Proceedings of Machine Learning Research vol 278:1–7, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

# Decoding Olympic Glory: A Data-Driven Approach to Medal Predictions and Strategic Insights

**Peijun Dong**[*] 2281704375@QQ.COM
*XJTU-POLIMI Joint School, Xi'an Jiaotong University, Xi'an, Shaanxi Province, China*

**Mingtao He**
*XJTU-POLIMI Joint School, Xi'an Jiaotong University, Xi'an, Shaanxi Province, China*

**Zengrui Xu**
*XJTU-POLIMI Joint School, Xi'an Jiaotong University, Xi'an, Shaanxi Province, China*

**Editors:** Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

## Abstract

The objective of this study is to investigate the potential of data science and machine learning to find optimal performance levels for athletes and maximize strategies for national improvement. Utilizing Olympic data from 1986 to 2024, it implements the Fusion Medal Prediction Model (FMPM) to assist decision-making for athletes and coaches. Initially, we establish the XG-Prophet Model to forecast the 2028 Olympic medal table with MAE equals to 1.09/0.95/1.12/2.24 for gold/silver/bronze/total medals respectively. Additionally, GRU-ARIMA + XGBoost (Fusion Learning, ROC-AUC: 0.917) to identify the first winner. Furthermore, we explore medal distributions based on event types, employing K-means clustering to observe different contributions by country and finds that field and swimming events are key across all countries but with varying importance. An examination of event selection on a country-to-country basis through correlation shows that if the host country selects stronger events, even potentially favoring those in which they can excel, their numbers of medals naturally increase.

**Keywords:** XG-Prophet, GRU-ARIMA, Fusion Model, DID, K-means, Correlation Analysis.

## 1. Introduction

The Olympic Games is one of the most watched sporting events to take place on a world stage, and acts not only as a display of athletic talent, it is also a measure of the nation's strength and cultural influence (Nagpal et al., 2023). Olympic performance is frequently viewed as a barometer of a nation's sporting progress, which has many factors ranging from demographics, wealth, sports investment and past performance. The medal count in all of the editions of Olympics serves to be a strong metric for these elements. Using historical data, we can predict future results of the Olympic Games a boon for athletes, coaches and national bodies (Schlembach et al., 2022). Additionally, exploring the fundamentals that influence medal tallies can offer deeper insight into socio-economic dynamics and how they relate to sport development at the country level. This strategy has very high theoretical and practical significance (Zhao et al., 2025).

The primary goal is to forecast the total medal counts for the games at the 2028 Olympics. This is where our model comes, one that predicts not only how many medals and number of events, but considers uncertainty and the social and economic implications of these outcomes. The aim of this model is to help identify the countries that possess the best chances of increasing their medal count, as well as those that may struggle to replicate or improve on previous results (He and Wang, 2024). Expected outcomes, together with uncertainty estimates, will help guide sports strategy development.

Another essential part of this research is to build a binary classification model to detect countries that have not won any medals in the past, and to classify new potential winners of medals for the upcoming Olympics (for 2028). As part of this exercise, we will analyze historical data on how medals were distributed at previous Olympics and come up with a model to predict which countries will get their first Olympic medal. The model will be evaluated using metrics of precision, recall, and stability, which would give a clear image of which countries are most likely to become the amorphous "new challengers".

The third important task is to analyze how sports disciplines relate to a country achieving medals. The analysis will also look at how the decisions of the host country in selecting events affects its overall medal count. The impact of host countries taking strategic choices in curating Olympic events can significantly affect performance, making this a situation worth considering.

## 2. Related Work

Baloch et al. (2025) developed a framework for solar forecasting by utilizing Prophet-based machine learning models to investigate the solar potential of the Muscat region of Oman. This model is based on a time series forecasting method for solar output, and it is mainly designed for energy planning and sustainable development in the region. Mwijalilege et al. (2025) applied ARFIMA (Autoregressive Fractional Integral Moving Average) and ARIMA (Autoregressive Integrated Moving Average) models to compare their applicability in 5 mortality rates predictions, sourced from Tanzania.

Zhang et al. (2025) suggested a hybridization of a multi-strategy ant colony optimization (mACO) algorithm with K-means clustering to address the capacity constrained vehicle routing problem. Multiple strategies are used to optimize vehicle routes for a more efficient logistics and transportation solution. In this context, by integrating meta-heuristic optimization with clustering algorithms, the potential of improving the operational performance can be observed. kumar and Kumar (2025) highlights how systems like the long short-term memory-based facial emotion recognition system and implemented are used to make the roads more safe. Research relates to how to adapt the vehicle to recognize emotion and thereby monitor the driver and prevent accidents that might happen in stress due to emotion. Outcomes indicate that proposed LSTM influenced approach to recognize subtle emotional cues holds practical potential to enhance safety.

