Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 11
TPCEE 2025

An Improved ARIMA-Neural Network Fusion Model for Multivariate Time-Series Prediction with External Factor Integration

Shaonan You, Jing Chen *, Liang Wei, Shuyi Zhou

Beijing Technology and Business University, Beijing, China

* Corresponding Author Email: 13310319353@163.com

Abstract. This study aims to solve the difficult problem of sports performance prediction and improve the accuracy and reliability of prediction. The improved ARIMA and neural network fusion technology was used to clean and standardize the sports data first, and then the improved ARIMA model was used to capture the time series features, and external factors such as athletes' physical fitness and training level were included in the model, and special cases were predicted by combining Bayes theory. Finally, neural network is used to further optimize the prediction results. The experimental results show that the fusion technology effectively integrates the advantages of the two models, and the prediction error is significantly reduced compared with the single model. The research conclusion is that the improved ARIMA and neural network fusion technology is feasible and effective in the field of sports performance prediction, which provides an innovative method for subsequent sports-related research and practice, and strongly promotes the development of sports data prediction technology.

Keywords: ARIMA Model; Neural Network; Bayesian theory; Fusion Model.

1. Introduction

In the realm of data - driven prediction and analytics, accurately forecasting complex time - series data with the influence of multiple factors is of utmost importance. This task has far - reaching implications across various fields such as finance, weather forecasting, and industrial production. With the exponential growth of data and the increasing complexity of real - world problems, traditional forecasting methods are facing significant challenges [1].

Traditional time - series analysis methods, like simple moving average and exponential smoothing, have limitations when dealing with non - stationary data and complex relationships. They often rely on the assumption of stationarity, which is rarely satisfied in real - world scenarios [2]. For instance, in financial time - series, stock prices can be affected by a multitude of factors such as economic policies, company earnings reports, and global market trends. These external factors can cause sudden changes in the data pattern, making it difficult for traditional methods to provide accurate forecasts.

Machine learning algorithms, on the other hand, have shown great potential in handling complex data. However, a single machine learning algorithm, such as a decision tree or a support vector machine, may not be sufficient when dealing with time - series data [3,4]. They lack the ability to capture the sequential nature and long - term dependencies inherent in time - series. For example, in weather forecasting, the temperature and precipitation values of consecutive days are closely related, and a model needs to be able to capture these relationships to make accurate predictions.

In recent years, there has been a growing trend of combining different models to leverage their respective strengths. Hybrid models that integrate time - series models and machine learning algorithms have emerged as a promising approach. Among these, the combination of the AutoRegressive Integrated Moving Average (ARIMA) model and neural networks has attracted significant attention. The ARIMA model is proficient in capturing the linear dependencies and trends in time - series data, while neural networks, with their powerful non - linear mapping capabilities, can handle complex non - linear relationships.

Despite the potential of this combination, few studies have focused on effectively integrating external factors into the ARIMA - neural network hybrid model to achieve more comprehensive and accurate predictions. This study aims to fill this research gap. By enhancing the ARIMA model and integrating it with neural networks, we construct a novel prediction model. This model can not only process the time - series characteristics of data but also incorporate external variables, enabling more accurate and reliable predictions. It is expected to provide a valuable solution for data - driven decision - making in various complex scenarios.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing

The research data comes from the public sports data statistics platform, covering a number of data indicators of athletes, such as physical fitness indicators, skill indicators and training performance. In the data preprocessing stage, the data is first cleaned to remove duplicates, excessive missing values and abnormal data records. For the data with a small number of missing values, interpolation method is used to fill in to ensure data integrity. Then, the data is standardized to convert the data of different dimensions into a unified scale to make the data more suitable for model training.

2.2. Model Overview

2.2.1. Model Selection

The predicted number of gold medals or total medals that each country may earn in the 2028 Olympic Games is closely related to the results of previous Summer Olympic Games. Therefore, we employ the time series ARIMA model, which is defined by three parameters:

p: The number of autoregressive terms (AR);

d: Indicates the number of differences, ensuring data stability；

q: The number of sliding average items. (MA).

