Molecular & Cellular Biomechanics 2025, 22(3), 1081.
https://doi.org/10.62617/mcb1081

Article

Research on sports injury risk assessment and rehabilitation strategy based on big data analysis

Jiawei Cao*, Hongxi Chen

SCHOOL OF SPORTS AND HEALTH, YUNNAN VOCATIONAL COLLEGE OF SPORTS, Kunming 650228, China
*** Corresponding author:** Jiawei Cao, caozinai@126.com

CITATION

Cao J, Chen H. Research on sports injury risk assessment and rehabilitation strategy based on big data analysis. Molecular & Cellular Biomechanics. 2025; 22(3): 1081. https://doi.org/10.62617/mcb1081

ARTICLE INFO

Received: 11 December 2024
Accepted: 31 December 2024
Available online: 18 February 2025

Abstract: By collecting athletes' basic information, exercise habits, historical injury records and sports performance data, this study constructs a random forest (RF) model to assess the risk of sports injuries. The model can effectively deal with high-dimensional data and capture nonlinear relationships, and has strong generalization ability. The study also defines a risk assessment index (RAI) to visually represent the risk level of athletes' sports injuries. In addition, this study identified the specific rehabilitation needs of patients with different injury types and degrees through association rule mining technology and cluster analysis, and made a personalized rehabilitation plan. In particular, biomechanical data, such as joint stability and muscle strength balance, are also included in this study to more accurately assess the risk of sports injury and guide rehabilitation training. Through comparative experiments, the results show that personalized rehabilitation plan based on big data analysis can significantly shorten the rehabilitation cycle and improve the quality of rehabilitation and patient satisfaction. The results of this study not only provide scientific sports guidance and rehabilitation suggestions for athletes and fitness enthusiasts, but also provide decision support for sports coaches, rehabilitation teachers and other professionals, which promotes the development of theory and practice in the field of sports injury prevention and rehabilitation.

Keywords: big data; sports injury; risk assessment; rehabilitation strategy; random forest; biomechanics

1. Introduction

Sports injury not only affects athletes' competitive performance and career, but also poses a threat to the health of ordinary sports enthusiasts. Therefore, how to effectively assess the risk of sports injury and formulate scientific and reasonable rehabilitation strategies has become an urgent problem in the field of sports medicine and sports science. According to the statistics of the World Health Organization, the number of injuries caused by sports in the world is huge every year, which not only brings physical pain and economic burden to individuals, but also poses great pressure on social medical resources. Therefore, how to effectively prevent sports injuries and how to recover quickly and scientifically after injuries have become the key problems to be solved urgently in the fields of sports science, medicine and rehabilitation. Big data analysis can process massive and multi-dimensional data and mine the laws hidden behind the data, which makes it possible to accurately predict the risk of sports injuries and customize rehabilitation programs. Compared with traditional research methods based on experience or small-scale samples, big data analysis can consider individual differences more comprehensively, improve the accuracy of evaluation and the effectiveness of rehabilitation strategies.

The sports injury assessment model based on big data network can simultaneously assess the horizontal and vertical injury risks, and clearly determine whether the injury site is a single injury or a compound injury [1]. These models complete the construction of the evaluation model through the determination of sports injury risk sources, identification of injury risk factors and sports injury evaluation based on injury risk factors [2]. Data mining technology plays a key role in sports injury prediction. By analyzing a large number of sports data, we can find out the potential risk factors of injury and take preventive measures in advance. For example, the whole body posture assessment can help rehabilitation workers find out the risk of sports injury in patients, such as muscle tension, limited joint activity and poor stability [3,4]. Functional Movement Screening (FMS) is a method used to evaluate an athlete's movement capabilities, and it has a certain relationship with the risk of sports injuries. By conducting FMS tests on elite athletes and recording instances of sports injuries, an FMS database for athletes can be established, thereby assessing the risk of sports injuries [5,6]. Research into physical factor-assisted motor and sensory rehabilitation represents a current frontier and hot topic. Utilizing physical stimuli such as electromagnetic and photoacoustic can modulate neural circuits associated with sensorimotor functions, promoting motor and sensory recovery. However, the mechanisms of existing physical factor rehabilitation interventions remain unclear, and optimal intervention parameters are still under investigation [7]. The study of sports injury risk assessment and rehabilitation strategies based on big data analysis is a multidisciplinary field that integrates knowledge from sports science, data science, rehabilitation medicine, and other areas [8]. Through big data analysis, it is possible to more accurately assess the risk of sports injuries and develop more effective rehabilitation strategies.

Age is an important factor that affects the type of sports injury and its impact on long-term sports career. Young athletes are in the stage of growth and development, and their skeletal and muscular systems are not yet fully mature, making them more prone to injuries such as growth plate injuries and ligament strains; However, middle-aged and elderly athletes are more susceptible to joint degenerative diseases, tendinitis, and other injuries due to decreased physical function. These injuries not only affect athletes' immediate performance, but may also limit their ability to participate in sports activities in the future [9]. Gender plays an important role in the incidence and recovery process of sports injuries. Due to differences in physiological structure and hormone levels, female athletes have a higher incidence of certain types of sports injuries than males, such as anterior cruciate ligament (ACL) injuries [10]. In addition, gender differences are also reflected in the recovery and rehabilitation process after injury. Female athletes may need longer recovery time, and there are differences in pain management and psychological adjustment during rehabilitation with men. From amateur to professional in different levels of sports, sports injury patterns and prevention strategies also show changes [11]. Amateur athletes are more prone to acute injuries due to unsystematic training and irregular techniques. Professional athletes, on the other hand, are more prone to overuse injuries due to long-term high-intensity training and competition. It is very important to formulate personalized prevention and rehabilitation strategies for athletes with different sports levels, so as to reduce the risk of injury, improve sports performance and prolong sports life.

