frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 12 July 2022
doi: 10.3389/fpsyg.2022.915108

Construction of Women's All-Around Speed Skating Event Performance Prediction Model and Competition Strategy Analysis Based on Machine Learning Algorithms

Meng Liu[1†], Yan Chen[1†], Zhenxiang Guo[2], Kaixiang Zhou[1,3], Limingfei Zhou[4], Haoyang Liu[5], Dapeng Bao[6*] and Junhong Zhou[7]*

[1]Sports Coaching College, Beijing Sport University, Beijing, China, [2]Department of Physical Education, Nanjing University of Aeronautics and Astronautics, Nanjing, China, [3]College of Sports, Chengdu University of Traditional Chinese Medicine, Chengdu, China, [4]School of Strength and Conditioning Training, Beijing Sport University, Beijing, China, [5]AI Sports Engineering Lab, School of Sports Engineering, Beijing Sport University, Beijing, China, [6]China Institute of Sport and Health Science, Beijing Sport University, Beijing, China, [7]Harvard Medical School, Hebrew SeniorLife Hinda and Arthur Marcus Institute for Aging Research, Boston, MA, United States

OPEN ACCESS

Edited by:
José Luis Losada, University of Barcelona, Spain

Reviewed by:
Rahim Alhamzawi, University of Al-Qadisiyah, Iraq
Christer Malm, Umeå University, Sweden

***Correspondence:**
Haoyang Liu
liuhaoyang@bsu.edu.cn
Dapeng Bao
baodp@bsu.edu.cn

†These authors have contributed equally to this work and share the first authorship

Specialty section:
This article was submitted to Movement Science and Sport Psychology, a section of the journal Frontiers in Psychology

Received: *07 April 2022*
Accepted: *20 June 2022*
Published: *12 July 2022*

Citation:
Liu M, Chen Y, Guo Z, Zhou K, Zhou L, Liu H, Bao D and Zhou J (2022) Construction of Women's All-Around Speed Skating Event Performance Prediction Model and Competition Strategy Analysis Based on Machine Learning Algorithms. Front. Psychol. 13:915108. doi: 10.3389/fpsyg.2022.915108

Introduction: Accurately predicting the competitive performance of elite athletes is an essential prerequisite for formulating competitive strategies. Women's all-around speed skating event consists of four individual subevents, and the competition system is complex and challenging to make accurate predictions on their performance.

Objective: The present study aims to explore the feasibility and effectiveness of machine learning algorithms for predicting the performance of women's all-around speed skating event and provide effective training and competition strategies.

Methods: The data, consisting of 16 seasons of world-class women's all-around speed skating competition results, used in the present study came from the International Skating Union (ISU). According to the competition rules, distinct features are filtered using lasso regression, and a 5,000 m race model and a medal model are built using a fivefold cross-validation method.

Results: The results showed that the support vector machine model was the most stable among the 5,000 m race and the medal models, with the highest AUC (0.86, 0.81, respectively). Furthermore, 3,000 m points are the main characteristic factors that decide whether an athlete can qualify for the final. The 11th lap of the 5,000 m, the second lap of the 500 m, and the fourth lap of the 1,500 m are the main characteristic factors that affect the athlete's ability to win medals.

Conclusion: Compared with logistic regression, random forest, K-nearest neighbor, naive Bayes, neural network, support vector machine is a more viable algorithm to establish the performance prediction model of women's all-around speed skating event; excellent performance in the 3,000 m event can facilitate athletes to advance to the final, and athletes with outstanding performance in the 500 m event are more likely competitive for medals.

Keywords: machine learning, speed skating, performance prediction, elite athletes, model construction

INTRODUCTION

Accurately predicting the performance during the actual competition can help develop training plans and determine optimal strategies for athletes, which is extremely important to winning the competition (Ofoghi et al., 2016; Bunker and Susnjak, 2022). For example, Novak et al. developed a multiple linear regression model and predicted Olympic distance cross-country mountain biking field performance. Then the knowledge obtained from the prediction helped design appropriate training programs for the athletes in this field (Novak et al., 2018). However, studies have shown that the prediction of athletic performance is challenging because of the complicated scoring system and competition rules of the sport [e.g., all-around speed skating event (Ofoghi et al., 2016)], the requirement of the multi-modal coordination of the physiological systems in athletes (Maier et al., 2018) for the performance of the event.

Specifically, women's all-around speed skating event consists of four successive individual subevents, namely the 500, 1,500, 3,000, and 5,000 m races. Only athletes who ranked top eight the scores in the first three events (i.e., 500, 1,500, and 3,000 m) can enter the final 5,000 m competition. The ranking is by calculating the average time of 500 meters for each event (i.e., the number of seconds the athlete costs is the number of points she scores), and the lower the score, the higher the ranking. This unique scoring system thus requires athletes to utilize different strategies of training and competitions for different goals of this event; that is, some may aim at entering in the last 5,000 m round, and then they aim at winning the medals. Therefore, an advanced prediction model is critical for women's all-around speed skating athletes by providing estimated performance in the following rounds for each athlete (Noordhof et al., 2016). Smyth and Willemsen (2020) previously proposed to use a case-based reasoning technique to analyze the competition results of skaters under different external environmental conditions (e.g., altitude) to help athletes adjust the taxiing rhythm in time to achieve the best sports performance. However, this approach is not suitable for all-around speed skating event. The determinants of entering a 5,000 m race and winning a medal may differ, so the athlete cannot obtain appropriate competition and training advice from this prediction method. Therefore, it is highly demanded to develop a novel prediction model for this event, which will ultimately help improve the athletic performance.