Due to the fact that the achievements of each country in the Olympic Games are also related to other factors, such as the host country, the events of each Olympic Games and the ability of the athletes, we can not simply use a basic ARIMA model, but must adopt an extended ARIMA model based on the basic ARIMA model [5]. We use the following regression equation:

$$Y_t = \alpha+\beta_1 X_{t-1}+\beta_2 X_{t-2}+\ldots\ldots+\beta_{t-1} X_1+\gamma Z_t+\epsilon_t \quad (1)$$

Among them, Y_t is the target variable (number of gold medals or total medals), X_{t-1}, X_{t-2},…… X_1 indicate the number of gold medals or medals won in previous Olympic Games and assigns different weights to them. Z_t is the external variables include the host country, the events of the Olympic Games and the ability of the athletes. ϵ_t is the error term is used for interval prediction.

2.2.2. The Establishment of the Model

First, it is necessary to process the data by reading it in. Since a single NOC may appear multiple times in the first column of data, it is important to deduplicate it during the reading process. Secondly, it is necessary to conduct a stability test on the data. If the data is found to be unstable, differencing treatment is required. After data processing, the appropriate values of p and q are determined through the autocorrelation function (ACF) and the partial autocorrelation function (PACF), and then the ARIMA model is fitted based on the training data and its accuracy is assessed [6].

Based on this model, for countries that have never won a medal in the Olympic Games, we will add Bayesian theory to predict the posterior probability of these countries winning medals at the next Olympic Games. We have assigned a prior probability to each country, representing the preliminary

probability of these countries winning medals, which is related to the rankings achieved by these countries in previous Olympic Games [7]. The higher the ranking and the more times it has been achieved, the greater this probability becomes. In addition, we consider three factors that influence the awarding of prizes, namely the athletes' abilities, the economic level of the country and the sports infrastructure. The quantitative value range for each factor indicates from zero to one that a larger value represents a stronger influence [8-10]. In order to comprehensively consider these three factors. We assigned a certain weight to each factor, with athlete ability accounting for 45%, national economic level 35% and sports infrastructure 25%. Then, based on the weights, we calculated the impact of these factors on the promotion of these countries in earning their first medal. Finally, using Bayes' theorem to calculate the posterior probability, the formula is:

$$P(A|B) = P(B|A) * P(A) / P(B) \tag{2}$$

In this formula, P(A) represents the prior probability, P(A|B) Represents a posterior probability, P(B|A) indicates the likelihood based on influencing factors and P(B) denotes the sum of the prior probabilities of all countries. The posterior probability calculated reflects the likelihood of the countries, which did not win a medal, earning a medal in the next Olympic Games under the known influencing factors.

Figure 1 is a flow chart of the Bayesian model to calculate the posterior probability.

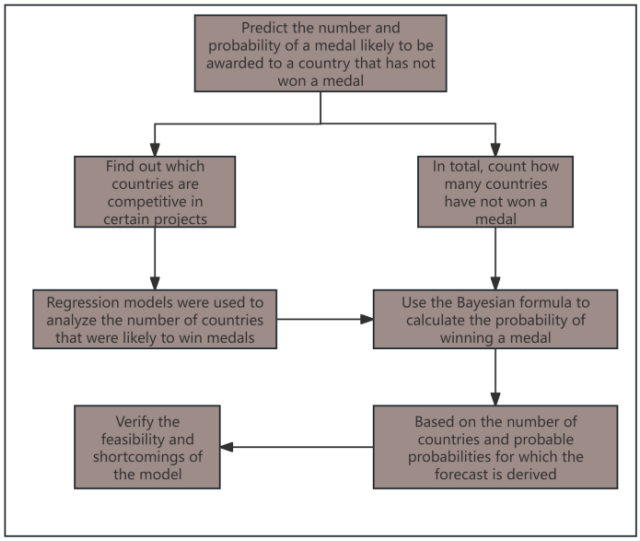

Figure 1. Bayesian Model Probability Calculation Process

2.3. The Evaluation of the Model

In order to measure the accuracy of the model, we divide the data into a training set and a testing set. We use the training set to train the model and the testing set to evaluate the model. The accuracy of the model is assessed by calculating the error between the predicted results and the actual data. The metrics we use for measurement are Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Their definitions and calculation formulas are as follows:

MSE: The mean of the squares of the errors between all predicted values and the actual values. The formula is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{3}$$

RMSE: The square root of the mean square error, which can be compared with the actual units of the data.