In this context, biomechanics, as a discipline to study the mechanical laws of objects (including human body), plays an important role in sports injury risk assessment and rehabilitation strategies. Biomechanics can help us to understand the mechanism of sports injury and identify high-risk factors that may lead to injury by analyzing the mechanical characteristics of human movement, such as the magnitude, direction, action point and kinematics parameters. For example, through the accurate measurement and analysis of joint stress, the possibility of joint injury under specific exercise or training mode can be predicted; Using biomechanical model to simulate the effects of different rehabilitation training programs is helpful to optimize the rehabilitation process, improve the rehabilitation efficiency and reduce the risk of re-injury. In addition, combined with big data analysis technology, a large number of biomechanical data can be deeply mined and analyzed, and potential injury risk patterns can be found, providing scientific basis for individualized sports injury prevention and rehabilitation. This interdisciplinary research method can not only improve the accuracy of sports injury risk assessment and the effectiveness of rehabilitation strategies, but also promote the innovative development of sports science and sports medicine.

Although the application prospect of big data in the field of sports injury is broad, the current research is still in the initial exploration stage, and there are many challenges and shortcomings. For example, how to effectively integrate and clean heterogeneous data from different sources, how to choose appropriate big data analysis models and algorithms to accurately assess the risk of sports injuries, and how to formulate and implement personalized rehabilitation strategies based on the results of big data analysis are all issues that need to be explored in depth. Therefore, this study explores sports injury risk assessment and rehabilitation strategies based on big data analysis, with special emphasis on the key role of biomechanics in this process. By integrating biomechanical principles and big data analysis technology, athletes and sports enthusiasts can be provided with more accurate and personalized sports injury prevention and rehabilitation services, thus ensuring their sports safety and health.

2. Risk assessment of sports injury based on big data

2.1. Data source

Collect the data of 1000 athletes aged between 16 and 35, covering different genders and sports, including amateur, semi-professional and professional athletes, to reflect the sports injury situation under the diversified training intensity and competition frequency. All participants passed the health examination before the start of the study to ensure that there were no major chronic diseases or serious previous sports injuries. Data sources not only include sports performance records obtained in cooperation with sports clubs and national teams, medical records obtained in cooperation with sports medical centers, lifestyle information of regular questionnaires, and sports and biofeedback data collected in real time through smart wearable devices, but also explicitly increase the collection of biomechanical data of athletes, such as ground reaction force, joint angle, muscle activity and so on. These data are obtained by motion capture system, dynamometer, electromyography and other equipment.

The study collected the basic information, exercise habits, historical injury records, sports performance data and biomechanical data of athletes from multiple sources, and ensured the effective use of these information through data preprocessing. This process includes data cleaning to remove duplicate, missing or abnormal values; Data conversion converts non-numerical data into numerical data; Feature selection selects variables highly related to sports injury risk according to statistical analysis, and especially emphasizes that biomechanical features, such as maximum joint mobility, muscle strength ratio, gait parameters, etc., should be included in the feature selection process besides basic information, exercise habits and historical injury records. Feature engineering transforms the original biomechanical data into useful features for sports injury risk assessment; and standardizing the feature variables to optimize the model performance. The data after pretreatment are shown in **Table 1**.

Table 1. Pre-processed sports performance index data.

Participant ID	age	gender	Sports	Training intensity	Competition frequency	Past injuries and illnesses	Ground reaction force (N)	Joint range of motion (°)	Muscle activation level (mV)	The fastest 100 m sprint time (s)	Maximum continuous running distance (m)	Damage occurs
A001	22	man	soccer	high	high	nothing	850	120	50	11.2	5000	no
A002	28	woman	basketball	secondary	secondary	have	700	110	45	12.5	4500	yes
A003	19	man	swim	low	low	nothing	650	130	55			no
...	...	...	...	...	...	...	...	...	...	...	...	...
A1000	35	woman	track and filed; athletics	high	high	have	900	115	60	10.8	6000	yes

2.2. Construction of risk assessment model

The sports injury risk assessment model based on big data needs to be able to handle high-dimensional data, capture nonlinear relationships and have strong generalization ability. Therefore, this study chooses RF (Random Forest) as the risk assessment model. RF is an integrated learning method. By constructing multiple decision trees and synthesizing their prediction results, it can effectively reduce the risk of over-fitting and improve the stability and accuracy of the model [12,13]. The principle of RF is shown in **Figure 1**.

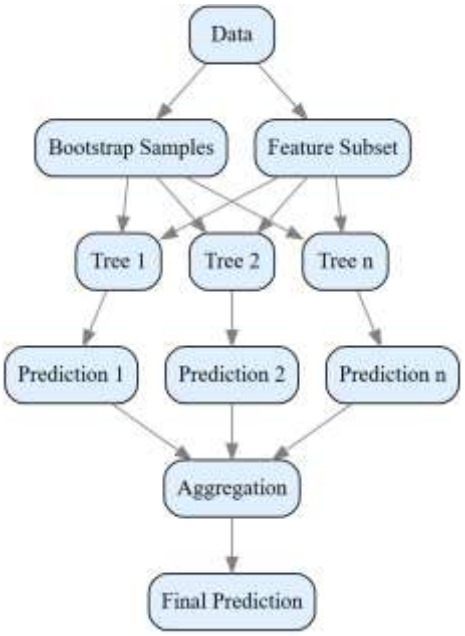

Figure 1. RF principle.

Using the preprocessed data set, the RF model is constructed. RF is an integrated model composed of multiple decision trees, and its basic construction formula can be expressed as:

$$\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b\,(x) \tag{1}$$

where $\hat{f}_{RF}(x)$ represents the prediction result of RF. B represents the number of decision trees. $\hat{f}_b(x)$ represents the prediction result of the b decision tree.

The construction process of decision tree involves extracting samples from the original data set by bootstrap sampling to form subsets, and then randomly selecting feature subsets at each node to split, and growing trees according to the criteria of information gain or Gini impurity until the preset stopping conditions such as maximum depth or minimum number of samples are reached [14].

In the RF model, each decision tree will classify the samples according to the characteristic variables, and finally get the prediction results through the voting mechanism. Assuming that the probability of an athlete's sports injury output by the model is P, the sports injury risk can be divided into different levels according to the value of P: low risk ($P < 0.3$), medium risk ($0.3 \leq P < 0.7$) and high risk ($P \geq 0.7$).

In order to express the risk assessment results more intuitively, a risk assessment index (RAI) is defined, and its calculation formula is as follows:

$$\text{RAI} = 100 \times P \tag{2}$$

where P is the probability of sports injury predicted by the model. The greater the value of RAI, the higher the risk of sports injury.

