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

Prediction of sports strength based on linear regression and random forest model

Yipeng Zhang [1, #], Zhuoyu Li [2, #], Haosheng He [1, *, #]

[1] Institute of Fintech, Shenzhen University, Shenzhen, China

[2] College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China

* Corresponding Author Email: 2023363020@email.szu.edu.com

[#]These authors contributed equally.

Abstract. This paper constructs a dynamic weighted comprehensive model that integrates linear regression and random forest algorithms. This model is used to predict the distribution of competitive performance indicators in international comprehensive sports events and analyze the relevant influencing factors. The model innovatively integrates multiple factors. It incorporates historical data and uses a mechanism to assign higher weights to recent performances, so as to better reflect the current situation. The model also takes into account geographical advantages, which may affect a country's performance in certain sports. Additionally, it includes the effectiveness of training strategy decision - making entities and the scale of participants. In the quantification process, the model first uses linear regression to predict the number of entities that are likely to achieve the best results for the first time. Then, it quantifies the competitiveness index of each project. Specifically, competitive achievements are converted into scores, and a dynamic weighting method is adopted to adjust the scores according to the time when the achievements are obtained, with more emphasis on recent achievements. Finally, the random forest algorithm is used to comprehensively consider various factors and predict the final results of participating units. This model has high application feasibility. Its unique dynamic weighting mechanism and multi - factor integration make its predictions more accurate. The model also predicts potential new - achiever units and evaluates prediction uncertainty. Future research could incorporate real - time individual data and socio-economic variables to enhance the model.

Keywords: Linear Regression; Random Forest Model; Dynamic Weighted Model; Exponential Decay Coefficient.

1. Introduction

Previous studies have extensively explored the prediction of performance - related metrics and the analysis of influencing factors, with a particular emphasis on leveraging various computer science techniques. In the realm of machine learning and statistical models, scholars like Maulud D, Abdulazeez A M [1], and Martin P [2] have studied linear regression, highlighting its significance in data analysis and prediction. Li Y and Mu Y [3] investigated a random forest - based feature selection algorithm in performance - related evaluation, demonstrating the potential of random forest algorithms in identifying important features. Gao Z and Kowalczyk A [4] used a random forest model to find key predictors in specific outcome predictions, emphasizing its practical use in analytics. Fernando K R M and Tsokos C P [5] proposed a dynamically weighted balanced loss approach, which might be useful for performance - related prediction considering data imbalance. Ahmad S et al. [6] introduced a homotopy perturbation method for complex equations, providing a basis for algorithm development in performance - related prediction. Christoph A H et al. [8] applied machine learning to predict performance - related changes over time, while Zhonghui T et al. [9] used a random forest algorithm in a certain application area, inspiring similar applications in performance - related analysis. Jiandong H et al. [10] used algorithms to predict the properties of materials, and the idea can be adapted for performance - related prediction. However, in performance - related prediction, existing studies have certain limitations. Many focus on single or a few influencing factors, and prediction

models lack comprehensiveness and accuracy. There is no complete system integrating multiple key factors like historical data, the effectiveness of decision - making entities in a certain strategy, and the scale of participants for accurate performance - related predictions.

Against this backdrop, this research constructs a comprehensive quantitative analysis model. It combines variables such as historical data, a factor related to a certain beneficial condition, the effectiveness of decision - making entities in a certain strategy, and the scale of participants to build a dynamic weighted prediction model. This model predicts new - achieving entities, quantifies the competitiveness index of each item, and forecasts overall performance. It also analyzes the impact of decision - making entity effects on performance - related indicators, explores indicator distribution characteristics, analyzes resource - allocation constraints, uncovers participating - unit contribution patterns, and assesses the influence of organizing - entity item - setting decisions, aiming to support strategic resource allocation among participating units.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing

The data utilized in this study were primarily sourced from historical records of past competitive sports events, encompassing multiple seasons and containing detailed information such as performance outcomes of participating entities across various events, the number of events entered, and participant-related data, including gender, age, etc. (https://www.olympics.com/en/olympic-games/paris-2024/medals).