The formula is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{4}$$

MAE: The average of the absolute values of the errors between all predicted values and the actual values.

The formula is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{5}$$

Among them: y_i is the i-th true value;

$\hat{y}_i$ is the i-th predicted value;

n is the number of samples.

Based on the number of gold medals and total medals won by each country in previous Olympic Games, we assessed the potential prediction result errors for each country under this model, as shown in the Figure 2 and Figure 3, where the Figure 4 represents the legend.

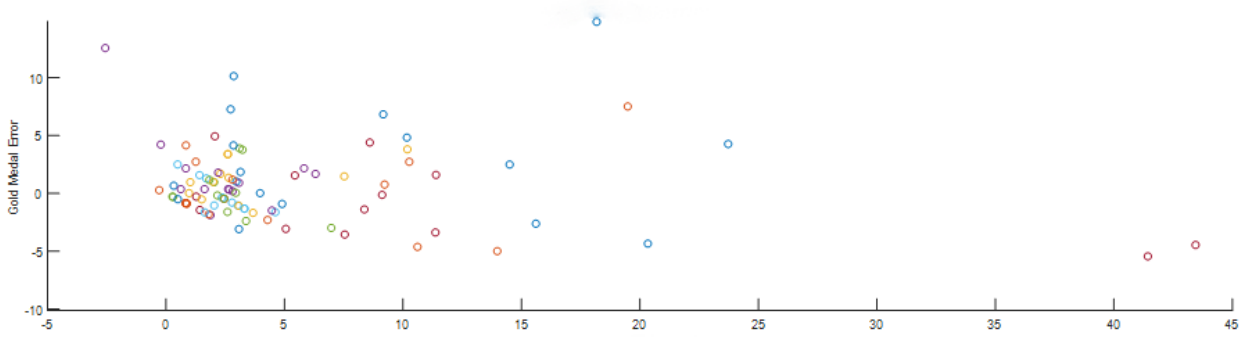

Figure 2. The Error of the Model Prediction Results (Gold Medals)

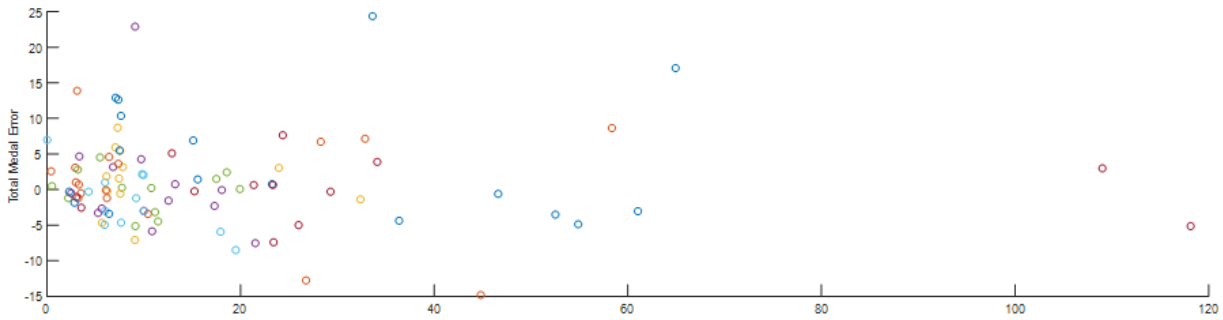

Figure 3. The Error of the Model Prediction Results (Total Medals)

Figure 4. Legend of Prediction Result Error

3. Result and Analysis

3.1. Model Configuration and Training Details

3.1.1. ARIMA Model Parameters

Taking China as an example, this study configured parameters of the improved ARIMA model and determined parameters through the analysis of autocorrelation function (ACF) and partial autocorrelation function (PACF), as shown in Figure 5 and Figure 6 below. After differentiating the non-stationary data (d=1), the optimal parameters p=2 and q=3 are selected as the model parameters for predicting total medals, and the optimal parameters p=0 and q=1 are selected as the model parameters for predicting gold medals. Finally, external factors such as athletes' physical fitness, training level and host country advantages are incorporated into ARIMA framework, which significantly improves the adaptability of ARIMA framework to dynamic time series model.