The RAI scale is shown in **Table 2**. The smaller the RAI value, the lower the risk of sports injury; The greater the RAI value, the higher the risk of sports injury. Through RAI scale, we can intuitively understand the risk level of athletes' sports

injuries, and provide basis for formulating corresponding prevention and rehabilitation strategies.

Table 2. RAI scale.

RAI value range	Risk grade of sports injury
$0 \leq RAI < 0.3$	Low risk
$0.3 \leq RAI < 0.7$	Medium risk
$RAI \geq 0.7$	high-risk

The parameters of RF model are optimized by grid search method. Grid search is an exhaustive method, and the optimal parameters are found by traversing all possible parameter combinations [15,16]. The formula can be expressed as:

$$BestParams = \underset{\theta \in \Theta}{\operatorname{argmin}}\, CVError\left(f_{RF}(x;\theta)\right) \tag{3}$$

where *BestParams* represents the optimal parameter combination. Θ represents the set of all possible parameter combinations. $CVError\left(f_{RF}(x;\theta)\right)$ represents the average error of RF model using parameter θ in cross-validation.

The performance of the model under different parameter combinations is evaluated by cross-validation, and the optimal parameter set is selected. The common K-fold cross-validation formula is expressed as:

$$CVError\left(f_{RF}(x;\theta)\right) = \frac{1}{k}\sum_{i=1}^{k} Error\left(f_{RF}(x_i;\theta)\right) \tag{4}$$

where k represents the number of folds. x_i represents the data set of the i fold. $Error\left(f_{RF}(x_i;\theta)\right)$ represents the error of RF model using parameter θ on the i -fold data set.

Finally, the optimized parameter set is used to train the RF model. The data set is divided into training set and test set, and the accuracy and generalization ability of the model are evaluated through the test set.

The following is the pseudo code of the RF modeling process:

Initialize RF model:

Input: training data set *D*, number of features *F*, number of trees *T*, maximum depth of trees *M*, minimum sample splitting number *S*, and size of random subset *k*

Initialize an empty forest collection forest = []

For $t = 1$ to T:
A random subset D_t with the same size as *D* is obtained by sampling with playback from the data set *D*.
Initialize an empty decision tree

Building decision tree:
Select an optimal feature for segmentation and the optimal segmentation point of this feature
Perform the following steps recursively until the stop condition is met (the depth of the node reaches *S* or the number of samples of the node is less than *M*):
k feature subsets are randomly selected for each feature.
Find the optimal segmentation point in the selected feature subset to minimize impurity
Segmentation of data sets according to optimal segmentation points

Recursively execute the process of building a decision tree for the segmented subset

Add the constructed decision tree to forest collection

Model training completed

Model prediction:
Input: sample x to be predicted, RF model forest

For every tree in forest:
Use tree to predict the sample x, and get a prediction result

Aggregate the prediction results of all trees to get the final prediction result

Return the final prediction result

2.3. Model optimization combined with biomechanical characteristics

In addition to considering the basic information of athletes, such as age, gender, sports events, exercise habits, such as training intensity, competition frequency and historical injury records, this paper also includes biomechanical characteristics, such as ground reaction, joint activity and muscle activation level. These biomechanical characteristics are obtained by motion capture system, dynamometer, electromyography and other equipment, which can directly reflect the mechanical characteristics of athletes in the process of sports.

In order to effectively integrate biomechanical characteristics with other characteristics to evaluate the risk of sports injury, firstly, data preprocessing is carried out, including cleaning, transforming and standardizing biomechanical data, and converting non-numerical data into numerical data; Secondly, feature selection, using statistical methods to select variables highly related to sports injury risk, especially those biomechanical features that have made significant contributions to risk assessment through correlation analysis and chi-square test; Finally, the feature engineering stage, where the original data are transformed into valuable features, such as calculating the coefficient of variation of joint activity and statistical data of muscle activation level, to better reflect the athletes' sports patterns and their potential injury risks.

In the RF model, the importance of each feature is measured by calculating the information gain or Gini impurity that the feature reduces in the process of model splitting. Through feature importance analysis, we can clearly understand which biomechanical features have important influence on sports injury risk assessment, thus providing scientific basis for model optimization.

Let T be a decision tree, S be the training set, S_t be the subset split by feature j, S_{t1}, S_{t2} be the two split subsets, $IG(S, S_{t1}, S_{t2})$ be the information gain, and N be the number of trees.

The importance I_j of the feature j is calculated by the following formula:

$$I_j = \sum_{t=1}^{N} \left(IG(S, S_{t1}, S_{t2}) \times \frac{|S_t|}{(S)} \right) \tag{5}$$

Among them:

$$IG(S, S_{t1}, S_{t2}) = Impurity(S) = \left(\frac{(S_{t1})}{|S|} \times Impurity(S_{t1}) + \frac{(S_{t2})}{|S|} \times Impurity(S_{t2}) \right) \quad (6)$$

For Gini impurity, the calculation formula is:

$$Impurity(S) = 1 - \sum_{k=1}^{k} p_k^2 \quad (7)$$

where p_k is the relative frequency of the class k in the set S.

For regression problems, the formula for calculating the mean square error (MSE) is:

$$Impurity(S) = \frac{1}{|S|} - \sum_{i \in S} (y_i - \bar{y}_i)^2 \quad (8)$$

where y_i is the target value of the i-th sample in the set S, and $\bar{y}_i$ is the average value of the target values in the set S.

The importance I_j of feature j is the average value of impurity reduction caused by feature j in all trees. In practical application, RF algorithm will calculate this value for each feature, and then rank the features according to this value, so as to get the importance ranking of the features. The result of feature calculation is shown in **Figure 2**.