Following data acquisition, a series of preprocessing steps were undertaken. First, the dataset was cleaned to remove outliers and anomalies, thereby ensuring completeness and accuracy. Subsequently, performance results across different events were standardized through a unified scoring scheme. In this process, the concept of a Competitive Achievement Index was introduced. A hierarchical quantification framework was established based on varying levels of performance, resulting in a consistent point-based system to support subsequent modeling and comparative analyses.

2.2. Methodological Framework

Objective 1: This study aims to predict the medal standings for the 2028 Global Sporting Event and to estimate the number of teams likely to achieve competitive success for the first time in future competition cycles. A regression analysis framework is constructed to facilitate these predictions. This model employs a dynamic weighting mechanism to evaluate the performance capabilities of different teams across various subcategories. Additionally, ensemble learning methods such as Random Forest are utilized to forecast future trends and to analyze the correlation between participating entities and competitive performance.

Objective 2: The study identifies teams with substantial performance fluctuations across historical records. A modeling approach is developed considering variables such as the rate of performance change, geographical advantage factors, and accumulated historical experience in specific disciplines. By introducing a control variable termed the "Training Strategy Decision Agent," the study evaluates its impact on performance volatility and assesses its explanatory power based on historical statistical means.

Objective 3: This study delves into the computational modeling of competitive achievement indices, focusing on the algorithmic analysis of performance distributions across various subcategories and geographical factors within international multi-sport events. The objective is to employ advanced correlation analysis techniques to explore the intricate relationships between subcategory achievements and regional characteristics. The model leverages machine learning algorithms, particularly ensemble methods like Random Forest, to uncover patterns and dependencies that may influence competitive outcomes. This approach not only enhances the understanding of how different regions and subcategories contribute to overall competitive performance but also provides a

computational framework for predicting and optimizing performance in future events. The study aims to demonstrate the utility of computational models in sports analytics.

A prominent feature of the proposed model is the incorporation of a historical data attenuation mechanism, which assigns higher weights to more recent historical performances. This enhances the model's responsiveness and real-world applicability. Moreover, the evaluation system is constructed based on participating entities rather than individual event outcomes, thereby reducing the influence of anomalous single-event results and ensuring a more stable and reliable comprehensive assessment.

To verify the predictive capability and adaptability of the model, a series of quantitative metrics are established. Sensitivity analysis is incorporated to examine the model's robustness under parameter perturbations (As shown in Fig. 1). Specifically, during the parameter setting phase, sensitivity analyses are conducted for core variables such as the historical weight control factor and the scoring system structure. Multiple weighting functions and scoring schemes are implemented in both control and experimental groups. The variations in model outputs are then compared to identify variable combinations that significantly impact prediction results, guiding the selection of the optimal configuration.

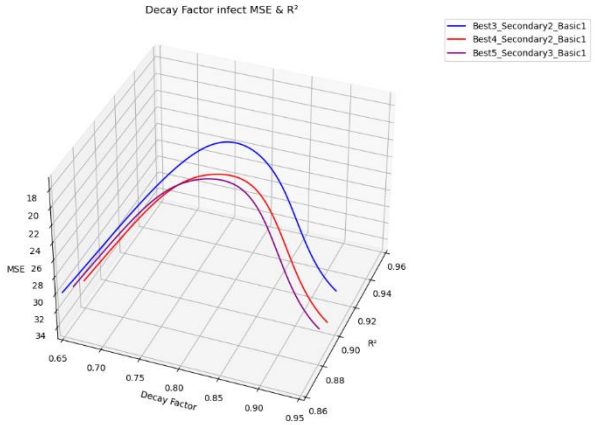

Figure 1. Decay Factor infect MSE and R^2

It is worth noting that although this study cannot precisely predict the specific countries (or regions) that will emerge as new competitive entities, the model employs parameterized techniques to reasonably estimate their perturbation effects on the overall competitive structure.

