensions of the problem more comprehensively. Models trained individually for a specific dataset are able to capture data features more effectively, resulting in more reliable predictions.

Future research directions will focus on two critical dimensions for enhancing the predictive accuracy and interpretability of the hybrid framework: (1) Expanding influential influences and incorporating emerging areas of research to select metrics that better measure model performance to evaluate models; (2) Development of a theoretically grounded weight optimization mechanism that adopts advanced computational intelligence algorithms to establish dynamic weight allocation schemes, validated through rigorous experimental comparisons against traditional gradient-based optimization approaches.

## References

[1] Jiang Y, Wan J P. Prediction of Freight Volume Based on Grey Correlation and Improved Grey Neural Network; proceedings of the 19th Annual Wuhan International Conference on E-Business (WHICEB), Wuhan, PEOPLES R CHINA, F Jul 05, 2020 [C]. 2020.

[2] Al Kindhi B, Dewi R A, Santosa N, et al. Prediction of the Unemployment and Bank Interest Rates on Changes in the Stock Price Index with Efficient Regression; proceedings of the 10th IEEE International Conference on Communication, Networks and Satellite (IEEE COMNETSAT), Electr Network, F Jul 17 - 18, 2021 [C]. 2021.

[3] Ratku A, Neumann D. Derivatives of feed-forward neural networks and their application in real-time market risk management [J]. Or Spectrum, 2022, 44 (3): 947 - 65.

[4] Yan Z, Chen H, Dong X H, et al. Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost [J]. Expert Systems with Applications, 2022, 207.

[5] Amiri A F, Oudira H, Chouder A, et al. Faults detection and diagnosis of PV systems based on machine learning approach using random forest classifier [J]. Energy Conversion and Mangement, 2024, 301.

[6] Rhodes J S, Cutler A, Moon K R. Geometry- and Accuracy-Preserving Random Forest Proximities [J]. IEEE Transactions on Pattern Analysys and Machine Intelligence, 2023, 45 (9): 10947 - 59.

[7] Jisi C, Roh B-h, Ali J. An effective scheme for classifying imbalanced traffic in SD-IoT, leveraging XGBoost and active learning [J]. Computer Networks, 2025, 257: 110939.

[8] Bansal S, Mehan V. Image retrieval of MRI brain tumour images based on SVM and FCM approaches [J]. 2021, 17 (3): 173 - 9.

[9] Shi C. Identifying Abnormal Corporate Financial Data Based on the Comparison of SVM and Logistic Algorithms; proceedings of the 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), F 8-9 April 2023, 2023 [C].

[10] Abe S. Do Minimal Complexity Least Squares Support Vector Machines Work? [Z]. Artificial Neural Networks in Pattern Recognition, ANNPR 2022. 2023: 53 - 64.10.1007/978 - 3 - 031 - 20650 - 4_5.