This study proposed a novel prediction model based upon machine learning (ML) techniques. The ML is believed to help make better predictions and formulate more reasonable strategies by learning mass data through its algorithms (Maier et al., 2018). It has been widely used in sport sciences, including analyzing injury risk (Karnuta et al., 2020; Huang and Jiang, 2021) and athletic performance (Sarlis and Tjortjis, 2020; Huang and Jiang, 2021). Recently, studies emerged to implement ML to predict sports competition (Blythe and Király, 2016; Kholkine et al., 2021) and the formulation of strategies for competition (Ofoghi et al., 2013b; Tian et al., 2020). However, no studies have focused on predicting the performance use ML of athletes in all-around speed skating.

This study aims to explore the feasibility of using ML to predict the competition performance in all-around speed skating. Six different ML algorithms —support vector machine (SVM), logistic regression (LR), random forest (RF), K-Nearest Neighbor (KNN), naive Bayes (NB), neural network (NN)—was used here to construct a 5,000 m-race model (i.e., to enter the 5,000 m round) and a medal model (i.e., to win the medals). The performance and functionality of these models were then explicitly examined and compared.

MATERIALS AND METHODS

Data Source and Feature Selection

The data for this study are acquired from the International Skating Union (ISU) official website (https://live.isuresults.eu/home), covering a total of 64 world-class women's all-around speed skating competition results in 16 seasons (i.e., 2003/04–2019/20, except for the 2009/10 season). After being counted, the dataset contains 71 features (**Supplementary Table S1**).

First, the competition result data (mm:ss) are converted into data with s as the unit; then, the data are normalized to be limited within the interval [0, 1] to ensure the model converges against the effect of outliers. The data normalization procedure is formularized as:

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}},\ i = 1, 2, \ldots, n$$

When using ML algorithms for modeling, one needs first to filter out the optimal features to improve the performance of model prediction. If all features are included, it will increase the computational complexity and reduce the model performance. Hence, dimensionality reduction becomes the key to solving the problem. This paper uses the lasso regression method to screen the features of the 5,000 m race and medal models. Lasso regression combines the advantages of both ridge regression and subset selection process so that its computation results reflect the interpretability of subset selection and the stability of ridge regression (Tibshirani, 1996; Hashem et al., 2016; Alhamzawi and Ali, 2018). Lasso regression adds to the minimum sum of squares of errors. Considering the 1-norm constraint on the regression coefficient, the formula can be given as follows:

$$(\alpha, \beta) = \arg\min \sum_{i=1}^{n} (y_i - \alpha_i - X_i\beta)^2 \text{ sbuject to} \|\beta\|_1 < t$$

Add the constraint in the above formula to get the following form:

$$(\alpha, \beta) = \arg\min \sum_{i=1}^{n} (y_i - \alpha_i - X_i\beta)^2 + \lambda \|\beta\|_1$$

where, X_i is the ith group of independent variables, which are row parameters; α and β are regression coefficients, and β is the column parameter, and $\|\beta\|_1$ represents the 1-norm, which is the sum of the absolute values of the elements in

the parameters; y_i is the value of the dependent variable of X_i; n is the size of the dataset used for regression modeling; λ and t are the parameters in different forms of lasso regression.

Machine Learning Model Building and Verification

Six instances of the 5,000 m race prediction model and the medal prediction model are established through SVM, RF, LR, KNN, NB, and NN algorithms (**Supplementary Figure S1**); the output of the model is whether the athlete can enter the 5,000 competition or win a medal. The fivefold cross-validation method was used to verify the model's performance. The specific process was splitting the dataset into five groups and assigning them each to an independent folder, four groups used as training data for building the model, and the remaining one used as test data to verify the model's effectiveness. Then, this process was repeated five times, and each of the five verifications was used as the result only once. Then take the average of the five results to get an estimate.

Among the algorithms, SVM adopts the linear kernel function as the primary function (Linear Support Vector Classifier, LSVC), given a set of labels corresponding to the instance, $i=1,\ldots,l, x_i \in R^P, y_i \in \{-1,+1\}$, which solves an unconstrained loss function optimization problem $\xi(\mathrm{w};,x_i;,y_i)$:

$$\min_w = \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi\left(w;,x_i;,y_i\right)$$

The L2-SVM loss function is used in this study:

$$\xi\left(w;,x_i;,y_i\right) = \max\left(1 - y_i w^T x_i, 0\right)^2$$

Naive Bayes adopts Gaussian Naive Bayes:

$$P\left(x_i \middle| y\right) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right)$$

where σ_y and μ_y are estimated using maximum likelihood estimation.

Logistic regression uses the L2 penalty logistic regression function:

$$\min_{w,c} \frac{1}{2} w^T w + C\sum_{i=1}^{n} \log\left(\exp\left(-y_i\left(X_i^T w + c\right)\right) + 1\right)$$

The KNN function can be expressed as (Euclidean distance):

$$p_{ij} = \frac{\exp\left(-Lx_i - Lx_j{}^2\right)}{\sum_{k\neq i} \exp - \left(Lx_i - Lx_k{}^2\right)},\ p_{ii} = 0$$

Uses the Bootstrap method to select n samples from the sample set and generates n classification trees to form a random forest (Breiman, 2001; Austin et al., 2013). The voting result of the classification tree determines the classification prediction result of the new data as expressed by the following formula:

$$f\left(x\right) = \arg\max_Y \sum_{i=1}^{n} I\left(h_i\left(X\right) = Y\right)$$

Where h_i represents the basic model of a single classification tree, Y represents the output variable, and I mean the indicative function.

