Proceedings of the 7th International Conference on Education Innovation and Philosophical Inquiries
DOI: 10.54254/2753-7048/2026.35640

Prediction of Olympic Gold Medalists Based on Machine Learning: A Comparative Study of Classification Models

Mingze Qin

College of Education and Human Development, University of Minnesota Twin Cities, Minneapolis, USA

daqingminghang1@163.com

Abstract. The main purpose of this study is to explore the factors that contribute to professional athletes winning gold medals at the Olympic Games. Given the global attention paid to the achievements of previous Olympic Games and the national focus on the cultivation of Olympic athletes and their medal situations, it is of significant analytical value to understand the differences between Olympic gold medalists and other medalists. This study will utilize the 120 Years of Olympic Athletes Dataset and obtain the basic data of Olympic athletes from it to construct a prediction model and identify the key factors influencing gold medal awards. The research structure indicates that the optimal physiological condition is usually associated with the optimal age group, and the optimal physiological condition will have a crucial impact on the competitive level of top athletes. Although machine learning models can effectively identify the success factors of top athletes, the absence of some athlete features and potential sample bias make the overall situation limited. Overall, this study demonstrates that predictive models can provide reliable predictive capabilities for sports analysis and help understand measurable athlete characteristics related to medal competition in the Olympics and other different events.

Keywords: Olympic gold medal prediction, machine learning, physiological characteristics, random forest

1. Introduction

As the highest stage of international sports events, the Olympic Games gather top athletes from all over the world and attract the attention of global audiences. However, among all the Olympic medalists, gold medalists usually represent the highest achievements and standards in their respective sports. Therefore, understanding the characteristic differences between gold medalists and other medalists has extremely high research value.

An analysis of previous research reveals that few studies have utilized machine learning methods to predict whether Olympic athletes can win gold medals by comparing their physical characteristics.

Therefore, the purposes of this study are as follows:

1. Construct a prediction model for Olympic gold medals by leveraging the characteristics of athletes

DOI: 10.54254/2753-7048/2026.35640

2. Identify which athlete characteristics have a positive impact on the competition for gold medals.

2. Literature review

Previous studies on Olympic performance prediction mainly focused on the medal-winning situation at the national level. For instance, based on the data from the Summer Olympics from 1896 to 2024, a random forest regression model was constructed. By integrating variables such as historical performance, the advantages of the host country, and the strength of athletes, the medal table for the 2028 Los Angeles Olympics was predicted [1]. There are also studies based on psychology that compare the differences between gold medalists and other top athletes to determine which psychological traits and intrinsic motivations can have a positive impact on winning gold medals [2]. In conclusion, existing research generally focuses on medal prediction at the national level and the prediction of gold medalists from a psychological perspective. There are relatively few studies that predict gold medals from the perspective of individual athletes using multiple machine learning models. The contribution of this study lies in that based on the individual data differences of athletes, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting were used for prediction and comparison to identify the important characteristic factors affecting athletes' gold medal wins. This will fill the research gap in gold medal prediction at the individual level.

3. Data and methodology

This study uses the "120 Years of Olympic Athletes" dataset from Kaggle. This dataset contains data information of a total of 277,116 participating athletes from Athens 1896 to Rio 2016, including demographic variables (age, gender, nationality) and physical characteristics (height, weight), as well as corresponding sports events and competition years.

Since the objective of this study is to identify the relevant factors of the athletes who won gold medals, we chose to remove the athletes who did not win medals from the dataset and code the gold medalists as 1, and the silver and bronze medalists as 0, in order to improve the model's prediction of the possibility of the winning athletes winning gold medals, rather than distinguishing all the Olympic participants.

Before building the predictive model, to ensure the consistency of the data, we preprocessed the data. Remove some missing data and useless variables such as ID, year and city from the dataset. Retain variables such as age, height, weight, and gender, as they represent the human characteristics related to the award-winning athletes. After data cleaning and preprocessing, the final sample for model training and testing consists of 39,783 medalists, rather than the original dataset of 277,116 Olympic athletes.