3.1.2. ARIMA Model Parameters

In order to verify the superiority of the fusion model, the performance of the fusion model was compared with that of the standalone ARIMA model on a test dataset. Using Germany as an example in Table I, the results show a 35.8% reduction in the RMSE predicted for Gold Medals and a 25.0% reduction in the RMSE predicted for Total Medals compared to the best single model.

Table 1. Prediction Error Comparison Across Models

Model	Gold Medals (RMSE)	Total Medals (RMSE)	Gold Medals (MAE)	Total Medals (MAE)
Single Model	4.32	9.87	3.12	8.59
Fusion Model	2.77	7.40	2.24	6.17

3.2. Medal Table Prediction

3.2.1. Interval Prediction Results

Based on the performance of each country in previous Olympic Games, combined with factors such as the host country effect, the events set for each Olympic Games and the ability of the athletes , a range-based prediction of the number of gold medals (as shown in the Figure 7) and total medals (as shown in the Figure 8) that each country may earn in the 2028 Olympic Games has been made using the ARIMA model.

The uncertainty interval in Figure 7 and Figure 8 reflects the influence of external variables on the ARIMA model. For example, due to the advantage of the host country, the uncertainty of the medal prediction of the United States is only about 10.4%, which is much lower than the uncertainty of other countries such as China's medal prediction of 17.9%, the Great Britain's medal prediction of 22.2%, and the Japanese medal prediction of 28.0%.

3.2.2. Partial Medal Table

According to the results predicted based on the aforementioned intervals, we use the midpoint of the interval as the possible number of gold medals or total medals that the countries may earn in the 2028 Olympic Games. We have combined both to create a chart of the predicted values for the number of gold medals and total medals earned by each country in the 2028 Olympic Games, as shown in the Figure 9.

According to the Figure 9, this study present a prediction of the top ten medal table for the 2028 Olympic Games, as shown in the Table II. The sorting here is done in descending order based on the number of gold medals and the total number of medals.

Table 2. Medal Table in the 2028 Olympic Games (Top 10)

RANK	NOC	Gold Medals	Total Medals
1	United State	46	135
2	China	37	84
3	Great Britain	26	73
4	Japan	20	51
5	Australia	17	52
6	France	12	51
7	Netherlands	12	33
8	Italy	10	38
9	South Korea	10	25
10	Germany	7	24