3. Analysis of the Sports Competitiveness Model

3.1. Model Construction and Analysis for Predicting Competitive Achievement Indicators

This section constructs a model framework for predicting relevant performance indicators of various countries in international comprehensive sports events, focusing on three key dimensions: predictive analysis of new award-winning entities, quantitative assessment of project competitiveness, and predictive deduction of final outcomes. The model leverages an algorithmic framework with multidimensional feature processing and high-efficiency predictive capabilities to enable modeling analysis of complex sports event data—by integrating historical performance data of each entity and dynamic project indicators, and abstracting specific event types and indicator categories, it enhances the algorithm's capturing capability for nonlinear relationships, thereby providing systematic solutions for performance forecasting of different entities in various projects.

3.1.1. Prediction of the first award-winning entity

Through the linear regression model, predict the number of subjects that may achieve the optimal competitive achievement for the first time in the future. The formula is as follows and symbol interpretation is shown in Table I:

$$N_f = \beta_0 + \beta_1 H + \beta_2 E_f + \beta_3 \bar{P} + \varepsilon \quad (1)$$

Table 1. Formula (1) Symbol Interpretation

N_f	the number of subjects that will newly obtain the optimal competitive achievements in the future
β_0	A constant term representing the basic number of subjects achieving competitive accomplishments
β_1	The contribution of geographical factors to the increase in the number of new successful bidders
β_2	The impact of future project distribution in various countries on the number of newly awarded winners
E_f	Number of future projects in each country
β_3	The impact of historical growth on forecasting
H	Take 0 or 1
P	Historical average fluctuation

3.1.2. Quantification of project competitiveness index

Convert the types of competitive achievements into corresponding scores for numerical analysis. The corresponding relationship is: The best competitive achievement is worth 3 points, the secondary competitive achievement is worth 2 points, and the basic competitive achievement is worth 1 point.

The formula for calculating dynamic weighted competitive achievements is:

$$W = \sum \lambda^{T-Y} \times S \quad (2)$$

Among them, W is the score adjusted based on competitive achievement rewards, reflecting the potential strength of the individual participants' performance. The exponential attenuation coefficient λ, where λ∈ (0,1), is a factor that reduces the weight of early competitive achievements over time, indicating that the value of early competitive achievements is relatively low. T is the target year and Y is the year when competitive achievements are made. S is determined based on the type of competitive achievement for the current year.

By analyzing the proportion of those who achieve competitive achievements for the first time among all subjects through linear regression, the number of subjects that will achieve competitive achievement indicators for the first time in international comprehensive sports events in the future is predicted. The competitive ability index of each subject in each competition event is quantified by using the dynamically adjusted scoring formula, and each achievement corresponds to a specific score. The competitive ability index of each entity in each competition project is processed by using the dynamic adjustment model and attenuation coefficient to reduce the influence of early data and calculate the expected competitive ability index of the entity in the future project (the higher the competitive ability index of the project, the stronger the competitive ability).

3.1.3. Final result prediction

Based on the number of relevant indicators for each project and the predicted values of relevant indices for each entity in the future, by leveraging the random forest—a model algorithm with excellent fitting and predictive capabilities—this study obtains the prediction results of relevant

performance indicators for each entity in each project in the future. In the analysis of the prediction results of competitive achievement indicators for future events generated by the dynamic weighting model, linear regression model and random forest model, it is found that some countries will still maintain a leading edge in competitive achievements, while the rankings of some countries have changed.

The dynamic weighting model, which assigns weights to historical medals using a decay factor of 0.85 to reflect the timeliness of data, shows that in terms of the distribution of competitive achievement indicators of different sports, some sports have stable and strong performances in optimal (gold), secondary (silver) and basic (bronze) competitive achievements. This stability is verified by the consistent high scores in the weighted medal calculation across multiple Olympic cycles. For other sports, the distribution of competitive achievement indicators shows diversity: the random forest model, which predicts specific awards based on predicted capability values and medal quotas, highlights that some sports stand out in the number of optimal competitive achievements due to high capability values, while others perform better in secondary or basic competitive achievements, reflecting different strategic focuses of countries.