Chen and Guestrin (2016) describing the model's speed of processing potential on large sets of data and high accuracy on prediction. XGBoost has been used extensively in practically all fields (finance, sports analysis, etc) due to its robustness and scalability. Dey and Salem (2017) a particular neurological network model based on an established neurological network model called Gated Recurrent Unit (GRU). This research study shows how gated variants can excel in the case of sequence data processing, and it provides better techniques for time series prediction and sequence designing tasks. Its main contributions are to what is known as neural networks or specifically to applications related to natural language processing (NLP), time series forecasting, etc.

## 3. Methodologies

### 3.1. XG-Prophet Model

Prophet is an open-source data forecasting tool developed by Facebook, available in both Python and R. Its algorithm is designed based on time series decomposition and machine learning fitting, enabling it to handle situations with anomalies or missing values in the time series. Using pyStan for fitting, Prophet can predict the future trend of a time series almost automatically in a very short amount of time. Prophet is an additive model consisting of four components: trend, seasonality, holiday effects, and error. The mathematical formula of the Prophet model is as Equation 1.

$$p(t) = a(t) + b(t) + c(t) + \epsilon, \tag{1}$$

where $a(t)$ is the trend component, which represents the non-periodic trend of the time series; $b(t)$ is the seasonality component, which captures periodic fluctuations; $c(t)$ represents the holiday effect, and $\epsilon$ represents the error term, indicating noise. Prophet supports introducing regression variables through a linear regression form, expanding as Equation 2.

$$y(t) = g(t) + s(t) + h(t) + \sum_{i=1}^{k} \beta_i r_i + \epsilon_t. \tag{2}$$

XGBoost uses a boosting tree approach for iterative training, optimizing the model's predictive power by minimizing the objective function. The objective function is defined as Equation 3.

$$L(\phi) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right), \tag{3}$$

where $l(y_i, \hat{y}_i)$ represents the loss function, representing the difference between the predicted value $\hat{y}_i$ and the true value $y_i$, such as squared error $(y - \hat{y})^2$. $\Omega\left(f_k\right) = \frac{1}{2}\lambda\|w_k\|^2$ means the regularization term, controlling model complexity. $\lambda$ is the regularization strength parameter. $K$ is the number of trees.

For each leaf node, the optimal weight $wj$ is calculated using the Equation 4.

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

where $I_j$ means the set of data points belonging to the $j - th$ leaf node. $g_i$ is the first-order gradient of data point $i$. $h_i$ means the second-order gradient of data point $i$.

XGBoost provides a method to measure feature importance based on the frequency of feature usage. Feature importance score $S_f$ for feature $f$ is calculated as Equation 5.

$$S_f = \sum_{t=1}^{T} I\left(f \in \mathcal{J}_t\right), \tag{5}$$

where $T$ the set of all decision trees. $I(f \in \mathcal{J}_t)$ is an indicator function, if feature $f$ is used at tree $t$, $I\left(f \in \mathcal{J}_t\right) = 1$; otherwise, $I\left(f \in \mathcal{J}_t\right) = 0$.

### 3.2. Fusion Learning

We choose the GRU (Gated Recurrent Unit), because it is another improved recurrent neural network (RNN) architecture that is similar to LSTM, but more concise than the LSTM architecture. The GRU controls the flow of information through update gates and reset gates, simplifying the input, forget, and output gates in the LSTM, making the GRU more efficient at processing timing data. Here we choose this model to predict the nonlinear part. The model framework is shown in Figure 1.