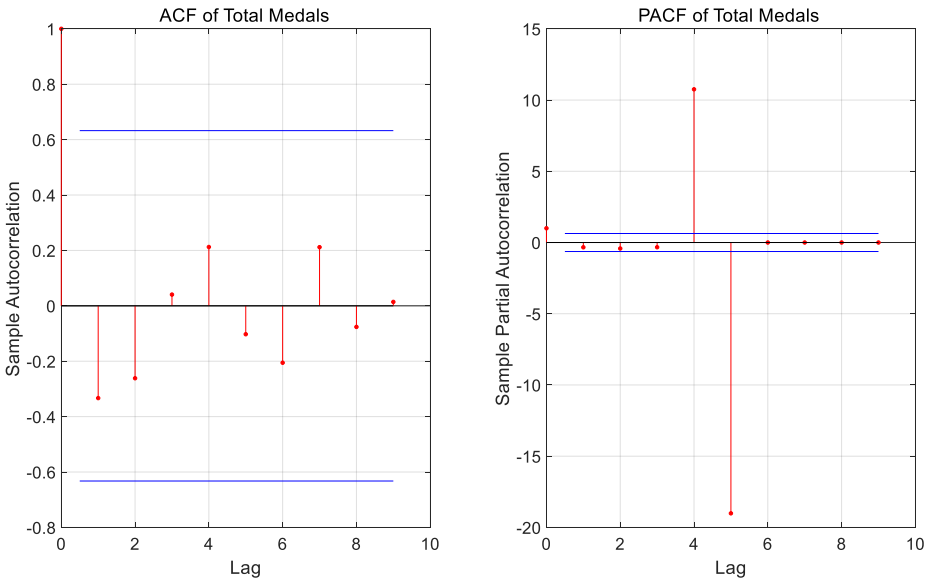

Figure 5. ACF and PACF of Total Medals

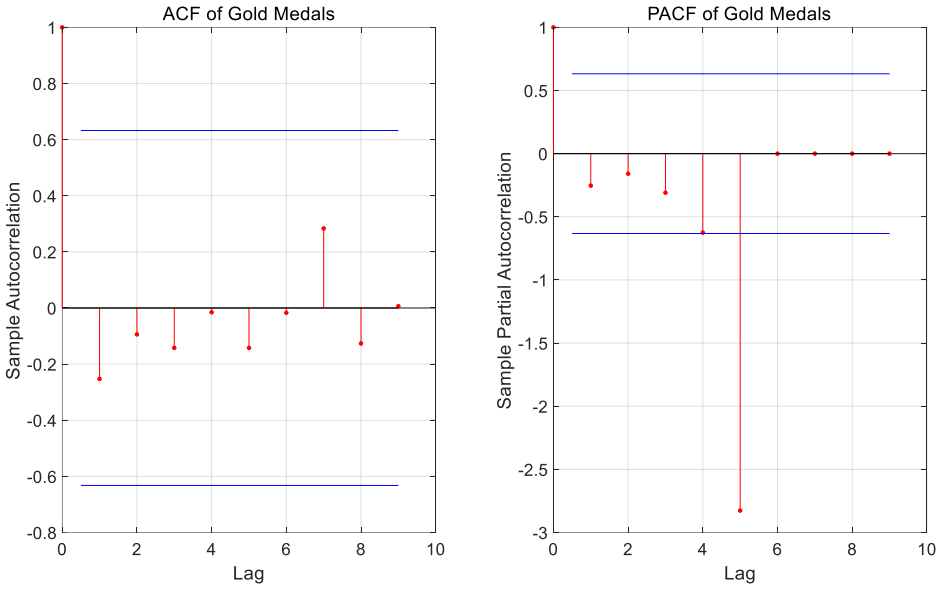

Figure 6. ACF and PACF of Gold Medals

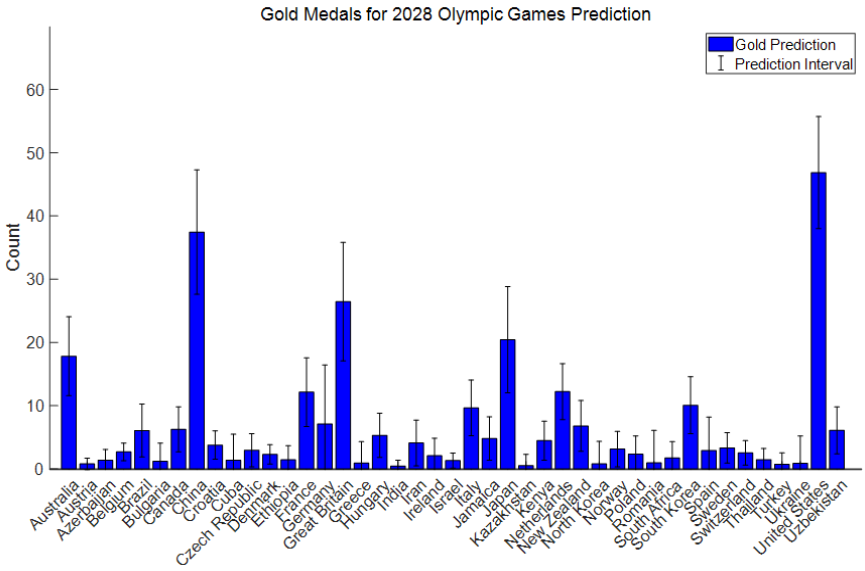

Figure 7. Model Interval Prediction of Gold Medals

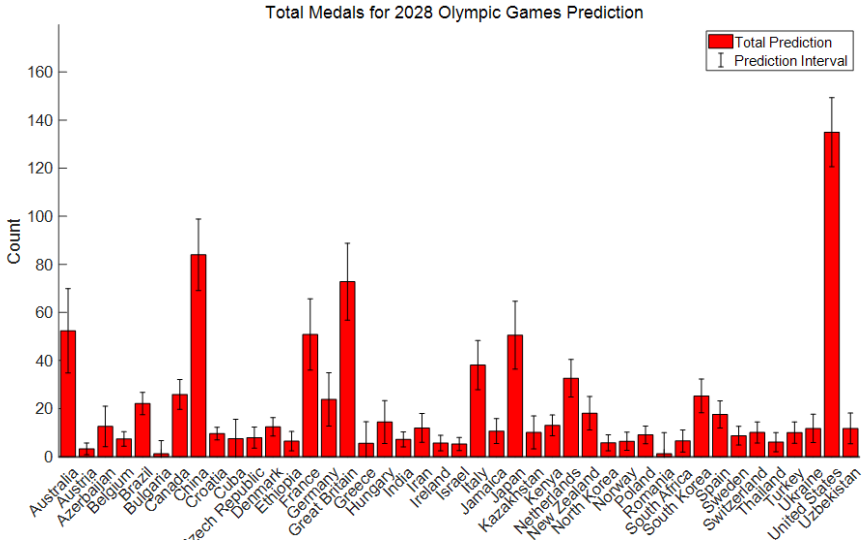

Figure 8. Model Interval Prediction of Total Medals

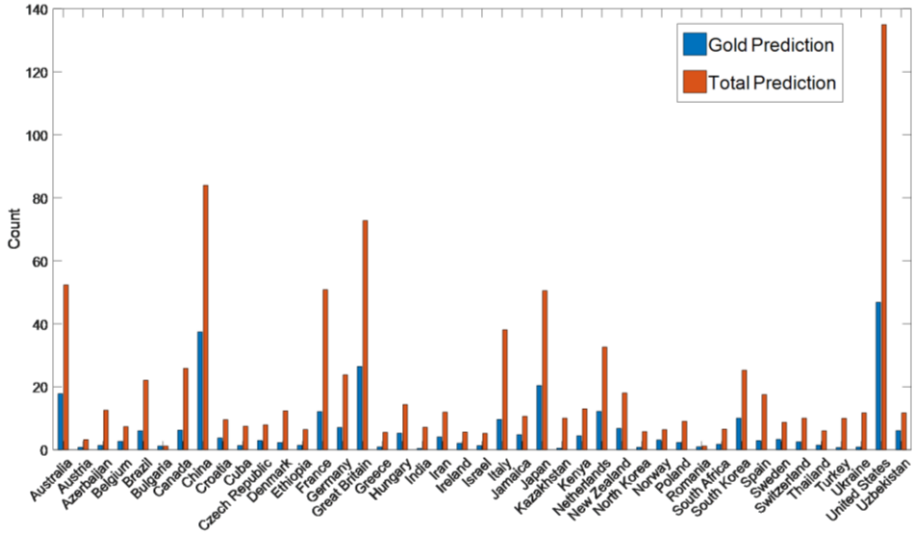

Figure 9. The Prediction of Gold Medals and Total Medals

3.3. Prediction of the Breakthrough from "0" to "1"

3.3.1. Results

The posterior probability of 20% or higher is set as the threshold for the possibility of winning the first medal. The model indicates that there are three countries most likely to win their first medal in the 2028 Olympic Games, as shown in the Table III.

Table 3. The probability of the countries that will win the first medal in the 2028 Olympic Games

NOC	Probability
South Sudan	0.267
Nepal	0.233
Bolivia	0.206

3.3.2. Bayesian Probability Calculation

The posterior probabilities for countries without prior medals were computed using formula (2). The likelihood function P(B|A) incorporated three normalized factors:

- Athletes' ability (x_1)

- Economic level (x_2)

- Sports infrastructure (x_3)

The joint likelihood was defined as:

$$P(B \mid A) = 0.45x_1 + 0.35x_2 + 0.25x_3 \tag{6}$$

The reason why South Sudan has such a high posterior probability is that it has recently invested heavily in sports infrastructure and its national economy may continue to grow in the future, so x_2 and x_3 will be larger in the range, so the posterior probability will increase.