Linear regression analysis of the total number of competitive achievement indicators among non-dominant countries reveals that their scores are relatively close, with a competition intensity index indicating fierce competition. This is consistent with the model's prediction that in the middle and upper reaches of the competitive achievement indicator ranking.

3.2. Model building idea

This study systematically analyzes the competitive achievement index score data of participating entities across ten international comprehensive sports events, integrating algorithmic methodologies to enhance technical rigor for empirical research.

First, the data pipeline commences with automated preprocessing: raw datasets undergo missing value imputation (using k-nearest neighbors for continuous variables) and outlier detection via Z-score standardization, ensuring data integrity. A custom Python script implements feature scaling (Z-score normalization) to harmonize metrics across heterogeneous sports categories, leveraging NumPy for vectorized computations to optimize efficiency. Subsequently, performance stability is quantified by calculating the sample standard deviation of competitive scores per sport, serving as a key indicator of score dispersion. High standard deviation signals volatility, prompting further causal analysis. In project selection, the top five entities with the highest stability indices (lowest score variability) are filtered using a k-sorting algorithm, isolating projects for multivariate linear regression modeling. The model specifies the competitive achievement index as the dependent variable, with independent variables including binary-coded training strategy decision-maker changes (1=transition, 0=continuity), dummy-coded geographic home advantage (host vs. non-host), and log-transformed participant counts to address heteroscedasticity. Scikit-learn's Linear Regression module is employed, with Standard Scaler ensuring β coefficients represent standardized effect sizes for interpretability. Ratio analyses are conducted via element-wise division to compute competitive score-to-event ratios, facilitating cross-sport comparisons. Historical score impacts are modeled through lagged regression, incorporating t-1 period ratios as predictors to capture temporal dependencies. Decision-maker change effects are evaluated by contrasting score trajectories before/after leadership transitions, while geographic gain factors are tested using two-sample t-tests between host and non-host nations. Participant count influences are estimated via log-log regression, yielding semi-elasticity coefficients to characterize resource allocation impacts. Crucially, the analytical framework includes model validation: 10-fold cross-validation assesses predictive reliability, and variance inflation factor (VIF) calculations mitigate multicollinearity. β coefficients from the regression model quantify the weighted influence of training strategy determinants, with statistical significance determined via t-tests ($p < 0.05$). All procedures are scripted in a reproducible Jupyter Notebook, adhering to FAIR data principles for transparency and replicability. By embedding algorithmic precision—through

standardized preprocessing pipelines, rigorous statistical modeling, and systematic validation—this study provides a methodologically robust framework for sports performance analysis.

Data preparation First, carry out data preparation. This dataset covers information over many years and contains the following columns:

Ath_X: The number of individuals participating in an international comprehensive sports event of year X,MS_X: In year X, the best competitive achievement is worth 3 points, the secondary competitive achievement is worth 2 points, and the basic competitive achievement is worth 1 point (taking the acquisition of the basic competitive achievement as an example),C_X: The number of training strategy decision-making entities participating in the international comprehensive sports event of year X,E_C_X: The number of items in year X,H_X: Whether the country was the host for year .

In each year's modeling, the goal is to predict the medal score for that year using these features.

Feature selection and Target variable: For each year (such as 1992, 1996, etc.), construct the feature matrix X and the target variable y based on the following features:

Feature matrix X: For 1992, the training strategy was adopted to determine the number of subjects (C _1992), Competitive achievement indicators (MS_1988), Project quantity (E_C_1992). The number of participating individuals (Ath_1992) And the geographical location gain factor(H_1992) as feature.