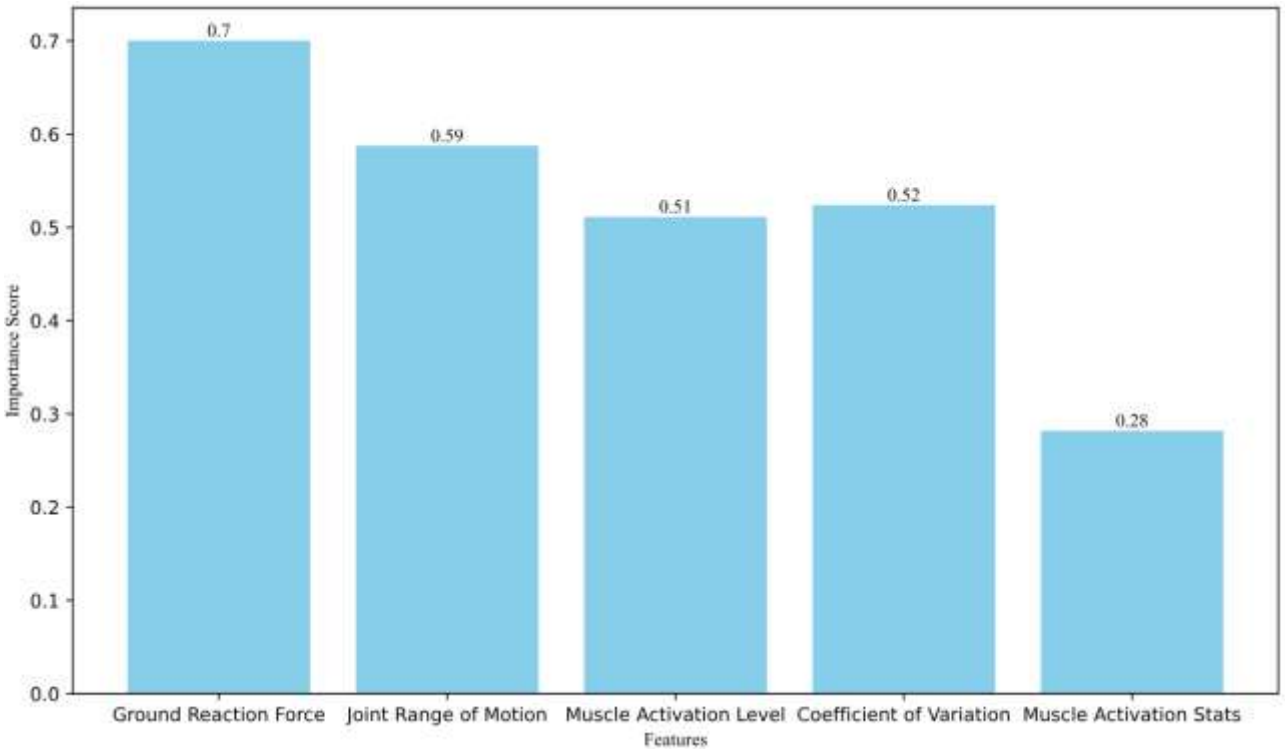

Figure 2. Characteristic calculation result.

The feature importance score calculated by RF model reveals the role of biomechanical features in predicting sports injury risk. These characteristics include ground reaction, joint activity, muscle activation level, its coefficient of variation and

statistical data, which respectively reflect the impact force, joint activity range, muscle activity degree and fluctuation of athletes. The importance score of a feature is measured based on its reduced information gain or Gini impurity in the model. The higher the score, the more critical the feature is to prediction.

In order to improve the accuracy of overall risk assessment, a separate prediction model based on biomechanics is constructed and integrated with RF model. A gait analysis model based on neural network is proposed, which can predict the risk of sports injury by analyzing the gait data of athletes.

The gait analysis model based on neural network is selected as the biomechanical prediction model (**Figure 3**). The model can deal with complex nonlinear relationships and capture subtle changes in gait data, thus accurately predicting the risk of athletes' sports injuries. Gait data are preprocessed to extract features related to sports injury risk, such as step size, step frequency, gait symmetry and so on. These features will be used as the input of the neural network model.

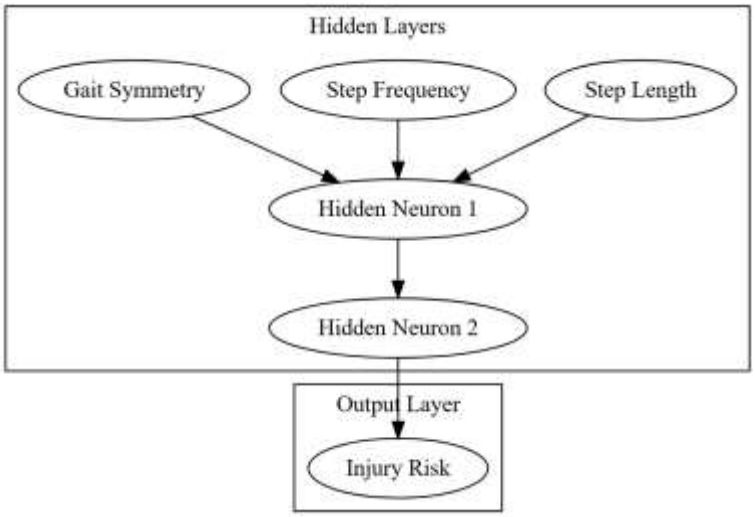

Figure 3. Gait analysis model structure based on neural network.

The neural network model is trained by using the preprocessed gait data, and the model performance is optimized by adjusting the model parameters. At the same time, cross-validation and other methods are used to evaluate the generalization ability of the model. The trained neural network model is fused with the RF model, and the fusion strategy adopts the weighted average method.

2.4. Model verification and application

The set-aside method is used to divide the data set into training set and test set, in which the training set is used for model training and the test set is used for model verification. The data of the test set is unknown in the process of model training, so the generalization ability of the model can be objectively evaluated [17]. Input the test set data into the trained RF model to obtain the prediction results of each sample.

By calculating the confusion matrix, the accuracy, precision, recall and F1 score of the model are obtained. **Table 3** shows four key performance indicators of model verification. The accuracy rate of 85% indicates that the prediction accuracy rate of the model is 85%. Precision92% shows that the model has high accuracy in predicting

positive classes; The recall rate of 78% indicates that the recognition ability of the model for all positive samples is relatively low, and there are cases of missed diagnosis; The F1 score of 84% is the harmonic average of precision and recall rate, which reflects that the overall performance of the model is good but there is still room for improvement, especially in improving the recall rate to enhance the recognition ability of positive samples.