During the model construction phase, the dataset is randomly divided into a 80% training set and a 20% test set. Four supervised classification models, namely logistic regression, decision tree, random forest, and gradient boosting, were constructed. The model performance was evaluated by Accuracy, Precision, Recall, and F1-score, while the AUC value was calculated based on the discriminative ability of the latest model.

DOI: 10.54254/2753-7048/2026.35640

4. Results

4.1. Model performance comparison

Table 1. Model performance summary

--Model Performance Summary--					
	Model	Accuracy	Precision	Recall	F1-score
0	Logistic Regression	0.5968	0.4283	0.5870	0.4952
1	Decision Tree	0.6473	0.4782	0.5123	0.4947
2	Random Forest	0.6949	0.5575	0.4577	0.5027
3	Gradient Boosting	0.6927	0.7439	0.1342	0.2274

As shown in Table 1, the performance parameters of four classification models. Random forest achieved the highest accuracy rate (0.6949), and Gradient Boosting ranked second (0.6927). It is noted that the accuracy rates of regression and decision trees are relatively low, being 0.5968 and 0.6473, respectively.

In terms of performance, Random Forest has the highest F1-score (0.5027) , which reflects that it is most balanced between Precision (0.5575) and Recall (0.4577). Although Gradient Boosting achieved the highest Precision (0.7439), its Recall (0.1342) was too low. The lower Recall rate indicates that Gradient Boosting is overly conservative in predicting gold medalists. Although it reduces false alarms for gold medalists, it will also cause the model to miss many true gold medalists. The Recall of the Logistic regression algorithm is relatively high (0.5870) , but its Precision (0.4283) makes its overall performance unable to match that of random forests.

Taking into account Accuracy, Precision, Recall, and F1-score comprehensively, the random forest has the most balanced and stable performance, and thus was selected as the final model.

4.2. Cross-validation and model stability

To evaluate the stability of the model, we conducted 5-fold cross-validation on the random forest model. The final validation accuracy was 0.6917, proving that this model has stable prediction performance in different datasets.

4.3. Determinants of gold medal attainment

Based on the analysis of athletes' characteristics by the random forest model, age, weight, and height are the three most important factors affecting the acquisition of gold medals. As illustrated in Figure 1, the importance of age accounted for the highest proportion (0.2126), followed by weight (0.2091) and height (0.1953). Gender, team, and specific sports events have a relatively small impact on the prediction results.