3.4. Analysis of the Medal Table

Based on the predicted results of the top ten medal table, we conducted a detailed analysis of the potential progress or regress of several countries in comparison to the 2024 Olympic Games.

3.4.1. Error Analysis

The prediction error of the fusion model for the traditional sports powers such as the United States and China is mainly due to the changes of potential external factors, which are difficult to be accurately embedded in the model, such as the change of the advantages of the host country, the change of the competitive level of athletes or coaches, and the investment of the country in different sports. The prediction error for emerging countries such as Azerbaijan is mainly due to the sparse historical data of new participants, which makes it difficult for the fusion model to adapt to fewer dynamic time series, which emphasizes the necessity of strengthening data collection in future work.

3.4.2. The Progress Country' Analysis

Firstly, the United States is most likely to make progress, as the 2028 Olympic Games will be held in Los Angeles, USA. The United States has the advantage of being the host country and they have added some events in which they excel, such as squash, baseball, lacrosse and flag football. They may remove some events in which they are not as proficient, such as weightlifting, boxing and modern pentathlon. Therefore, we predict that the United States will achieve significant progress, thereby dominating the gold medal and overall medal table.

Secondly, another country that may achieve progress is Great Britain, as there are new events in the next Olympic Games in which the Great Britain excels, such as lacrosse and flag football. Therefore, we predict that the Great Britain will also make certain advancements.

3.4.3. The Regress Country' Analysis

Firstly, France is most likely to regress and to regress the most, as the 2024 Olympic Games was held in Paris, France, giving France the advantage of being the host nation. At that time, they include many events in which they excel, such as breakdancing, climbing, skateboarding and surfing. They achieve excellent results in these events. Therefore, France achieve its best performance in the history of the Olympics in 2024. However, in the 2028 Olympic Games, France will no longer have the advantage of being the host nation and will be unable to include its advantageous events. Thus, we predict that France is most likely to regress and to regress the most.

Secondly, it is possible that China may experience a decline, as some events in which China excels, such as weightlifting, may be removed in the next session. This will affect China's ability to secure gold and other medals. However, considering China's rapid development in recent years in areas such as economy, technology, and military, along with the emphasis on becoming a strong sporting nation, a series of policies have been introduced to motivate athletes to achieve good results. Additionally, Chinese athletes are known for their resilience, fearlessness in the face of difficulties, and strong patriotic spirit, which often leads them to perform well in events where they do not have a competitive advantage. Therefore, we predict that China will experience a decline, but it will not be significant.

Additionally, it is also possible that the German may regress. In recent years, the performance of the German in the Olympic Games has been declining, which is due to the heavy burden on German coaches, low salaries, limited financial resources for sports programs in schools, and a lack of emerging athletes. Following this trend, it will be difficult for Germany to achieve significant improvement in the next four years. Therefore, we predict that the German will continue to decline in the 2028 Olympic Games.

3.5. The Impact of the Advantage Events on Countries

According to the data, each country have different advantageous events and they often achieve better results in their strong events, almost guaranteeing medal wins. The advantageous events of each country ensure a baseline number of medals, including gold medals and these events are typically the ones most valued by the people of the respective countries(As shown in the Table IV). Often, these events can attract the attention of the audience, instilling great confidence in their countries and raising expectations. Our model also predicts that the advantageous events of each country will consistently win medals, including gold medals, in the 2028 Olympic Games. However, the

performance of these different countries in the Olympic Games is influenced by various factors, including historical and cultural backgrounds, sports development and resource investment, geographical environment, technical expertise and the support of the government. For example, China's advantage in table tennis stems from its profound table tennis culture and strong training system, having almost dominated the field since 1988. Japan's outstanding performance in judo is attributed to the cultural foundation of judo as a traditional sport, especially since the 1964 Tokyo Olympic Games, where Japanese athletes have almost continuously won gold medals.