For 1996, the training strategy was adopted to determine the number of subjects (C_1996)、Competitive achievement indicators (MS_1992_chu)、Project quantity (E_C_1996)、The number of participating individuals (Ath_1996) And the geographical location gain factor (H_1996) as feature. Similarly, for each year, the model takes the competitive achievement indicators of the previous year as one of the input features.

Target variable y: The target variable is the annual competitive achievement indicator, such as the competitive achievement indicator for 1992, the competitive achievement indicator for 1996, etc. The basic idea of linear regression model construction is to use a set of features (independent variables) to predict a continuous target variable. For each year X_t, a linear regression model of the following form is constructed:

$$S_t = \beta_0 + \beta_1 \cdot C_t + \beta_2 \cdot S_{t-4} + \beta_3 \cdot E_{C_t} + \beta_4 \cdot A_t + \beta_5 \cdot H_t \quad (3)$$

S_t is the target variable (the competitive achievement indicator of year t). β_0 is the intercept term, representing the baseline competitive achievement indicator when all independent variables are zero. β_1 is the coefficient of the variable "the number of training strategy decision-making subjects", representing the influence of the number of training strategy decision-making subjects on the competitive achievement index.β_2 is the coefficient of the "previous year's competitive achievement indicators", reflecting the influence of past competitive achievement indicators on the present.β_3 is a coefficient of the "number of projects", indicating the impact of the number of projects on the competitive achievement indicators.β_4 is a coefficient of the "number of participating individuals", indicating the impact of the number of participating individuals on the competitive achievement indicators.β_5 is the coefficient of the variable "geographical location gain factor", representing the impact of being a geographical location gain factor (the host country) on the competitive achievement indicators.

Model Training and Fitting: For the data of each year, a linear regression model is adopted for training. The training process fits the optimal regression model by minimizing the sum of squared errors (the least square method). The specific steps are as follows:

Feature matrix: The feature matrix X is constructed annually.

Target variable: Construct the target variable y for each year.

Linear regression fitting: Train a linear model using LinearRegression in sklearn to obtain the regression coefficients （β1, β2 ,β3） .

Coefficient extraction: Extract the regression coefficients for each year, with particular attention to those representing the "number of decision-making subjects for the training strategy" (β1...)

3.2.1. The interpretation of regression

Coefficients in the annual linear regression model, the main coefficient of concern is β1. It represents the influence of the number of decision-making subjects for training strategies on the indicators of competitive achievements. The explanation of the regression coefficient is as follows: If a certain feature (such as the number of decision-making subjects in the training strategy) increases by one unit while all other features remain unchanged, the competitive achievement index will increase or decrease the value of this coefficient. For example, in the model of a certain year, if β1= 0.2. It means that for each additional training strategy decision-making subject, the competitive achievement index will increase by 0.2 points. In this way, the influence of various characteristics such as the number of decision-making subjects for training strategies and the number of participating individuals on competitive achievement indicators can be analyzed.

3.2.2. Storage and analysis of results

The regression results (i.e., coefficients) of each year are stored for further analysis. Finally, all the results are saved to a CSV file for future data analysis or report writing.

3.2.3. Result

Through this modeling process, the regression coefficients of different years can be analyzed to understand the impact of characteristics such as the number of decision-making subjects for training strategies, the number of participating individuals, and the number of projects on competitive achievement indicators. These analysis results are helpful for formulating targeted strategies to enhance performance in future international comprehensive sports events.

