Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

Explainable Prediction of Non-Contact Injury Risk Using Machine Learning

Xiaohong Ye[1#*], Yuanqi Huang[2#]
[1]Chengyi University College, Jimei University, Xiamen 361021, China
[2]School of Physical Education and Sport Science, Fujian Normal University, Fuzhou 350117, China
[#]These authors contributed equally to this work and should be considered co-first authors.
*Corresponding Author.

Abstract:

Injuries not only hinder players' competitive performance and increase team medical expenses, but also affect players' long-term or even lifelong physical activity. The use of machine learning algorithms to model the relationship between an player's training data and risk of injury can help to assess an player's risk of injury and provide a basis for decision making on training load adjustment. However, existing research reports mostly focus on the prediction accuracy improvement of the model while ignoring the explainability and reliability issues of the model, which makes the prediction model in practical application rate reduced. Therefore, this study collected data on the training load, subjective perceived wellness, menstrual status, athletic ability test and injury information of 18 young female basketball players in Fujian Province during their preparation for the 2021 National Games of the People's Republic of China. Multiple machine learning algorithms were used to construct non-contact injury risk prediction models for young female basketball players. Experiments show that the model based on eXtreme Gradient Boosting (XGBoost) has the best performance and can detect approximately 89.4% of non-contact injury risk with a prediction of 66.6%, outperforming other counterparts. Four factors influencing the risk of non-contact injury were screened by the model's importance variable analysis, and the results were highly consistent with research reported in sports science, demonstrating the reliability of the model. The application of the model can provide a reliable basis for decision making in sports injury prevention practice.

Keywords: *Machine learning, explainability, Sports injury, Training load, Analysis of risk factors.*

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

I. INTRODUCTION

With the continuous improvement of the competitive level of modern basketball, higher requirements are put forward for the physical quality of basketball players. Basketball players need to obtain sufficient stimulation through progressive load training and develop sports ability and physical quality to meet the needs of special sports. Investigation shows that non-contact injury (NCI) accounts for 47.0% of basketball players' training injuries [1]. The incidence of NCI in the center position is as high as 86.1%, and 28.0% of the NCI will lead to the loss of more than 7 days of training time [2]. Injuries not only hinder the development of players' competitive level, increase the medical expenses of teams, but also affect the long-term and even life-long physical activities of players. Gabbett et al. [3] made a long-term longitudinal study on the training load, subjective perceived wellness, sports ability and the NCI risk of players in different events and found that unreasonable training load arrangement is an important reason for the high incidence of NCI. Therefore, the method of injury risk assessment based on data monitoring and analysis is undoubtedly an important way to early identify sports injury risk [4].

In recent years, many scholars have explored the law of injury occurrence through the observation data in training practice, trying to establish the prediction model of sports injury risk by using the parametric model (such as logical regression) in classical statistics, so as to achieve the purpose of early identification of sports injury [5-8]. However, there are many problems in real world research data, such as large amount of data, many influencing factors, complex data structure, interaction among variables, multicollinearity and unbalanced sample distribution [9]. The classical statistical modeling methods are not enough in the discussion of injury risk factors, such as the model is not fit, the accuracy is low, the complex nonlinear relationship between variables can not be identified, and the nonlinear characteristics need to be transformed. This hinders the development of injury risk assessment methods based on data monitoring and analysis, The prevention strategy of sports injury is still based on experience rather than data [10]. With the development of machine learning and data science, it is possible to predict the risk of sports injury by using machine learning model. Rossi et al. [11] pointed out that modeling and analyzing the relationship between an player's training data and the risk of sports injury using machine learning algorithms can help to assess an player's injury risk and can provide a basis for making decisions about an player's training load adjustment.

Rossi et al. [6] used Decision Tree (DT), Random Forest (RF) and Logistic Regression (LR) algorithm to predict 26 Italian football players' NCI. The results of 2-fold cross-validation of the model showed that the DT could detect about 80% of NCI with about 50% accuracy. LR can detect only about 60% of NCI with an accuracy of about 18%. Bryan et al. [7] used

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

algorithms such as eXtreme Gradient Boosting (XGBoost) and LR to model the injury prediction of 2322 hockey players and used 10-fold cross-validation to assess their performance. The results show that XGBoost is superior to LR in predicting injury risks for players and goalkeepers at all locations next season. However, the application of machine learning algorithms in injury risk prediction is relatively small, and most of the studies are based on seasonal technical and tactical analysis data and pre-game screening data for prediction and modeling, and their results have low reference value for injury risk prediction during actual sports training, as well as poor interpretability and other problems. Therefore, the performance of machine learning algorithms in training practice data needs to be further investigated.