The neural network model uses Multi-layer perceptron (MLP). A set of training examples $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, $x_i \in R^n$ are given in the MLP, $y_i \in \{0,1\}$, one hidden layer and one hidden neuron MLP learning function.

$$f\left(x\right) = W_2 g\left(W_1^T x + b_1\right) + b_2$$

with $W_1 \in R^m$, $W_2, b_1, b_2 \in \mathrm{R}$ being the model parameters. W_1 *and* W_2 represent the weights of the input layer and the hidden layer, respectively; b_1 *and* b_2 represent the deviations added to the hidden layer and the output layer, respectively; $g(\cdot): R \to R$ is the activation function, set by default as the hyperbolic tangent given by:

$$g\left(z\right) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

For binary classification, $f\left(x\right)$ yields an output value between 0 and 1 through the logic function $g\left(z\right) = 1/\left(1 + e^{-z}\right)$. Samples are assigned to the positive class if having an output value greater than or equal to the threshold 0.5, else to the negative class.

The algorithm and evaluation are implemented using Scikit-learn based on Python 3 (Pedregosa et al., 2011). In the training process, the main parameters of different instances of the models are adjusted. The grid search method is used to adjust the hyperparameters to find the parameter value corresponding to the highest accuracy provided that the training data exist.

Model Evaluation

Evaluation indicators include the area under the receiver operating characteristic curve (ROC) AUC, accuracy, sensitivity, precision, and balanced F1 score. AUC is used to evaluate the discriminative ability and performance of the model. When the value of AUC is 1, it means that the model is perfect; a value of 0.5 means the deficient performance of a random classifier, i.e., the random classifier does not have any discriminative ability; a value of 0.90–1 means excellent, 0.80–0.90 good, 0.70–0.80 fair, 0.60–0.70 poor, and 0.50–0.60 failure (Bruce et al., 2020). The correct rate is the proportion of the samples judged correctly by the classifier among all samples. The higher the correct rate, the better the classifier; sensitivity is the proportion of all positive examples judged correctly by the classifier, which measures the classifier's ability to recognize positive examples; accuracy represents the proportion of positive

examples judged to be positive by the classifier; the F1 score is the weighted average of model accuracy and recall; the maximum of the four indicators is 1, the minimum is 0, and the higher the value, the better the model (Stehman, 1997). Among the results of judgment, TP = true positive, TN = true negative, FP = false positive, FN = false negatives.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2TP}{2TP + FP + FN}$$

Feature Weight Calculation

The present study quantifies the impact of the included features on model performance by computing weights (Li et al., 2020). To this end, the LSVC model is used in Python 3 Scikit-learn.

RESULTS

Feature Selection Results of Lasso Regression

Feature Inclusion in the 5,000 m Competition Model

Features other than those associated with the 5,000 m race were filtered using Lasso regression analysis to determine the best features to build the model. When λ was equal to 0.051, the model based on the following six features performed best: 3,000 m 1st lap score (3,000 m1), 3,000 m 7th lap score (3,000 m7), 1,500 m 1st lap score (1,500 m1), 3,000 m 8th split timer (3,000 ms8), 1,500 m 2nd split timer (1,500 ms^2), and 3,000 m points (3,000 m Points; **Figure 1**).

Feature Inclusion in the Medal Model

Lasso regression analysis was used to screen all features. When λ was equal to 0.0054, the model based on the following 21 features performed best: 500 m 2nd lap score (500 m2), 3,000 m 2nd lap score (3,000 m2), 3,000 m 3rd lap score (3,000 m3), 3,000 m 4th lap score (3,000 m4), 3,000 m 5th lap score (3,000 m5), 3,000 m 7th lap score (3,000 m7), 1,500 m 2nd lap score (1,500 m2), 1,500 m 3rd lap score (1,500 m3), 1,500 m 4th lap score (1,500 m4), 5,000 m 1st lap score (5,000 m1), 5,000 m 2nd lap score (5,000 m2), 5,000 m 4th lap score (5,000 m4), 5,000 m 5th lap score (5,000 m5), 5,000 m 9th lap score (5,000 m9), 5,000 m 11th lap score (5,000 m11), 5,000 m 13th lap score (5,000 m13), 5,000 m 3rd split timer (5,000 ms3), 500 m ranking, 3,000 m ranking, 1,500 m ranking, and 5,000 m ranking (**Figure 2**). In order to facilitate the actual operation, the 500 m ranking, 3,000 m ranking, 1,500 m ranking, and 5,000 m ranking from which features cannot be directly extracted in the test process are excluded, and the remaining 17 features were retained.

Performance Prediction Model Results

Evaluation and Comparison of the 5,000 m Race Model for Women's All-Around Speed Skating Event

According to the plotted ROC curve (**Figure 3**), the AUC values of the six instances of the 5,000 m race model for women's all-around speed skating event established by SVM, RF, LR, KNN, NB, and NN are 0.86, 0.85, 0.85, 0.83, 0.64, and 0.85, respectively. It can be observed that the overall better-performing algorithms are SVM, RF, LR, and NN. SVM had the most balanced classification through a comprehensive comparison of accuracy, sensitivity, and F1 score (**Table 1**).