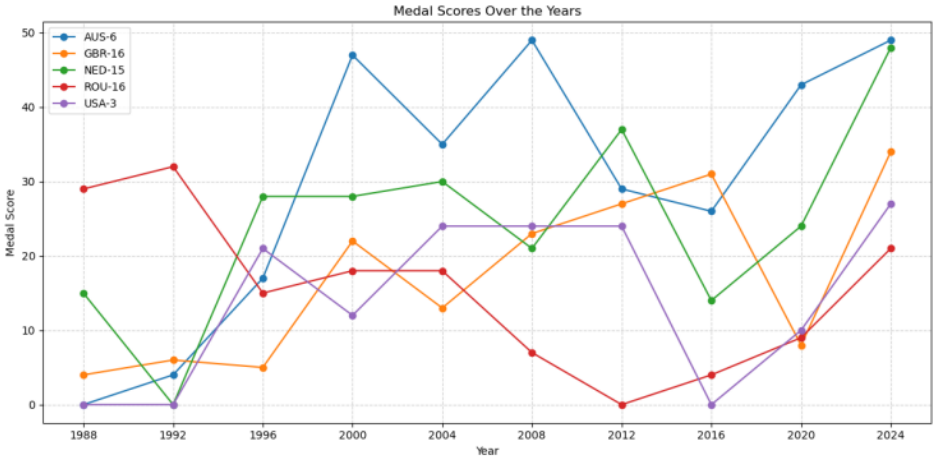

Figure 2. Medal scores over the years

This paper adopts the method of selecting the five items with the most prominent performance stability indicators (standard deviations) among the five participating entities to reveal the impacts of performance fluctuations, as shown in Fig.2. This analysis applies competitive achievement scoring rules, assigning 3 points, 2 points, and 1 point to optimal, secondary, and basic achievements respectively. The models established in this study utilize these scores and incorporate advanced algorithms with dynamic weight adjustment to predict future performance trends. These performance patterns are modeled through dynamic weights, not only considering fluctuations but also able to project future outcomes. The sustained leading position of the US, which rose from approximately 40 points in 1988 to about 50 points in 2024, further confirms the predictive accuracy of the model.

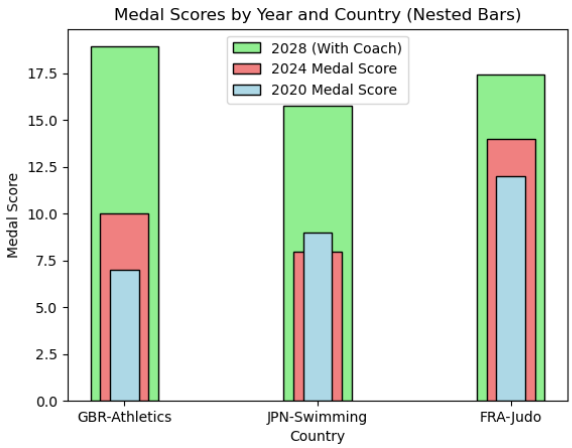

Figure 3. Medal scores by year and country

The overall trend analysis is shown in Fig. 3. In 2028, the competitive achievement indicators for GBR-Athletics, JPN-Swimming, and FRA-Judo are significantly higher than those for 2024 and 2020. The prediction models capture this growth trend effectively. Sub-event data indicate a consistent upward trend in competitive achievement indices for GBR Track and Field, JPN Swimming, and FRA Judo between 2020 and 2028, with all three approaching or exceeding 18 points by 2028, reflecting significant performance improvements across disciplines.

Modeling the influence of external factors, such as training resource optimizations, suggests an average increase of approximately 1 point per event, resulting in an average competitive achievement index of 10.38 points. Trend analysis indicates a significant upward trajectory across all entities from 2024 to 2028. The predictive algorithms, incorporating dynamic adjustment mechanisms, successfully capture both the gradual and rapid performance shifts across different sports categories. The modeling results confirm that structural optimization of key influencing variables leads to notable improvements in prediction accuracy for international competitive performance.