Table 3. Model verification results.

index	value
Accuracy	0.85
Precision	0.92
Recall	0.78
F1 Score	0.84

K-fold cross-validation method is used to evaluate the generalization ability of the model. In this method, the data set is randomly divided into k subsets. In each iteration, k-1 subsets are used as training data, and the remaining subset is used as test data. This process is repeated for k times, and finally the average of k results is taken as the estimation of model performance. The results of K-fold cross-validation are shown in **Table 4**.

Table 4. K-fold cross-validation results.

Fold number	Accuracy	precision	Recall	F1 Score
1	84.5%	91.8%	77.5%	83.6%
2	86.2%	92.5%	79.1%	84.8%
3	85.8%	91.2%	78.3%	84.0%
4	83.9%	92.3%	76.8%	83.1%
5	85.1%	91.6%	77.9%	83.5%

Table 4 shows that the performance of the model is relatively stable in 5 iterations in K-fold cross-validation. The accuracy rate is maintained between 84% and 86%, with an average of 85%; Precision ranges from 91.2% to 92.5%, with an average of 92%; The recall rate is between 76.8% and 79.1%, with an average of 78%, indicating that there is a certain situation of missed diagnosis; F1 scores ranged from 83.1% to 84.8%, with an average of 84%. The model performs well in accuracy and precision, but there is still room for improvement in improving the recall rate to better identify positive samples.

The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are used to evaluate the classification performance of the model across different thresholds. The closer the ROC curve is to the upper left corner, and the larger the AUC value, the higher the reliability of the model. **Figure 4** shows the ROC curve of the model, with an AUC value of 0.92, indicating a high level of model reliability.

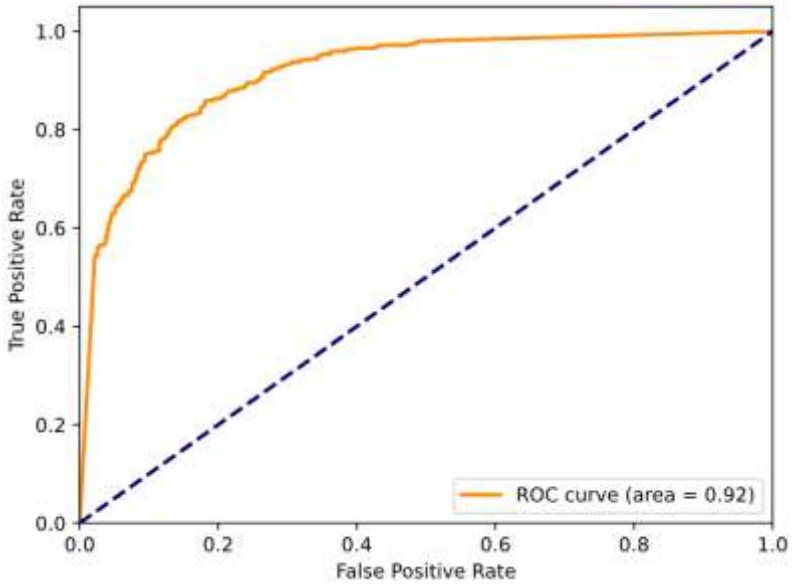

Figure 4. ROC curve of the model.

During exercise, the head and neck are less stressed and are not the main stress parts; However, the shoulders, elbows, spine, hips, knees and ankles are under great pressure, especially when the arms are used frequently and involve twisting, bending and load-bearing movements. The wrist is moderately stressed, depending on whether it is necessary to hold the instrument. This stress distribution shows (see **Figure 5**) that some parts such as shoulders, elbows, spine, hips, knees and ankles may face higher risk of injury, so special attention should be paid to the protection and targeted training of these high-risk parts in training and competition to reduce the possibility of injury.

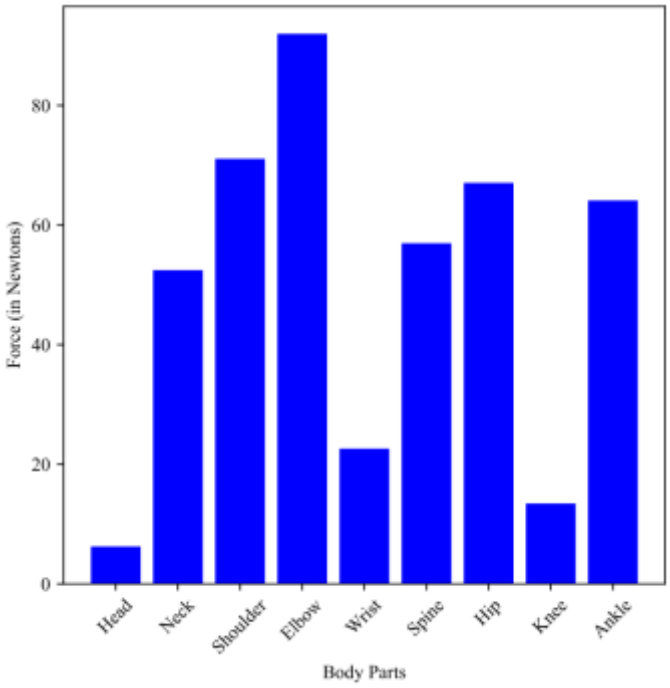

Figure 5. The stress situation of athletes in different parts during the competition.

Input the newly collected athlete data into the model, calculate the probability of each athlete's sports injury, and divide the athletes into three levels according to RAI: low risk, medium risk and high risk. The results of risk assessment and grading of athletes are shown in **Table 5**. According to the data table, athletes are divided into three risk grades according to the RAI value: athletes with low risk (RAI value close to 0) have ID numbers of 5 and 7, and can train as planned but need continuous monitoring; Athletes with medium risk (RAI value 0.3 to 0.7) have ID 1, 4, 6, 8 and 10, so it is suggested to adjust the training plan and strengthen preventive measures; Athletes with high risk (RAI close to 1) have ID numbers of 2, 3 and 9, so it is necessary to immediately reduce the amount of exercise and take professional rehabilitation and other intervention measures to prevent sports injuries.

Table 5. Risk assessment and grading results of athletes.