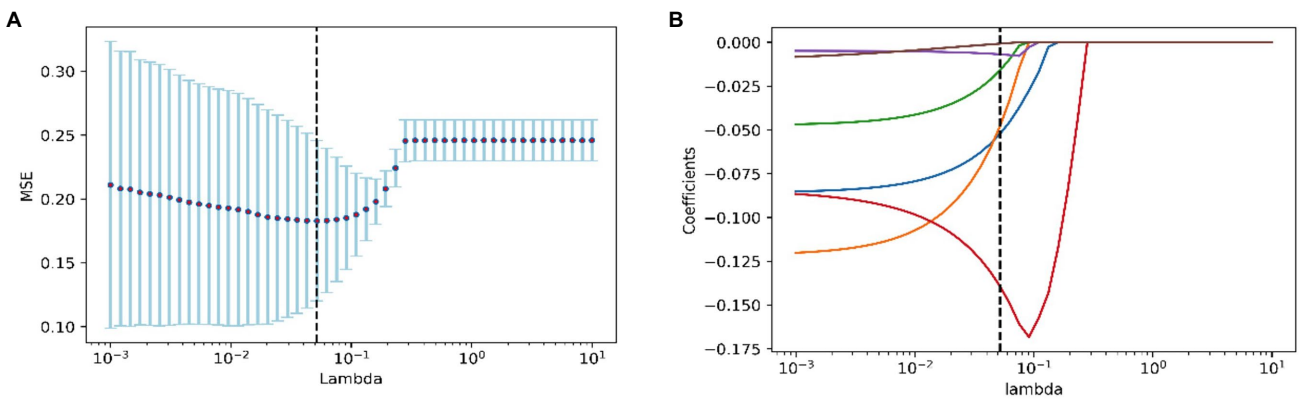

FIGURE 1 | Feature screening of the 5,000 m Race Model. **(A)** The figure of test MSE by lambda value; **(B)** the path diagram of lasso regression coefficients.

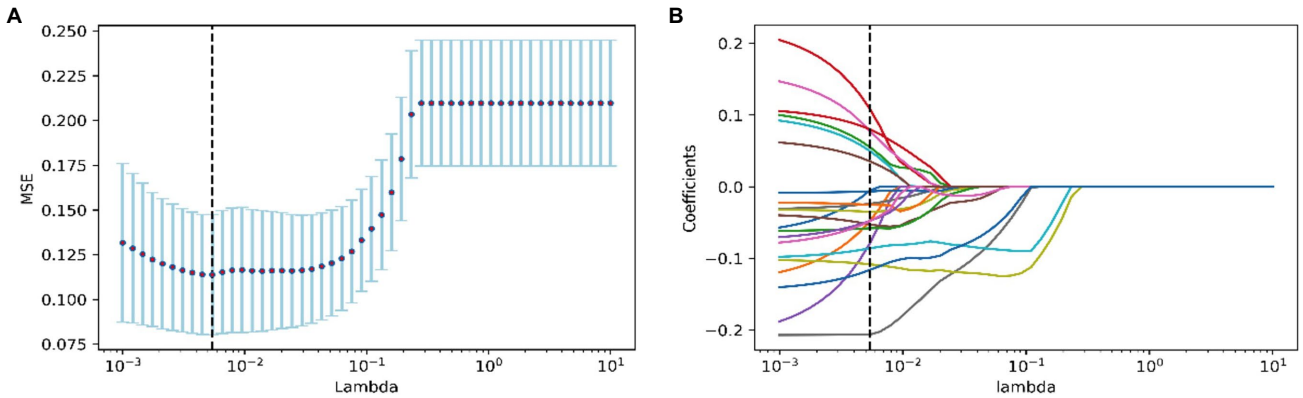

FIGURE 2 | Feature screening of the Medal Model. **(A)** The figure of test MSE by lambda value; **(B)** the path diagram of lasso regression coefficients.

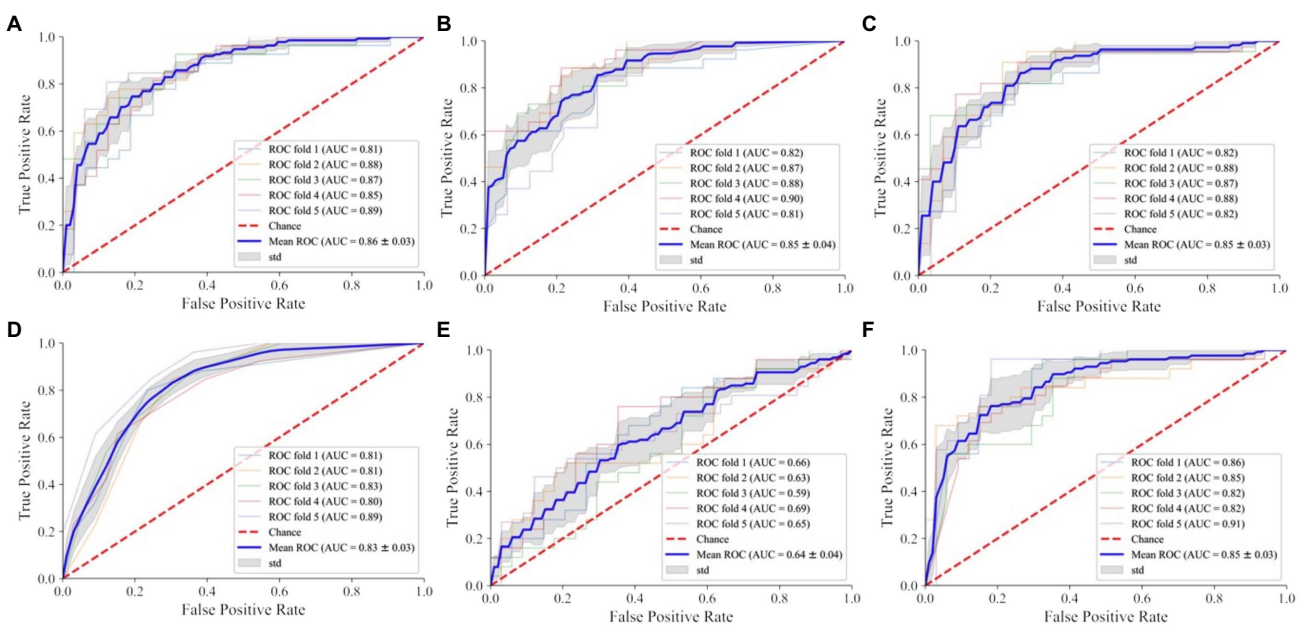

FIGURE 3 | Receiver operating characteristic (ROC) curves of different models in the 5,000 m final of women's all-around speed skating event. **(A)** SVM: support vector machine; **(B)** RF: random forest; **(C)** LR: logistic regression; **(D)** KNN: K-nearest neighbor; **(E)** NB: naive Bayes; **(F)** NN: neural network.