DOI: 10.54254/2753-7048/2026.35640

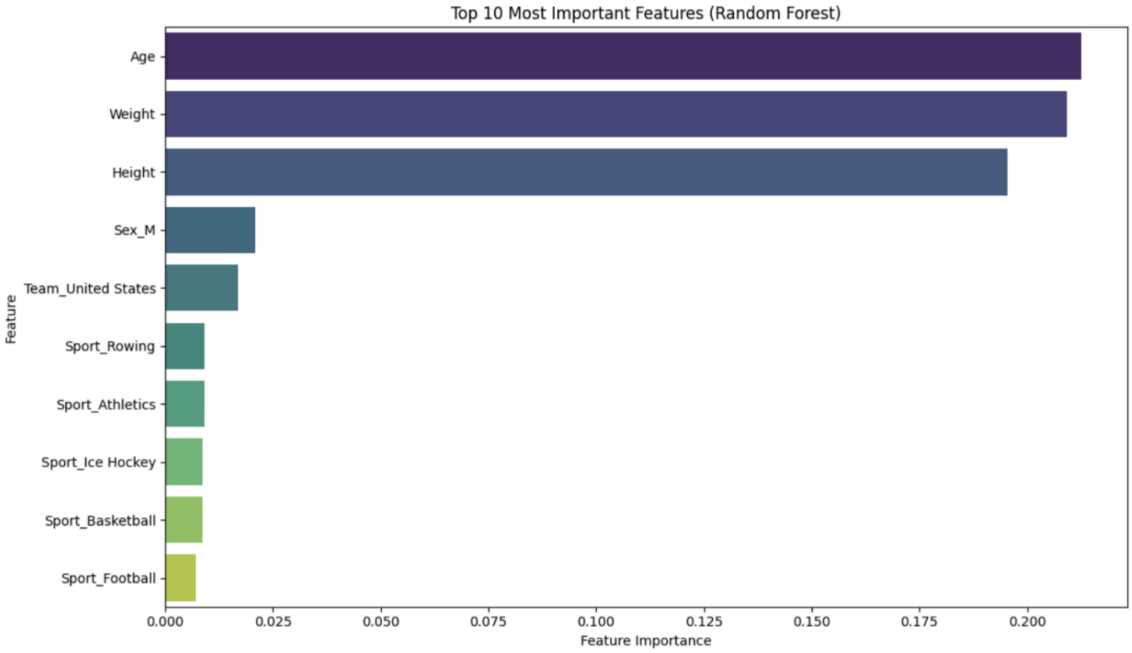

Figure 1. Top 10 most important features (random forest)

The AUC score of this model is 0.7051, which means that the model has a 70.51% probability of correctly distinguishing gold medalists. This also indicates that although the model has sufficient discrimination ability, it still has certain deficiencies to some extent.

5. Discussion

Through this study, it was found that age, height, and weight are three important predictive variables related to Olympic gold medalists. These physiological characteristic variables are closely related to the prediction of gold medals in the model, indicating that these three measurable physical features play a very important role in predicting gold medalists.

The age of an athlete is the most important variable factor in the prediction of medals. The physical fitness and performance of an athlete tend to peak at a certain age in competitive sports [3]. This means that young athletes generally have superior physical fitness and sport-specific competencies than their older counterparts. Furthermore, young athletes have a superior post-competition recovery, which enables them to engage in high-intensity and high-frequency near bouts in order to enhance their performance. However, young athletes lack some of the matured psychological qualities as well as the tactical understanding and experience of competition that older athletes possess.

It is worth noting, however, that the dataset of this study includes the winners of all the Olympic Games from 1896 to 2016. Over the past 120 years, the training methods, dietary nutrition, sports science, and competition standards of top athletes have undergone tremendous changes. For instance, modern athletes are more likely to receive professional and scientific training methods and dietary combinations, while early Olympic athletes found it difficult to receive such systematic training and protection [4]. Therefore, the peak age of athletes may vary in different years. In this study, the findings related to the age of athletes reflect the comprehensive results from 1896 to 2016. Secondly, the concept of "optimal age" does not necessarily apply to all athletes. There are physiological differences between male and female athletes, and their peak ages may vary.

Moreover, different sports have different physical requirements for athletes. For instance, some sports that require high explosive power may be more suitable for athletes of different age groups, while those that emphasize endurance are not. Therefore, the findings regarding age advantage in this study may vary by gender and competition event. In the future, we will focus on the recent Olympic athlete datasets to reduce the differences in diet, training and competition standards among athletes in competitive sports due to the huge time span. Meanwhile, the model is subdivided by gender to observe whether there is a difference in the importance of age for prediction.

It is obvious that the height and weight of Olympic athletes also have a relatively significant predictive ability. These human characteristics are closely related to physical strength and biomechanical efficiency. For instance, a taller athlete usually has the ability to utilize their arm span to achieve greater effects and cover a greater distance with each step during running. In addition, athletes with a heavier body weight usually generate more force in each movement they complete, which is particularly important in weightlifting and many other sports that require athletes' explosive power [5]. Therefore, an athlete's physical fitness may play a huge role in their ability to perform at the highest level of competition.

Although the overall impact of sports events in the comprehensive model is relatively small, the physical requirements for different sports events vary. For instance, marathon runners need a lighter weight and to match it with strong endurance, while weightlifters require powerful muscles and explosive power [5]. If all sports events are combined and added to the prediction model, it may weaken the feature recognition of the physiological characteristics of specific events. In future research, we will subdivide athletes from different sports, such as endurance and strength types, or make predictions for individual sports to test the significance of height and weight for predictions.