3.3. Intra- and Inter-Regional Correlation Analysis of Event-Specific Capabilities

To ensure analytical accuracy, the study focused on top-performing teams across continents, selecting major contributors for regional event capability correlation analysis:

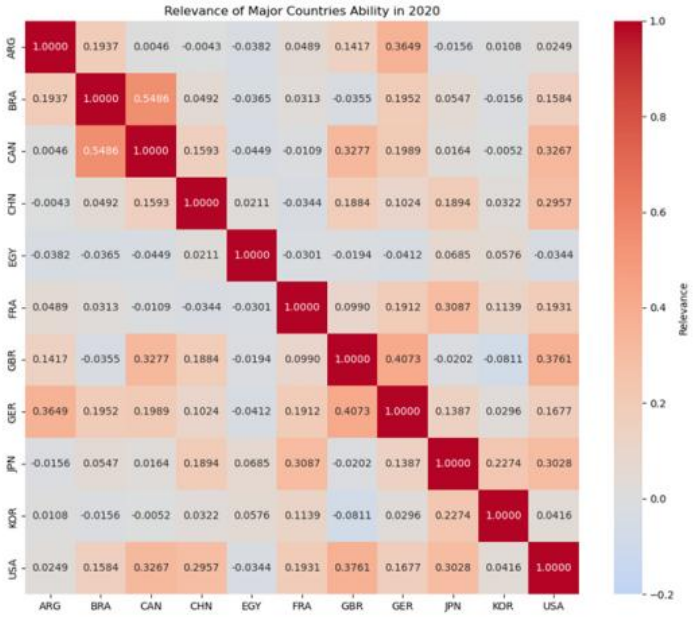

Figure 4. Correlation between Continents in 2020

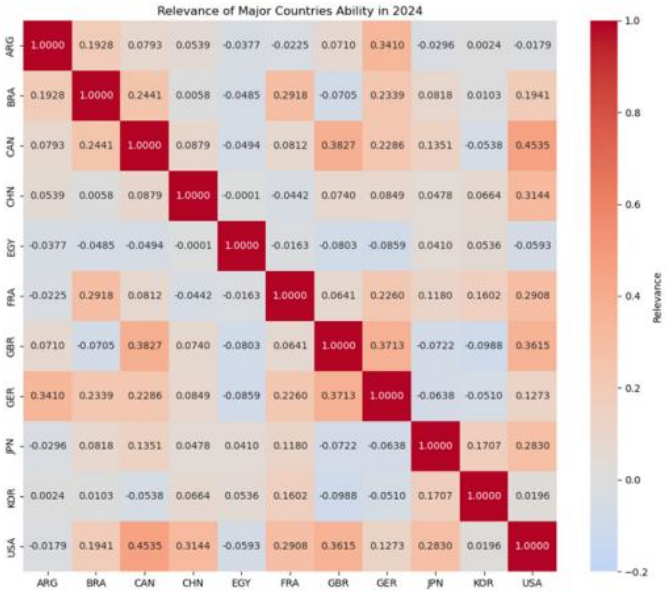

Figure 5. Correlation between Continents in 2024

From 2016 to 2024, intra-regional correlation coefficients remained consistently higher than inter-regional values, indicating stronger internal consistency in event capabilities within regions, especially in East Asia and North America.

The following chart was also employed, where lines of different colors represent different teams, and line thickness indicates the overall correlation between them:

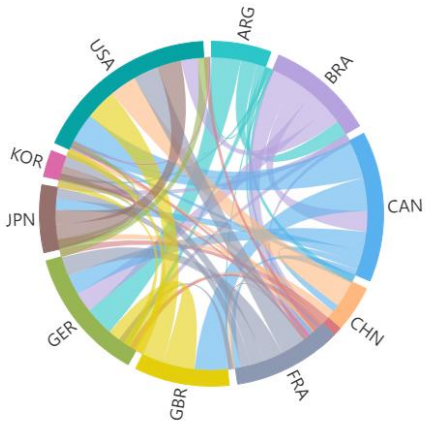

Figure 6. Correlation of National Capacity Values in 2020

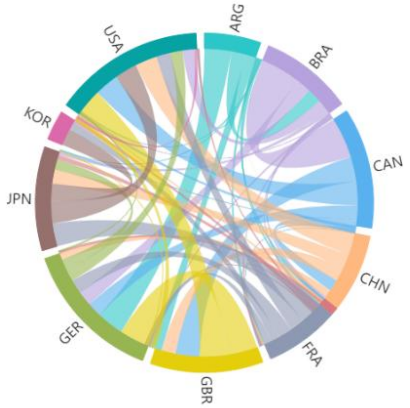

Figure 7. Correlation of National Capacity Values in 2024

Fig.6 and Fig.7 demonstrate that teams within the same region often exhibit similar performance patterns across events, which the model captures as a measurable internal correlation. These correlations are used as key inputs for enhancing the model's regional synergy coefficients. Notably, Western European teams show persistently low inter-team correlation, a feature that the model interprets as a sign of strategic diversification, prompting the adjustment of inter-regional influence parameters.

To account for hosting effects, the model introduces a hosting impact coefficient that dynamically adjusts competitive achievement predictions not only for host nations but also for neighboring countries. This mechanism enhances the model's sensitivity to external performance drivers while controlling for fairness through correlation-weighted influence scaling. The inclusion of region-specific correlation structures and hosting response patterns significantly improves the predictive robustness and adaptability of the algorithm across different geopolitical contexts.

4. Conclusion

This study developed a dynamic weighted prediction model integrating linear regression and random forest algorithms to forecast the distribution of competitive achievement indicators in international multi-sport events. The algorithm identifies potential ranking shifts—such as between the United Kingdom and Australia—through event-specific optimization parameters and hosting effect modeling. Key innovations of the model include a dynamic weighting mechanism that emphasizes recent performance trends and an attenuation function that reduces the influence of outdated historical data. These features significantly improve the model's forecasting precision and responsiveness to structural changes. Although current limitations include a lack of sensitivity to previously unrecognized emerging entities, future enhancement could be achieved by incorporating real-time athlete-level data and multi-dimensional performance indicators.

Overall, the framework offers a scalable and data-driven tool for resource allocation optimization and strategic planning, supporting adaptive governance in competitive sports systems.

References

[1] Maulud D, Abdulazeez A M. A review on linear regression comprehensive in machine learning[J]. Journal of applied science and technology trends, 2020, 1(2): 140-147.

[2] Martin P. Linear regression: An introduction to statistical models[J]. 2022.

[3] Li Y, Mu Y. Research and performance analysis of random forest-based feature selection algorithm in sports effectiveness evaluation[J]. Scientific Reports, 2024, 14(1): 26275.

[4] Gao Z, Kowalczyk A. Random forest model identifies serve strength as a key predictor of tennis match outcome[J]. Journal of Sports Analytics, 2021, 7(4): 255-262.

[5] Fernando K R M, Tsokos C P. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(7): 2940-2951.

[6] Ahmad S, Ullah A, Akgül A, et al. A Novel Homotopy Perturbation Method with Applications to Nonlinear Fractional Order KdV and Burger Equation with Exponential-Decay Kernel[J]. Journal of Function Spaces, 2021, 2021(1): 8770488.

[7] Christoph S, L. S S, Dominik S, et al. Forecasting the Olympic medal distribution – A socioeconomic machine learning model[J]. Technological Forecasting & Social Change,2022,175Christoph A H, K A B, Bergita G. *Learning from machine learning: prediction of age-related athletic performance decline trajectories*. GeroScience, 2021, 43(5): 1–13.

[8] Christoph A H, K A B, Bergita G. Learning from machine learning: prediction of age-related athletic performance decline trajectories. [J]. GeroScience,2021,43(5):1-13.

[9] Zhonghui T, Juan H, Shuo M, et al. Estimating cloud base height from Himawari-8 based on a random forest algorithm[J]. International Journal of Remote Sensing,2020,42(7):2485-2501.

[10] Jiandong H ,Tianhong D ,Yi Z , et al. Predicting the Permeability of Pervious Concrete Based on the Beetle Antennae Search Algorithm and Random Forest Model[J]. Advances in Civil Engineering, 2020.