Evaluation and Comparison of the Medal Model

In training the instances of the medal model, the NN-based instance fails due to the excess data size. According to the plotted ROC curve (**Figure 4**), the AUC values of the five medal events for women's all-around speed skating event established by SVM, RF, LR, KNN, and NB are 0.81, 0.73, 0.73, 0.70, and 0.60, respectively. Among these model instances, the SVM instance proves high-performing and is the only instance that demonstrates good stability through a comprehensive comparison of accuracy, sensitivity, and F1 score of the five models (**Table 2**).

Feature Weight Analysis

According to the feature weights calculated by LSVC, the scores of the 3,000 m laps 1st and 7th and the individual points of the 3,000 m are the most critical features that affect whether an athlete can enter the 5,000 m competition (**Figure 5A**). The results of 5,000 m lap 11th, 500 m lap 2nd, and 1,500 m lap 4th are positive characteristics that affect whether athletes can win medals, while the results of 5,000 m laps 9th, 13th and 3,000 m lap 3rd are negative characteristics that affect whether athletes can win medals (**Figure 5B**).

DISCUSSION

This study examined six ML algorithms approaches based upon a real competition database of split times and ranking in women's all-around speed skating athletes. The results have shown that it is feasible to predict the performance ranking by ML algorithms. Through comparison in the performance

of the instances of the models built by different algorithms, it has been observed that the SVM-based instance can effectively predict the performance of the women's all-around speed skating event, suggesting that this model would help athletes to set appropriate training programs, improving the quality of their strategic decision-making and competitive performance.

The model for performance prediction in women's all-around speed skating event established through ML can provide more direct suggestions for the training and competition of this event. For instance, coaches and athletes can input the daily test results into the model to obtain the probability of athletes entering the 5,000 m competition or winning medals to help athletes and coaches in training better. This is more generic than Smyth's (Smyth and Willemsen, 2020) use of specific case-based reasoning. The selection of features is the key to building this more general model (Horvat et al., 2020). Research has shown that lasso regression has advantages over traditional Stepwise Regression methods in feature selection (Yarkoni and Westfall, 2017). Applied for feature screening in this study, the lasso regression method is more conducive to eliminating unimportant related features and accurately screening out relatively important ones. Combined with weight calculation, the model can be more accessible and interpretable. In this study, the features of the 5,000 m race model and the medal model are distinct. In deciding whether athletes are eligible for entering the 5,000 m competition, the 3,000 m score has the most relevant features, while in determining whether the final result suffices to win a medal, things are different. Laps 1, 2, 4, 7, 11 at 5,000 m, 5, 7 laps at 3,000 m, lap 2 at 500 m, and before 3,000 m, the speed of the three laps has an important influence on whether the athlete can win a medal. This reminds coaches that the training emphasis of athletes should be highlighted for different competition purposes. If the athlete's goal is to enter the 5,000 m race, she should first develop the long-distance racing ability until scoring high enough in the 3,000 m race for entering the 5,000 m race. However, if the athlete's goal is to win a medal, she should also pay attention to the development of speeding ability. Athletes must not have apparent shortcomings; otherwise, the final ranking will probably be affected by the 500 m score. Athletes with outstanding 500 m scores are easier to win a medal. Moreover, one can notice that the medal winners of laps 9 and 13 of the 5,000 m race do not outspeed the non-winners. This seems to reveal that having faster speed in the first half of the 5,000 m race can be more conducive to good results. Previous studies have reported that active start-up acceleration and forward speed are conducive to achieving better athletic performance (Muehlbauer et al., 2010). This revelation also provides a reference for athletes to formulate competitive strategies. Previous studies have also shown that the decrease in the second half

TABLE 1 | Validity evaluation of different prediction models for the 5,000 m final of women's all-around speed skating event.

ML	Accuracy	Sensitivity	Precision	F1 Score
SVM	0.78 ± 0.03	0.77 ± 0.05	0.73 ± 0.04	0.75 ± 0.03
RF	0.76 ± 0.04	0.81 ± 0.06	0.67 ± 0.03	0.73 ± 0.03
LR	0.77 ± 0.04	0.76 ± 0.05	0.66 ± 0.08	0.70 ± 0.05
KNN	0.72 ± 0.01	0.71 ± 0.11	0.68 ± 0.04	0.69 ± 0.06
NB	0.62 ± 0.01	0.57 ± 0.04	0.66 ± 0.05	0.63 ± 0.04
NN	0.72 ± 0.03	0.65 ± 0.04	0.71 ± 0.04	0.68 ± 0.05

SVM, support vector machine; RF, random forest; LR, logistic regression; KNN, K-nearest neighbor; NB, naive Bayes; NN, neural network.