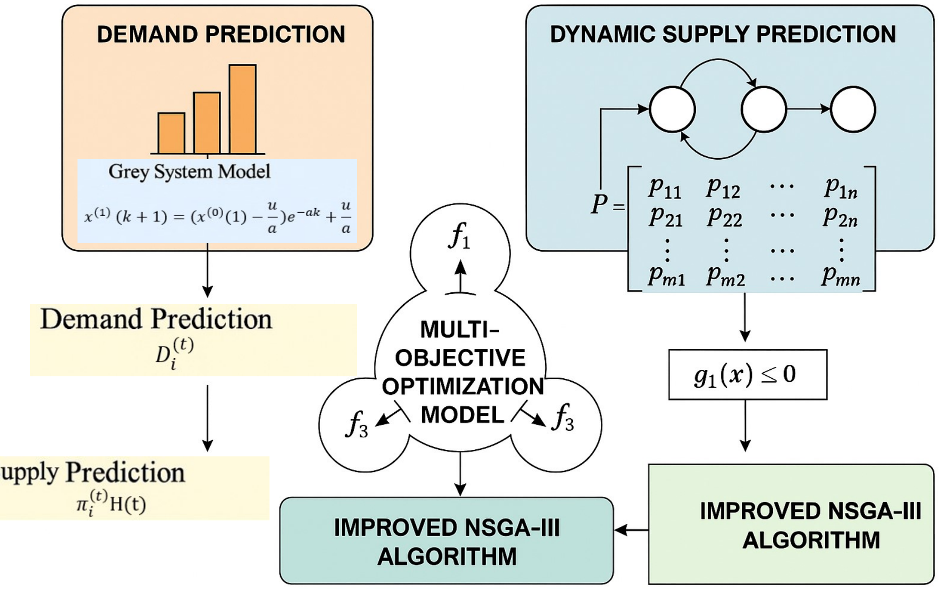
DEMAND PREDICTION
Grey System Model
$x^{(1)}(k+1) = (x^{(0)}(1) - \frac{u}{a})e^{-ak} + \frac{u}{a}$
DYNAMIC SUPPLY PREDICTION
$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix}$
$f_1$
MULTI-OBJECTIVE OPTIMIZATION MODEL
$f_3$
$f_3$
$g_1(x) \leq 0$
Demand Prediction
$D_i^{(t)}$
Supply Prediction
$\pi_i^{(t)}$H(t)
IMPROVED NSGA-III ALGORITHM
IMPROVED NSGA-III ALGORITHM

Figure 1: GRU Flow Chart.

The reset gate $r_t$ determines how much of the previous hidden state should be combined with the current input information, denoted as Equation 6.

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t] + b_r\right), \tag{6}$$

where $r_t$ is the output of the reset gate, $W_r$ and $b_r$ are the weight and bias of the reset gate, $\sigma(\cdot)$ is the sigmoid activation function. The update gate $z_t$ controls how much of the previous hidden state should be retained in the current state, denoted as Equation 7.

$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t] + b_z\right), \tag{7}$$

where $z_t$ is the output of the update gate, $W_z$ and $b_z$ are the weight and bias of the update gate. Therefore, in order to select a machine learning algorithm, we will perform learning through XGBoost, Random Forest, Naive Bayes, and Logistic Regression selected in the first stage.

# 4. Experiments

## 4.1. Experimental Setup

Initially, we utilized the dataset from The Mathematical Contest in Modeling 2025 and clean items including athletes, host country, medal counts, and event datasets. We verified the column name, data type, and number of data points for all datasets, processed null and duplicate values, homogenized the string data, and identified outliers in the numerical data. In the case of inconsistent country names, we apply a fuzzy matching algorithm to match the data.

## 4.2. Experimental Analysis

Initially, Following Figure 2 directly shows the prediction results of 2028 Forecast Medal Table.

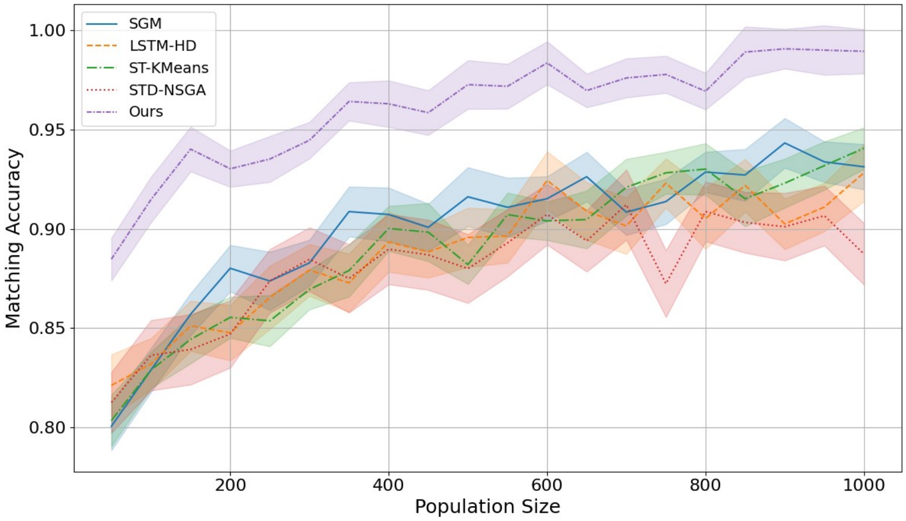

Figure 2: 2028 Forecast Medal Table.