Therefore, through the use of the training monitoring and injury registration database of Fujian young women basketball players during the preparation for the 14th National Games of the People's Republic of China, this study intends to use the data science method and machine learning algorithm to construct the NCI risk prediction model of young women basketball players. Using the prediction model, this paper discusses the important influencing factors of the NCI risk of young women basketball players in Fujian Province, analyzes the influence of the interaction between important factors on the NCI risk, and makes a preliminary discussion on the prevention of NCI of young women basketball players combined with training practice. The purpose of this paper is to provide ideas for the future sports data analysis, sports training support, sports injury prevention and so on.

II. MATERIALS AND METHODS

2.1 Materials

Our research uses the training monitoring and injury registration data of Fujian young female basketball players during the preparation of the 14th National Games of the People's Republic of China (November 2020 to April 2021) to construct a data set. The data set contains the general information, training load, subjective perceived health, athletic ability test and injury status of 18 female basketball players in Fujian Province.

2.1.1 Training load

The training load is collected every day. The training load quantification method uses session Ratings of Perceived Exertion (sRPE) to quantify. The Ratings of Perceived Exertion (RPE) questionnaire designed by Foster et al. [12]. It was used to inquire about the player's subjective fatigue after each training. According to formula (1), calculate the sRPE by

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

multiplying the PRE score and duration. The validity and reliability of this quantitative method have been confirmed in a large number of studies [13].

$$sRPE = duration \times RPE \tag{1}$$

Considering the cumulative effect of training load on the duration of physical stimulation, the authors refer to the recommendations of Williams et al. [14]. According to formula (2), the authors use Exponential Weighted Moving Average (EWMA) calculation method to calculate the Acute Load (AL) and Chronic Load (CL) of each player.

$$EWMA_{today} = \frac{2}{N+1} \times \left(Load_{today}\right) + \left(1 - \frac{2}{N+1}\right) \times EWMA_{yesterday} \tag{2}$$

According to the study, use formula (3) calculate the Training Monotonicity (TM) and formula (4) to calculate the Acute: Chronic Workload Ratio (ACWR) of each player.

$$TM = \frac{\frac{1}{7}\left(\sum_{i=1}^{7} Load_i\right)}{\sqrt{\frac{1}{7}\sum_{i=1}^{7}\left(Load_i - \frac{1}{7}\left(\sum_{i=1}^{7} Load_i\right)\right)^2}} \tag{3}$$

$$ACWR = \frac{WAL}{WCL} \tag{4}$$

2.1.2 Subjective perception of wellness

Subjectively perceived wellness is collected every day like training load. Refer to the consensus statement on exercise training load monitoring [15]. The wellness status survey questionnaire designed by Hooper et al. [16] is used to quantify subjectively perceived wellness. The scale uses Likert's 5 grades to score, including five aspects: fatigue, sleep, Muscle Soreness (MS), stress and desire. Each item is assigned a value ranging from very poor recovery to very good recover with 1 to 5 points respectively. The sum of the player's self-reported fatigue, sleep, MS, stress, and desire is the Hooper Index (HI). This scale is widely used to quantify the training response of player and has good reliability and validity [17]. This study tested the reliability of the scale through preliminary investigation. After

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

testing, the Cronbach coefficient α of the scale is 0.804, which has high reliability.

Considering the accumulation and attenuation characteristics of subjective perceived wellness over time, calculate the acute accumulation and chronic accumulation of players' subjective perceived wellness according to formula (2). Including weekly cumulative fatigue (WFaitgue), weekly cumulative sleep (WSleep), weekly cumulative muscle soreness (WMS), weekly cumulative stress (WStress), weekly cumulative desire (WDesire) and weekly cumulative HI (WHI); 4 weeks cumulative fatigue (4WFatigue), 4 weeks cumulative sleep (4WSleep), 4 weeks cumulative muscle soreness (4WMS), 4 weeks cumulative stress (4WStress), 4 weeks cumulative desire (4WDesire), and 4 weeks cumulative HI (4WHI).