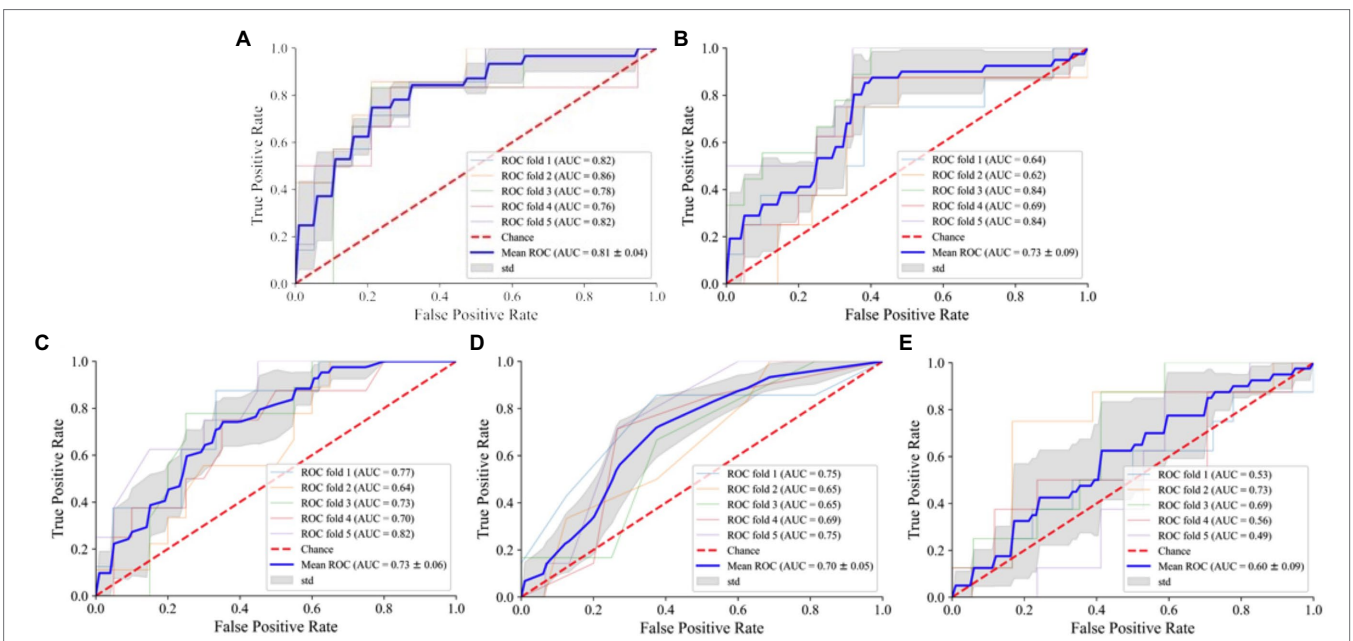

FIGURE 4 | ROC curve of different models of women's speed skating medals. **(A)** SVM: support vector machine; **(B)** RF: random forest; **(C)** LR: logistic regression; **(D)** KNN: K-nearest neighbor; **(E)** NB: naive Bayes.

of the competition speed may increase the push-off angle associated with fatigue (Noordhof et al., 2013). Therefore, improving the technical stability of athletes in a fatigued state is crucial to improving sports performance. This also provides a particular idea for the election of athletes. When all-around speed skaters are elected among women athletes, sufficient attention should be paid to those with excellent aerobic capacity and explosive power.

Since modeling in this study aims to determine the probability of athletes entering the finals and winning medals, six classification algorithms were selected when each model was established. The most commonly used ML algorithms in sports include SVM, RF, LR, KNN, NB, and NN (Horvat et al., 2020). Some of these algorithms have been applied to predict the performance of some events. For example, Ofoghi et al (2013a) used *K*-means combined with traditional statistical methods to model the performance prediction of athlete election after the all-around track cycling competition system was changed, which effectively helped coaches elect athletes and develop the appropriate training plan. The research by Dwyer et al. found that the triathlon performance prediction model based on the NB algorithm is also effective. This model helps coaches and athletes formulate reasonable competitive strategies to optimize athletes' sports performance (Ofoghi et al., 2016). In addition, other researchers have also used different ML algorithms for performance prediction in different events (Richter et al., 2021). In this study, ML proves effective and feasible in predicting the performance ranking in women's all-around speed skating event by learning from past competition data and establishing a viable model. Theoretically, these six models can predict performance, but the comparison has revealed differences in the predicted performance between the prediction models built on different ML algorithms. For the 5,000 m final prediction model, the AUC values of SVM, RF, LR, KNN, and NN are similar. The SVM-based instance model has achieved the best overall performance, while the AUC value of the NB-based instance is only 0.64, and its accuracy, sensitivity, and F1 score are also low (**Table 1**). For the medal model, the SVM-based instance of the medal model has also shown a good performance (**Table 2**), while the NB-based instance has performed relatively poorly, and the NN-based instance has failed.

TABLE 2 | Effectiveness of the prediction models for women's all-around speed skating medal.