Athlete ID	age	gender	Sports	Training hours per week	Exercise years	Past injury times	Time of last injury (month)	Maximum speed (m/s)	Maximum endurance time (min)	RAI	risk level
1	30	man	swim	18.05	5	0	7	6.86	33.73	0.67	Medium risk
2	33	woman	run	19.68	8	4	9	4.60	41.18	0.97	high-risk
3	18	man	swim	16.99	4	1	3	2.31	26.90	0.88	high-risk
4	21	woman	run	11.92	3	4	6	7.07	43.74	0.51	Medium risk
5	21	woman	soccer	16.71	8	1	23	9.67	25.86	0.06	Low risk
6	25	man	run	6.77	3	2	11	7.22	48.92	0.45	Medium risk
7	27	man	soccer	14.60	1	2	14	7.08	57.48	0.02	Low risk
8	22	woman	swim	7.15	1	0	18	9.96	43.13	0.44	Medium risk
9	24	woman	basketball	19.17	5	1	0	6.65	10.68	0.98	high-risk
10	30	woman	soccer	12.83	6	1	14	5.31	41.14	0.36	Medium risk

Based on the results of risk assessment, individualized prevention strategies are formulated for athletes with different risk levels, especially for high-risk athletes. By analyzing their exercise habits and historical injury records, targeted measures such as strengthening specific muscle training, improving techniques and adjusting equipment are formulated. In addition, the dynamic monitoring and adjustment mechanism is implemented, the risk assessment is updated regularly and the prevention strategy is optimized according to the latest data to ensure its effectiveness and accuracy. In practical application, a professional sports team uses RF model for risk assessment, and takes preventive measures for high-risk athletes, including knee stability training and running posture adjustment, which ultimately significantly reduces the incidence of sports injuries and improves the overall performance of athletes.

The RF model based on big data analysis in this study shows high accuracy and reliability in sports injury risk assessment, which can provide athletes with personalized risk assessment results and prevention strategies, and is of great

significance for reducing the incidence of sports injuries and improving athletes' sports performance.

3. Rehabilitation strategy optimization based on big data

3.1. Analysis of rehabilitation needs based on biomechanics

In the analysis of rehabilitation demand based on biomechanics, the parameters such as joint stability and muscle strength balance are focused on to identify the key points in the rehabilitation process, and the relationship between biomechanical characteristics and rehabilitation demand is analyzed through association rule mining technology, which provides the basis for personalized rehabilitation plan [18]. Collect and classify a large number of data from hospitals and sports rehabilitation centers, including different types of sports injuries (such as muscle strain, ligament tear, etc.), injury degree, age, gender, weight and exercise habits of patients, so as to count common symptoms and recovery cycles, and analyze high-risk factors. Through cluster analysis of patients, we can identify groups with similar rehabilitation needs, such as acute injury of young athletes and chronic injury of the elderly. Finally, according to these analysis results, the specific rehabilitation needs of different types and degrees of injuries are determined, including physical therapy, nutritional intervention and psychological rehabilitation, and the core role of biomechanical evaluation in formulating rehabilitation strategies is emphasized to ensure that the plan not only targets at the injury itself, but also optimizes the biomechanical characteristics of patients.

3.2. Biomechanics-guided rehabilitation planning

Based on the evaluation of joint mobility and muscle strength, targeted rehabilitation actions such as stretching and strength training are designed, and biomechanical monitoring technology (such as wearable devices) is introduced to track the patient's rehabilitation progress in real time to ensure timely adjustment of the plan. Personalized rehabilitation strategies cover physical therapy, nutritional intervention and psychological rehabilitation, and use big data analysis to meet individual needs. Customize the physical therapy plan by evaluating the type and degree of injury, and predict the effect of patient response optimization; Provide personalized dietary advice according to nutritional status; Assess mental state, identify risks and intervene in advance. Biomechanical evaluation plays a central role in physical therapy, ensuring the safety and effectiveness of rehabilitation actions, and dynamically adjusting the plan with real-time monitoring technology in order to achieve the best rehabilitation effect. Integrate these three aspects of comprehensive rehabilitation programs, continuously track and adapt to patients' recovery through the big data platform, and provide all-round support.

3.3. Evaluation of strategy effect

In order to evaluate the effectiveness of the optimized rehabilitation strategy, a comparative experiment was designed, and 60 patients were randomly divided into two groups: the experimental group adopted a personalized rehabilitation plan based

on big data analysis, while the control group followed the traditional rehabilitation method. By setting rehabilitation cycle, recovery quality and patient satisfaction as evaluation indicators, data including rehabilitation progress, symptom improvement and patient feedback were collected regularly, and the rehabilitation effects of the two groups were analyzed and compared by statistical methods to determine whether the optimized strategy significantly improved rehabilitation efficiency and patient satisfaction.

The results showed that the average recovery period of the experimental group was 45.6 days and the standard deviation was 12.3 days, which was shorter and more concentrated than the average recovery period of 60.8 days and the standard deviation of 15.4 days in the control group. The median of 43 days in the experimental group is also lower than that in the control group, with the minimum and maximum values of 30 days and 70 days respectively. Compared with the control group's 45 days to 90 days, it shows that the experimental group not only has a shorter average and median rehabilitation cycle, but also has a smaller range of rehabilitation cycles, and the overall rehabilitation performance is more consistent and efficient. The statistical results of rehabilitation cycle are shown in **Table 6**.

Table 6. Statistical results of rehabilitation cycle of experimental combination control group.

group	Average recovery period (days)	Standard deviation (days)	Median (days)	Minimum value (days)	Maximum value (days)
Experimental group	45.6	12.3	43	30	70
Control group	60.8	15.4	62	45	90

The patients in the experimental group performed better than those in the control group in terms of functional recovery, pain relief and motor ability recovery after rehabilitation, indicating that personalized rehabilitation plan is more conducive to patients' comprehensive recovery. Through the investigation of patients' satisfaction, the satisfaction of patients in the experimental group was significantly higher than that in the control group, reflecting that the individualized rehabilitation plan better met the rehabilitation needs of patients (**Table 7**).