2.1.3 Athletic ability test

According to the data provided by the Basketball and Volleyball Management Center, taking into account the athletic ability test can reflect the characteristics of basketball. From the data, the authors chose 1RM test for squat, 1RM test for bench press, 5.8-meter 6-round shuttle run, maximum vertical jump (MVJ) test. Among them, the squat 1RM test is used to quantify the maximum muscle strength of the lower limbs; the bench press 1RM test quantifies the maximum muscle strength of the upper limbs of an player; The 6 rounds shuttle run test of 5.8 meters quantifies the players' rapid start, emergency stop and conversion capabilities; Touch height by rmaximum vertical jump test is used to assess the explosive power and coordination of the lower limbs.

Each test is regarded as the players' athletic ability over a 5-week period, and data is collected every 5 weeks for a total of 4 times. In order to facilitate the horizontal comparison of the athletic ability test indicators of different individuals. The authors divide the player's squat 1RM weight and bench press 1RM weight by the body weight to calculate the relative strength of the squat and the bench press.

2.1.4 Injury information

With reference to the injury data collection procedure of Fuller et al. [18], the injury was diagnosed by medical staff of the Fujian Basketball Volleyball Management Center through physical examination and other methods. Record the nature of the injury, diagnosis method and other information according to the injury registration form. In this study, NCI was defined as damage caused by mechanisms other than direct contact with any part of the body (including overuse injury or chronic injury). Studies have shown that there may be a time lag between the peak fluctuation of training load and the increase in injury risk [19]. Therefore, the NCI risk is

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

defined as whether there is an NCI next week (0 is negative/no NCI; 1 is positive/NCI), and the NCI risk is used as the dependent variable for the study. TABLE I provides a description of the characteristics considered.

TABLE I. Dataset feature description

FEATURE	DESCRIPTION
AGE	The number of years from the player's date of birth to today
POSITION	The position of a player on the field
SPORT LEVEL	The skill level of an player
HEIGHT	Player's height (in meters)
WEIGHT	Player's weight (in kg)
BMI	Body Mass Index: ratio between weight and the square of height
TRAINING YEARS	The number of years an player has trained
MENSTRUAL	Whether the player is menstruating
TM	The ratio between the mean and standard deviation of the training load in the past 7 days
AL	Average training load over the past 1 week
CL	Average training load in the past 4 weeks
ACWR	The ratio between the AL and CL
WFATIGUE	Average fatigue over the past 1 week
WSLEEP	Average sleep quality over the past 1 week
WMS	Average muscle soreness over the past 1 weeks
WSTRESS	Average stress level over the past 1 week
WDESIRE	Average desire over the past 1 week
WHI	Average Hooper Index (HI) over the past 1 week
4WFATIGUE	Average fatigue over the past 4 weeks
4WSLEEP	Average sleep quality over the past 4 weeks
4WMS	Average muscle soreness over the past 4 weeks
4WSTRESS	Average stress level over the past 4 weeks
4WDESIRE	Average desire over the past 4 weeks
4WHI	Average HI over the past 4 weeks
SQUAT	The ratio of the player's squat 1RM weight to body weight
BENCH	The ratio of the player's bench 1RM weight to body weight
SHUTTLE	The time for the player to complete the 6 rounds of 5.8 meters shuttle

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

RUN TIME	run
MVJ	The height difference between the MVJ height and the ground

2.2 Pre-processing

Data pre-processing includes:

(1) Missing Value Analysis and Processing: Analyzing the missing values in the collected data, using sequence mean interpolation at the individual level for missing values of less than 10%, and eliminating missing values of more than 10%.

(2) Standardization: Z-score standardization of continuous feature data with different dimensions, eliminating the dimension differences between features to ensure data comparability. In order to explore the co-changing relationship between NCI risk and various indicators for young women basketball players.

(3) Data integration: In the study, the data of general information, training load data, subjective perceived wellness data of players and sports ability test data are integrated and processed with training day as the unit of time interval to improve the utilization of data information resources. Due to the computational requirements of Training Monotony (TM) and cumulative indices, the first 4 weeks of the dataset are excluded.