ML	Accuracy	Sensitivity	Precision	F1 score
SVM	0.80 ± 0.07	0.71 ± 0.06	0.63 ± 0.04	0.67 ± 0.08
RF	0.73 ± 0.02	0.43 ± 0.08	0.58 ± 0.02	0.49 ± 0.05
LR	0.78 ± 0.06	0.42 ± 0.08	0.8 ± 0.2	0.55 ± 0.08
KNN	0.75 ± 0.07	0.51 ± 0.06	0.63 ± 0.8	0.55 ± 0.8
NB	0.60 ± 0.07	0.59 ± 0.06	0.42 ± 0.08	0.49 ± 0.06

The above differences may be ascribed to the characteristics of different algorithms. Based on conditional probability, the NB algorithm uses Bayes' theorem to calculate the probability by determining the combination of the frequency and the historical data values. It also rests on the assumption of a given output and that the interclass attributes are independent, but this assumption is difficult to hold in practice (Rish, 2001). The same is true in this study. The NN algorithm has very high requirements on data size, which may be the reason for not being able to establish the medal model. The SVM is highly applied in solving relatively small sample predictions and is more sensitive to data. Given the relatively small dataset in this study, the final decision function of SVM has been determined by only a few support vectors. The computational complexity depends on the number of support vectors rather than on the dimensionality of the sample space, and the direct association between the input variables in this study avoids the "curse of dimensionality" in some sense (Shalev-Shwartz et al., 2011). The SVM algorithm is widely used in the domain of sports. For example, the maximum oxygen uptake prediction model established by the SVM algorithm has good prediction accuracy (Abut and Akay, 2015), the gait diagnosis model established by Begg et al. through SVM is also of high applied value (Begg et al., 2005), and the Chinese Super League ranking model built on the SVM algorithm also has high accuracy (Li et al., 2020). From the results of this research, the SVM algorithm is also feasible for performance prediction. The NB algorithm has shown application prospects for predicting the performance of complex events in previous studies, such as all-around track cycling (Ofoghi et al., 2013a), triathlon (Ofoghi et al., 2016), decathlon (Trevor et al., 2002). However, because the NB algorithm assumes that the sample attributes are independent, its effect is not satisfactory when the sample attributes are correlated. In this study, the included features may have a strong correlation, such that the NB algorithm

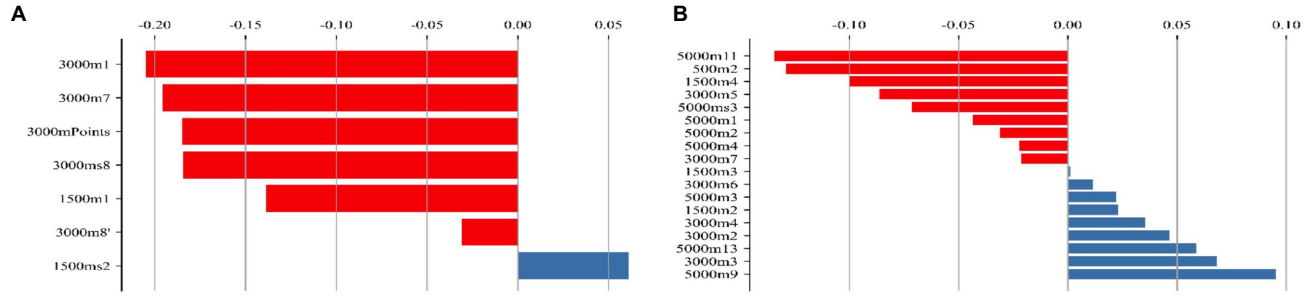

FIGURE 5 | Weight of feature. **(A)** The feature weights of the 5,000 m race model; **(B)** the feature weights of the medal model.

becomes less suitable for the prediction model. NN is considered an excellent ML algorithm, but the model has poor interpretability due to the extremely high data requirements and the "black box" problem. Still, the research results show that NN does not necessarily outperform other ML algorithms in performance prediction (Bunker and Susnjak, 2022).

To sum up, the present work is that it provides information that can be used to predict future performances in women's all-around speed skating with a certain level of accuracy. The mathematical models that form the basis for these predictions were developed from an analysis of historical race data. We believe that our analytical approach is reasonable to be confident about the accuracy of our results. Although we have performed a great deal of work, this study still had some limitations. First, this research has overfitted the available data when using NN to build the medal prediction model due to the relative lack of data. In the future, with the increase in data size, neural networks will be helpful in prediction. Secondly, ignoring the different event settings, this study failed to explore men's all-around speed skating event. Future research can conduct a comparative study between men's and women's events.

CONCLUSION

The ML algorithm has proven feasible in predicting women's all-around speed skating competition performance. The prediction model built on SVM has proven more suitable for predicting women's all-around speed skating competition performance compare with LR, RF, KNN, NB, and NN. Female speed skaters with excellent results in the 3,000 m race are entitled to enter the all-around final, while athletes with outstanding results in the 500 m race are strong competitors for a medal.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

KZ, DB, ML, and JZ: design and/or conceptualization of the study. KZ, ML, YC, LZ, JZ, and DB: analysis and/or interpretation of the data. KZ, JZ, and DB: drafting and/or revising the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the National Key Research and Development Program of China (Grant Numbers 2018YFC2000602 and 2019YFF0301803).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.915108/full#supplementary-material

REFERENCES

Abut, F., and Akay, M. F. (2015). Machine learning and statistical methods for the prediction of maximal oxygen uptake: recent advances. *Med Devices (Auckl)* 8, 369–379. doi: 10.2147/mder.S57281

Alhamzawi, R., and Ali, H. T. M. (2018). The Bayesian adaptive lasso regression. *Math. Biosci.* 303, 75–82. doi: 10.1016/j.mbs.2018.06.004

Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., and Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J. Clin. Epidemiol.* 66, 398–407. doi: 10.1016/j.jclinepi.2012.11.008

Begg, R., and Kamruzzaman, J. J. J. (2005). A machine learning approach for automated recognition of movement patterns using basic, kinetic and kinematic gait data. *J. Biomech.* 38, 401–408. doi: 10.1016/j.jbiomech.2004.05.002

Blythe, D., and Király, F. J. (2016). Prediction and quantification of individual athletic performance of runners. *PLoS One* 11:e0157257. doi: 10.1371/journal.pone.0157257

Breiman, L. (2001). Random forests. *JML* 45, 5–32.