The comparison results of classification models show that the random forest model performs the most evenly and stably. Unlike logistic regression, which can only handle simple linear relationships, random forest can deal with more complex variable relationships. Compared with decision trees, random forests reduce risks by combining multiple trees. Therefore, for the situation where sports performance is affected by multiple factors, compared with other models, the more flexible random forest is more suitable for capturing information and reflecting the patterns therein, thereby achieving the best prediction effect.

Although the Gradient Boosting model achieved the highest precision rate (0.7439), its recall rate (0.1342) was too low. This indicates that the model's ability to predict gold medalists is overly conservative. In other words, the Gradient Boosting model will determine an athlete as a gold medalist only when it is absolutely certain of the prediction result. Although this ability can significantly reduce false positives for gold medalists, it will also cause the model to ignore many true gold medalists. Therefore, accuracy and recall highlight the significance of multiple metrics in evaluating model performance, reducing the performance imbalance caused by a superior score in a single metric while others are too low.

Based on the above discussion, it is found that there are some limitations in this study. Firstly, the dataset adopted by this prediction model lacks some important variables related to competitive sports, including the training patterns of athletes, the coaching level of athletes' coaches, the psychological quality of athletes, and the support level of the athletes' countries for them (funds, training equipment, training venues), etc. All the above factors are key to whether an athlete can win a gold medal at the Olympic Games. In future research, the methods of this study can be adopted and more variables related to the research objectives can be included. Secondly, different sports all have unique physical characteristics, and there are also differences in the peak age groups [3]. Therefore, future research can conduct different model predictions for different Olympic events to

DOI: 10.54254/2753-7048/2026.35640

achieve more accurate prediction results and better fit the characteristics of different Olympic events.

6. Conclusion

This research shows that machine learning models can help identify meaningful patterns related to Olympic gold medalists. In this test of the model, the random forest model provided us with the most balanced and stable prediction performance. Age, height, and weight, these three measurable physical parameters, have been identified as the most important predictor variables, indicating that predictable physical characteristics are closely related to gold medal results.

Through this research, it is found that athlete coaches, various sports organizations around the world, and athlete training institutions will gain beneficial insights from it, as the results identify measurable physical characteristics of athletes directly related to gold medalists, which will help these industry stakeholders provide assistance and make decisions in athlete development, talent selection, and other work.

However, this model still has limitations because we did not incorporate factors such as the training quality of each athlete, past injury history, coaching experience of coaches, psychological characteristics of athletes, and even the degree of support from the state for athletes into the model dataset. But all these factors may have a huge impact on the performance of athletes. In the future, we will incorporate more variables and new data, which can further enhance the accuracy of predictions and help us gain a deeper understanding of the performance of top athletes.

References

[1] Mehta, K. (2024, August 1). Here is what separates gold medalists from the rest. *Forbes*. https://www.forbes.com/sites/kmehta/2024/07/30/here-is-what-separates-gold-medalists-from-the-rest/

[2] Wang, Y. (2025). Research on Olympic medal prediction using random forest regression model. *CAMMIC '25: Proceedings of the 2025 5th International Conference on Applied Mathematics, Modelling and Intelligent Computing*, 231–236. https://doi.org/10.1145/3745533.3745572

[3] Allen, S. V., & Hopkins, W. G. (2015). Age of peak competitive performance of elite athletes: A systematic review. *Sports Medicine (Auckland, N.Z.), 45*(10), 1431–1441. https://doi.org/10.1007/s40279-015-0354-3

[4] Haake, S. J. (2009). The impact of technology on sporting performance in Olympic sports. *Journal of Sports Sciences, 27*(13), 1421–1431. https://doi.org/10.1080/02640410903062019

[5] Nuryadin, I., Doewes, R. I., Adi, P. W., & Wijanarko, B. (2023, May 22). Performance characteristics of male athletes in 35 different sports. *Riped-Online*. https://www.riped-online.com/articles/performance-characteristics-of-male-athletes-in-35-different-sports-99074.html#:~:text=The%20distinguishing%20characteristics%20are%20briefly,leg%20muscle%20power%20for%20basketball