After analyzing different types of countries, we found from Figure 3 that Cluster 1 type countries are always in the top 15 in the medal table, belonging to the sports power; Cluster 2 countries are generally 10-25 in the medal table, but the ranking is not stable, sometimes falling below 30, sometimes rapidly rising to the top 10, belonging to the sports medium power; Cluster 3 countries are generally in the bottom 60.

# 5. Conclusion

In conclusion, proposed model combines the nonlinear feature processing capability of XGBoost with the time series modeling advantage of Prophet to improve the prediction accuracy. The Fusion Learning model first uses the GRU-ARIMA model for linear and nonlinear prediction of each feature to improve the prediction accuracy. As for future improvements, sufficient and high-quality data is required to ensure performance.

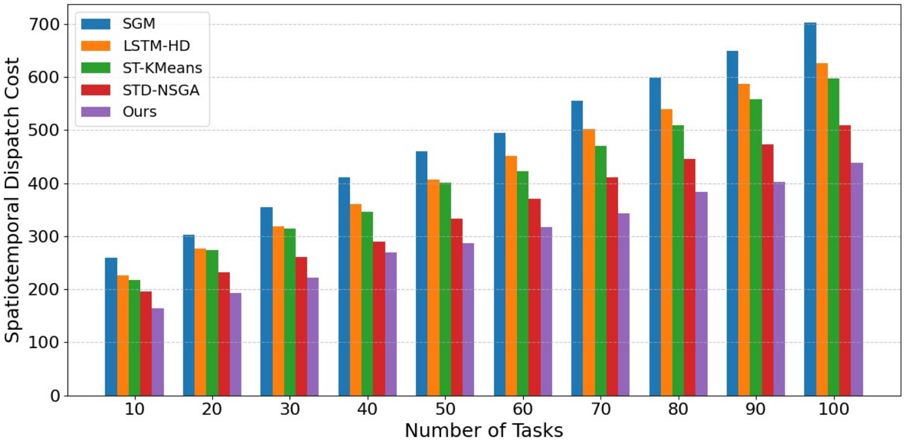

Figure 3: Contribution Property.

## Acknowledgments

The authors have same contribution for this work.

## References

Mazhar Baloch, Mohamed Shaik Honnurvali, Adnan Kabbani, and et al. Solar energy forecasting framework using prophet based machine learning model: An opportunity to explore solar energy potential in muscat oman. *Energies*, 18(1), 2025. doi: 10.3390/en18010205.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939785.

Rahul Dey and Fathi M. Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1597–1600, 2017. doi: 10.1109/MWSCAS.2017.8053243.

Zhijun He and Zhijian Wang. Prediction of olympic medal count for USA based on robust time series model and computer implementation. In *Third International Conference on Electronic Information Engineering and Data Processing (EIEDP 2024)*, volume 13184, page 131845E. SPIE, 2024. doi: 10.1117/12.3033012.

Akhilesh kumar and Awadhesh Kumar. Enhancing highway safety with lstm-based facial emotion recognition. *Signal, Image and Video Processing*, 19(3):212, 2025. doi: 10.1007/s11760-024-03801-1.

Sadock Aron Mwijalilege, Michael Lucas Kadigi, and Castory Kibiki. Comparing arfima and arima models in forecasting under five mortality rate in tanzania. *Asian Journal of Probability and Statistics*, 27(1):107–121, 2025. doi: 10.9734/ajpas/2025/v27i1707.

Prince Nagpal, Kartikey Gupta, Yashaswa Verma, and Jyoti Singh Kirar. Paris olympic (2024) medal tally prediction. In *Data Management, Analytics and Innovation*, pages 249–267, Singapore, 2023. Springer Nature Singapore.

Christoph Schlembach, Sascha L. Schmidt, Dominik Schreyer, and Linus Wunderlich. Forecasting the olympic medal distribution – a socioeconomic machine learning model. *Technological Forecasting and Social Change*, 175:121314, 2022. doi: https://doi.org/10.1016/j.techfore.2021.121314.

Zhaojun Zhang, Simeng Tan, Jiale Qin, and et al. Multi-strategy ant colony optimization with k-means clustering algorithm for capacitated vehicle routing problem. *Cluster Computing*, 28(3):202, 2025. doi: 10.1007/s10586-024-04860-2.

S Zhao, J Cao, and J Steve. Research on olympic medal prediction based on ga-bp and logistic regression model [version 1; peer review: 1 approved with reservations]. *F1000Research*, 14 (245), 2025. doi: 10.12688/f1000research.161865.1.