Bruce, P., Bruce, A., and Gedeck, P. (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. Newton, MA: O'Reilly Media.

Bunker, R., and Susnjak, T. (2022). The application of machine learning techniques for predicting results in team sport: a review. *J. Artif. Intell. Res.* 73, 1285–1322. doi: 10.48550/arXiv.1912.11762

Hashem, H., Vinciotti, V., Alhamzawi, R., and Yu, K. (2016). Quantile regression with group lasso for classification. *ADAC* 10, 375–390. doi: 10.1007/s11634-015-0206-x

Horvat, T., Job, J., and Discovery, K. (2020). The use of machine learning in sport outcome prediction: a review. *Wiley Interdiscip. Rev. Data Min. Knowl.* 10:e1380. doi: 10.1002/widm.1380

Huang, C., and Jiang, L. (2021). Data monitoring and sports injury prediction model based on embedded system and machine learning algorithm. *Microprocess. Microsyst.* 81:103654. doi: 10.1016/j.micpro.2020.103654

Karnuta, J. M., Luu, B. C., Haeberle, H. S., Saluan, P. M., Frangiamore, S. J., Stearns, K. L., et al. (2020). Machine learning outperforms regression analysis to predict next-season Major League Baseball player injuries: epidemiology and validation of 13,982 player-years From performance and injury profile trends, 2000-2017. *Orthop. J. Sports Med.* 8:2325967120963046. doi: 10.1177/2325967120963046

Kholkine, L., Servotte, T., de Leeuw, A. W., De Schepper, T., Hellinckx, P., Verdonck, T., et al. (2021). A learn-to-rank approach for predicting road cycling race outcomes. *Front. Sports Act. Living* 3:714107. doi: 10.3389/fspor.2021.714107

Li, Y., Ma, R., Gonçalves, B., Gong, B., Cui, Y., Shen, Y. J. C., et al. (2020). Data-driven team ranking and match performance analysis in Chinese Football Super League. *Chaos Solit. Fractals* 141:110330. doi: 10.1016/j.chaos.2020.110330

Maier, T., Meister, D., Trösch, S., and Wehrlin, J. P. (2018). Predicting biathlon shooting performance using machine learning. *J. Sports Sci.* 36, 2333–2339. doi: 10.1080/02640414.2018.1455261

Muehlbauer, T., Panzer, S., and Schindler, C. (2010). Pacing pattern and speed skating performance in competitive long-distance events. *J. Strength Cond. Res.* 24, 114–119. doi: 10.1519/JSC.0b013e3181c6a04a

Noordhof, D. A., Foster, C., Hoozemans, M. J., and de Koning, J. J. (2013). Changes in speed skating velocity in relation to push-off effectiveness. *Int. J. Sports Physiol. Perform.* 8, 188–194. doi: 10.1123/ijspp.8.2.188

Noordhof, D. A., Mulder, R. C., de Koning, J. J., and Hopkins, W. G. (2016). Race factors affecting performance times in elite long-track speed skating. *Int. J. Sports Physiol. Perform.* 11, 535–542. doi: 10.1123/ijspp.2015-0171

Novak, A. R., Bennett, K. J. M., Fransen, J., and Dascombe, B. J. (2018). A multidimensional approach to performance prediction in Olympic distance cross-country mountain bikers. *J. Sports Sci.* 36, 71–78. doi: 10.1080/02640414.2017.1280611

Ofoghi, B., Zeleznikow, J., Dwyer, D., and Macmahon, C. (2013a). Modelling and analysing track cycling Omnium performances using statistical and machine learning techniques. *J. Sports Sci.* 31, 954–962. doi: 10.1080/02640414.2012.757344

Ofoghi, B., Zeleznikow, J., MacMahon, C., and Dwyer, D. (2013b). Supporting athlete selection and strategic planning in track cycling omnium: a statistical and machine learning approach. *Inf. Sci.* 233, 200–213. doi: 10.1016/j.ins.2012.12.050

Ofoghi, B., Zeleznikow, J., Macmahon, C., Rehula, J., and Dwyer, D. (2016). Performance analysis and prediction in triathlon. *J. Sports Sci.* 34, 607–612. doi: 10.1080/02640414.2015.1065341

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.48550/arXiv.1201.0490

Richter, C., O'Reilly, M., and Delahunt, E. (2021). Machine learning in sports science: challenges and opportunities. *Sports Biomech.* 1–7. doi: 10.1080/14763141.2021.1910334

Rish, I. (2001). An empirical study of the naive Bayes classifier. *Int. J. Recent Innov.* 1:127.

Sarlis, V., and Tjortjis, C. J. I. S. (2020). Sports analytics—evaluation of basketball players and team performance. *Inf. Syst.* 93:101562. doi: 10.1016/j.is.2020.101562

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2011). Pegasos: primal estimated sub-gradient solver for SVM. *Math. Program.* 127, 3–30. doi: 10.1007/s10107-010-0420-4

Smyth, B., and Willemsen, M. C. (2020). "Predicting the personal-best times of speed skaters using case-based reasoning," in *International Conference on Case-Based Reasoning* (Springer), 112–126.

Stehman, S. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.* 62, 77–89. doi: 10.1016/S0034-4257(97)00083-7

Tian, C., De Silva, V., Caine, M., and Swanson, S. J. A. S. (2020). Use of machine learning to automate the identification of basketball strategies using whole team player tracking data. *Appl. Sci.* 10:24. doi: 10.3390/app10010024

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Trevor, F., Ryan, T., and Cox, D. (2002). An analysis of decathlon data. *J. R. Stat. Soc. Series D* 51, 179–187. doi: 10.1111/1467-9884.00310

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.