Table 7. Statistical results of patient satisfaction in experimental combination control group.

group	Average satisfaction score	Standard deviation (days)	Median (days)	Minimum value (days)	Maximum value (days)
Experimental group	4.3	0.8	4.5	3.0	5.0
Control group	3.6	1.2	3.7	2.0	5.0

According to the data provided, the average score of patients' satisfaction in the experimental group was 4.3, the standard deviation was 0.8, and the median was 4.5. Compared with the average score of 3.6, the standard deviation was 1.2 and the median score was 3.7 in the control group, it showed higher satisfaction and the score was more centralized and consistent. The satisfaction score range of the experimental group (3.0–5.0) is also narrower than that of the control group (2.0–5.0), which further

shows that the satisfaction of the patients in the experimental group is not only higher, but also less varied and more stable as a whole.

For some specific types of sports injuries (such as muscle strain, ligament tear, etc.), the rehabilitation effect of patients in the experimental group is particularly significant, indicating that the personalized rehabilitation plan is more targeted for such injuries. Although the individualized rehabilitation plan may involve more data analysis and customized services, in the long run, the overall rehabilitation cost of patients in the experimental group may be lower than that of the control group due to the improvement of rehabilitation efficiency and the shortening of rehabilitation cycle.

4. Discussion

Through the analysis method based on big data, this study deeply explored the risk assessment and rehabilitation strategy of sports injury. The research results show that after the individualized rehabilitation plan based on big data analysis is adopted in the experimental group, the rehabilitation cycle is significantly shortened, the quality of rehabilitation is significantly improved, and the patient satisfaction is also greatly improved. These results fully prove the significant advantages of big data analysis in sports injury risk assessment and rehabilitation strategies.

Big data analysis can integrate massive data from multiple sources, including patients' exercise habits, physiological indicators, historical injury records, etc., which provides a rich foundation for building an accurate risk assessment model. By mining the potential association between these data, high-risk groups and individuals can be identified more accurately, so as to take preventive measures in advance and reduce the probability of sports injuries [19]. At the same time, big data analysis can also customize personalized rehabilitation plans according to the specific conditions of patients to ensure the pertinence and effectiveness of rehabilitation measures. Big data analysis also played an important role in the formulation of rehabilitation strategies. By monitoring the patients' rehabilitation progress in real time, the rehabilitation plan can be adjusted in time to meet the patients' recovery needs. In addition, big data analysis can also predict the rehabilitation trend of patients and provide scientific decision support for doctors, thus improving the rehabilitation efficiency and the quality of life of patients. By combining biomechanical characteristics, such as joint stability and muscle strength balance, high-risk groups and individuals can be identified more accurately, and then more scientific and reasonable prevention and rehabilitation strategies can be formulated. Biomechanics research not only helps to understand the mechanism of sports injury, but also guides the design of rehabilitation training to ensure the safety and effectiveness of training.

Biomechanics research has broad application prospects in the field of sports injury prevention and rehabilitation, but it also faces some challenges. For example, how to collect and analyze biomechanical data more accurately, how to effectively translate biomechanical knowledge into clinical application, and how to overcome technical and cost obstacles so that the achievements of biomechanical research can benefit more athletes and ordinary people. Future research needs to continue to explore these issues in order to give full play to the potential of biomechanics in the prevention and rehabilitation of sports injuries.

This study provides scientific basis for sports injury risk assessment and rehabilitation strategy through big data analysis, which has the following practical significance: First, the model can identify high-risk athletes and potential injury types, and help coaches and trainers implement preventive measures in training plans; Secondly, support the formulation of individualized training and rehabilitation programs, adjust the intensity and recovery strategies according to individual conditions, improve the training effect and accelerate rehabilitation; Furthermore, provide decision support for health care professionals, assist in accurate diagnosis, treatment and monitoring rehabilitation progress; Finally, the research findings can be incorporated into education and training courses to improve the ability of future professionals to cope with sports injuries. These applications not only help to enhance the health and performance of athletes, but also provide strong support for relevant professionals and promote the overall progress of the sports industry.

Although this research has achieved remarkable results, there are still some limitations and challenges. Data quality is a key factor affecting the analysis effect of big data. In practical application, problems such as missing data, wrong data or inconsistent data may be encountered, which will affect the accuracy of risk assessment and the effectiveness of rehabilitation strategies. Therefore, how to improve the data quality and ensure the accuracy, integrity and consistency of the data is a problem that needs to be focused on. The generalization ability of the model is also a big challenge [20]. Although we have constructed effective risk assessment models and rehabilitation strategies in this study, whether these models and strategies can maintain the same effect in different groups of people and different sports still needs further verification and optimization. In order to improve the generalization ability of the model, it is necessary to collect more diversified data and adopt more advanced algorithms and technologies to build the model. Privacy protection and ethical issues are also challenges that big data analysis cannot ignore in sports injury risk assessment and rehabilitation strategies. When collecting, processing and using patient data, we must strictly abide by relevant laws, regulations and ethical norms to ensure that the privacy of patients is fully protected.

Future research directions include constructing a refined risk assessment model integrating multi-dimensional data such as genetics and psychology to improve the accuracy and pertinence of the assessment; Develop an intelligent rehabilitation system combining artificial intelligence and internet of things technology to realize real-time monitoring and personalized guidance of patients' rehabilitation progress and improve rehabilitation efficiency and quality of life; Explore the fusion analysis method of cross-disciplinary data such as sports science, medicine and biomechanics, deeply understand the mechanism of sports injury, and provide scientific support for risk assessment and rehabilitation strategies; At the same time, strengthen the research on privacy protection and ethical norms to ensure that the rights and interests of patients in big data applications are fully guaranteed.

5. Conclusion

By using the big data analysis method, the study deeply explored the risk assessment and rehabilitation strategy of sports injury. The research shows that the

risk assessment of sports injury by using RF model has high accuracy and reliability, and can provide personalized risk assessment results and preventive strategies for athletes, significantly reducing the incidence of sports injuries and improving athletes' sports performance. In addition, the optimization of rehabilitation strategy based on big data shows that the rehabilitation period of patients in the experimental group is significantly shortened, and the quality of recovery and patient satisfaction are greatly improved after receiving personalized rehabilitation plan. These achievements fully prove the significant advantages of big data analysis in sports injury risk assessment and rehabilitation strategies. This study also emphasizes the importance of biomechanical research in sports injury risk assessment and rehabilitation strategy formulation. By integrating biomechanical data, such as joint stability and muscle strength balance, high-risk groups and individuals can be identified more accurately, so as to formulate more scientific and reasonable prevention and rehabilitation strategies. Biomechanics research not only enhances the accuracy of risk assessment, but also provides a scientific basis for the design of rehabilitation training, ensuring the safety and effectiveness of training. However, despite remarkable achievements, there are still some limitations and challenges in this study. Data quality, model generalization ability, privacy protection and ethical issues are issues that need to be focused on. In addition, the collection and analysis of biomechanical data also face technical challenges, and further research and innovation are needed to overcome these obstacles. Future work will be devoted to solving these problems, so as to further improve the accuracy and effectiveness of sports injury risk assessment and rehabilitation strategies.