Table 4. The Achievement of the Advantage Project for Each Country

Conutry	Advantage projects	Number of gold medals won	Number of medals won
United States	Track and field, swimming	233	573
China	Table tennis, diving	70	111
Germany	Fencing, sailing	64	121
Japan	Judo, swimming	29	70
Kenya	long-distance race	15	36
Australia	swimming	63	177

The strong performance of the United States in swimming and athletics is attributed to its systematic training system, substantial resource investment and diverse talent pool. Especially in swimming, athletes such as Michael Phelps have led the United States to historic achievements. Kenya's outstanding performance in long-distance running is related to its unique geographical environment and high-altitude training, which has cultivated a large number of top marathon and middle-distance runners.

The success of Germany in fencing and sailing events can be attributed to its long-standing sports tradition and well-established training system, which these countries typically enhance through support from the government and sports organizations to concentrate efforts on improving the competitiveness of certain disciplines.

In summary, the Olympic performances of various countries are not only related to the individual abilities of the athletes but are also closely linked to the cultural background of the nation, resource investment, and a systematic training framework.

4. Conclusions

In this study, this study successfully built a sports performance prediction model based on the fusion of improved ARIMA and neural network. In the data processing stage, the data quality is improved through data cleaning and standardization, which lays a good foundation for model training. The improved ARIMA model combined with external factors can effectively capture the time series characteristics of sports performance data. Bayes theory provides strong support for the prediction of special cases. The neural network further optimizes the prediction results and improves the accuracy and reliability of the prediction. The experimental results show that the fusion model has significant advantages in sports performance prediction, and can predict athletes' sports performance more accurately than the traditional method and the single model.

Future studies can further optimize the model parameters and explore more reasonable external factor selection and weight allocation methods to improve the model performance. Attempts are made to introduce more advanced neural network architectures such as recurrent neural networks (RNN) and their variants LSTM, GRU, etc., to better handle long-term dependencies in motion data. In addition, the application scope of the model is expanded to conduct more in-depth prediction research on sports performance in different types of sports events and different training scenarios, so as to provide more

comprehensive and accurate prediction technical support for the field of sports science and promote the continuous development of sports science.

References

[1] Wang J, Liu Y, Li Y. A parallel differential learning ensemble framework based on enhanced feature extraction and anti-information leakage mechanism for ultra-short-term wind speed forecast[J].Applied Energy,2024,361122909-.

[2] He N ,Yang Z ,Qian C , et al.Remaining useful life prediction of lithium-ion battery based on fusion model considering capacity regeneration phenomenon[J].Journal of Energy Storage,2024,85111068-.

[3] Grasa R P, Rodriguez F R, Novelli G, et al. Satellite image classification with neural quantum kernels[J]. Machine Learning: Science and Technology,2025,6(1):015043-015043.

[4] Hashemi A, Izadkhah A. A graph neural network simulation of dispersed systems[J].Machine Learning: Science and Technology,2025,6(1):015044-015044.

[5] Wang T, Tan X, Tian Y, et al. Risk assessment based on Bayesian Network for the typhoon-storm surge-flood-dike burst disaster chain: A case study of Guangdong, China[J]. Journal of Hydrology: Regional Studies,2025,58102251-102251.

[6] Zheng R ,Yang B ,Qian Y , et al.Joint SOH and RUL estimation for lithium-ion batteries via optimal deep belief network with Bayesian algorithm[J].Journal of Energy Storage,2025,114(PB):115891-115891.

[7] Wang Y, Yan L, Zhou T. Deep learning-enhanced reduced-order ensemble Kalman filter for efficient Bayesian data assimilation of parametric PDEs[J]. Computer Physics Communications,2025,311109544-109544.

[8] Sepehry N ,Ehsani M ,Amindavar H , et al.A novel enhanced Superlet Synchroextracting transform ensemble learning for structural health monitoring using nonlinear wave modulation[J].Engineering Applications of Artificial Intelligence,2025,147110341-110341.

[9] Shen S, Cheng J, Liu Z, et al. Bayesian inference-assisted reliability analysis framework for robotic motion systems in future factories[J]. Reliability Engineering and System Safety,2025,258110894-110894.

[10] Zhang Y, Liu L, Qiao Q, et al. A Lie group Laplacian Support Vector Machine for semi-supervised learning[J]. Neurocomputing,2025,630129728-129728.