In our research, it is suggested that future research can include the following aspects:

1) Carry out long-term follow-up research to monitor the changes of sports injury risk of athletes in different training stages and competition cycles. This will help to verify the stability and accuracy of our model in practical application.
2) Expand the research scope, including athletes from different regions and different sports, so as to improve the universality and representativeness of the research results.
3) Explore the use of deep learning techniques, such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), to process complex sports data, so as to improve the accuracy of sports injury risk assessment.
4) Based on the results of big data analysis, customize a personalized rehabilitation program for each athlete to minimize the risk of re-injury.
5) Combined with the physiological, psychological and social factors of athletes, the prevention strategies are optimized to reduce the incidence of sports injuries.

Author contributions: Conceptualization, JC and HC; methodology, JC; software, JC; validation, JC and HC; formal analysis, JC; investigation, JC; resources, JC; data curation, HC; writing—original draft preparation, HC; writing—review and editing, JC; visualization, JC; supervision, JC; project administration, JC; funding acquisition, HC. All authors have read and agreed to the published version of the manuscript.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Tuakli-Wosornu YA, Grimm K, Macleod JG. Expanding sports injury prevention to include trauma and adversity. British journal of sports medicine. 2022; 56(15), 835–836.
2. Theisen, Malisoux, Genin, Delattre, Seil, Urhausen. Influence of midsole hardness of standard cushioned shoes on running-related injury risk. British journal of sports medicine. 2020; 48(5), 371–376.
3. Mausehund L, Krosshaug T. Knee biomechanics during cutting maneuvers and secondary acl injury risk: a prospective cohort study of knee biomechanics in 756 female elite handball and soccer players. The American Journal of Sports Medicine. 2024; 52(5), 1209–1219.
4. Anonymous. Return to the pre-injury level of sport after anterior cruciate ligament reconstruction: a practical review with medical recommendations. International journal of sports medicine. 2024; 45(8), 572–588.
5. Andrade R, Wik EH, Rebelo-Marques A, Blanch P, Whiteley R, Espregueira-Mendes J, et al. Is the acute: chronic workload ratio (acwr) associated with risk of time-loss injury in professional team sports? a systematic review of methodology, variables and injury risk in practical situations. Sports Medicine. 2020; 50(9), 1613–1635.
6. Fitzpatrick JD, Chakraverty R, Patera E, James SLJ. Is there a need to reconsider the importance of myoaponeurotic injury within the nomenclature of sports-related muscle injury?. British journal of sports medicine. 2022; 56(23), 1328–1330.
7. Bache-Mathiesen LK. Improving statistical methodology in training load and injury risk research (phd academy award). British journal of sports medicine. 2023; 57(21), 1403–1404.
8. Wik EH. Injuries in elite male youth football and athletics: growth and maturation as potential risk factors (phd academy award). British journal of sports medicine. 2023; 57(21), 1405–1406.
9. Pang J, Li X, Zhang X. Coastline land use planning and big data health sports management based on virtual reality technology. Arabian Journal of Geosciences. 2021; 14(12), 1–15.
10. Li C, Cui J. Intelligent sports training system based on artificial intelligence and big data. Mobile Information Systems. 2021; 2021(1), 1–11.
11. Yang T, Yuan G, Yan J. Health analysis of footballer using big data and deep learning. Scientific Programming. 2021; 2021(2), 1–8.
12. Chia L, Fuller CW, Taylor D, Pappas E. Mastering the topic, the message, and the delivery: leveraging the social marketing mix to better implement sports injury prevention programs. The Journal of orthopaedic and sports physical therapy. 2022; 52(2), 55–59.
13. Keays SL, Mellifont DB, Keays AC, Stuelcken MC, Lovell DI, Sayers MGL. Long-term return to sports after anterior cruciate ligament injury: reconstruction vs no reconstruction—a comparison of 2 case series. The American Journal of Sports Medicine. 2022; 50(4), 912–921.
14. Myklebust G, Funnemark K, Moseid CH. Closing the gap on injury prevention: the oslo sports trauma research centre four-platform model for translating research into practice. British Journal of Sports Medicine. 2022; 56(9), 482–483.
15. Chalmers PN, Mcelheny K, D'Angelo J, Rowe D, Ma K, Curriero FC, et al. Effect of weather and game factors on injury rates in professional baseball players. The American Journal of Sports Medicine. 2022; 50(4), 1130–1136.
16. Howell DR, Seehusen CN, Carry PM, Walker GA, Reinking SE, Wilson JC. An 8-week neuromuscular training program after concussion reduces 1-year subsequent injury risk: a randomized clinical trial. The American Journal of Sports Medicine. 2022; 50(4), 1120–1129.
17. Butler LS, Janosky JJ, Sugimoto D. Pediatric and adolescent knee injuries: risk factors and preventive strategies. Clinics in sports medicine. 2022; 41(4), 799–820.
18. Edouard P, Ruffault A, Bolling C, Navarro L, Martin S, Frédéric.Depiesse, et al. French athletics stakeholders' perceptions of relevance and expectations on injury prevention. International journal of sports medicine. 2022; 43(12), 1052–1060.
19. Jauhiainen S, Kauppi JP, Leppnen M, Pasanen K, Parkkari J, Vasankari T, et al. New machine learning approach for detection of injury risk factors in young team sport athletes. International journal of sports medicine. 2021; 42(2), 175–182.
20. Biese KM, Winans M, Fenton AN, Hernandez M, Schaefer DA, Bell DR. High school sport specialization and injury in collegiate club-sport athletes. Journal of athletic training. 2021; 56(12), 1271–1277.