(4) Feature selection: Feature selection before machine learning classification reduces the dimension of the feature space, thereby reducing the risk of overlap. The specific method is to use the the Generalized Estimation Equation (GEE) to analyze the correlation between features and dependent variables for feature selection. Variance Inflation Factor (VIF) was used to analyze the multicollinearity between features. The larger the VIF value, the greater the possibility of collinearity between features. Based on the collinearity diagnostic criteria, features with tolerance less than 0.1 and VIF greater than 10 were excluded. The final remaining features and their collinearity diagnostic results are shown in TABLE II.

TABLE II. Colinearity diagnostic check results

FEATURE	TOLERANCE	VIF
SPORT LEVEL (X1)	0.656	1.525
TM (X2)	0.636	1.566

ACWR (X3)	0.705	1.419
4WFATIGUE (X4)	0.289	3.462
4WSLEEP (X5)	0.494	2.024
4WSTRESS (X6)	0.388	2.577
4WDESIRE (X7)	0.679	1.472
MVJ (X8)	0.985	1.015

(5) Unbalanced processing: the dataset contains 172 damage samples and 1642 non-damage samples. There is a problem of data class imbalance. Because of data imbalance, the model can not classify a few samples correctly. In order to reduce the negative impact of data imbalance on model training, Synthetic Minority Oversampling Technique (SMOTE) [20] technology was used to synthesize samples for a few classes in the training set.

2.3 Model Construction

In this study, parameterized model and non-parametric model were used to model the NCI of Fujian young female basketball players. The parameterized model uses the widely used LR model, the non-parametric model uses the DT, RF and XGBoost algorithm, which are widely reported in the literature. A grid search method based on stratified 10-fold cross-validation was used to optimize the model for hyperparameters.

III. RESULTS AND DISCUSSIONS

3.1 Evaluation Method

In this study, the dataset is randomly divided into training set and test set at a 9:1 scale, the model is built and the internal validity of the model is evaluated using stratified 10-fold cross-validation in the training set, and the performance of the optimal model is validated using the model test set to evaluate the external validity of the model. Since the cost of missed diagnosis of injury is much higher than that of misdiagnosis of injury in the detection of injury, it is not appropriate to use only accuracy as an evaluation index of the model. In order to better evaluate the classification performance of the model, the precision, recall, F2 score and Area Under receiver operating characteristic Curve (AUC) were selected as indicators to measure model performance.

In practice, accuracy and recall rates are mutually constrained. When accuracy is high, recall rates tend to be low. In order to balance precision and recall, F2 score was used to

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

reconcile. Since the cost of missed diagnosis of injury is much higher than that of misdiagnosis of injury, it is necessary to improve the recall rate as much as possible while ensuring the accuracy. The F2 score fully reflects the overall prediction performance of the prediction model on most and a few samples. The horizontal and vertical coordinate axes of the receiver operating characteristic Curve (ROC) represent the False Positive rate (FP) and True Positive rate (TP), respectively. The closer the ROC curve is to the upper left corner, the better the performance of the classifier. However, when comparing different classification models, using the ROC curve directly cannot quantitatively compare the classification performance of different classification models. Therefore, AUC is used as a quantitative indicator.

The formulas for calculating the accuracy and recall rates are as follows, where FN is false negative rate and FP is false positive rate:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

The formula for calculating the F2-score is as follow:

$$F2 = \frac{5 \times Precision \times Recall}{4 \times Precision + Recall} \tag{7}$$

3.2 Performance Evaluation

The characteristics in TABLE II were used as independent variables, and the NCI risk was used as the dependent variable to construct a predictive model. LR, DT, RF and XGBoost were used to model the NCI risk of young female basketball players.

3.2.1 Internal validity assessment

A stratified 10-fold cross-validation was used to evaluate the internal validity of the model. The Kruskal-Wallis test was performed on the stratified 10-fold cross-validation performance evaluation results of the model. In TABLE III, the results show that there are very significant differences between the precision, recall, F2 and AUC indicators of different models ($p < 0.01$). Compared with LR and DT, the precision, recall, F2 and AUC indexes of RF and XGBoost

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

were significantly higher than LR and DT ($p < 0.05$). Compared with RF, although XGBoost's recall, F2, and AUC indicators increased by 6.56%, 3.09%, and 2.44%, respectively, and the Precision indicator decreased by 7.84%, the performance difference between XGBoost and RF did not reach statistical significance ($p > 0.05$). The ROC curve of XGBoost is shown in Fig 1. In the actual situation of damage detection, the cost of missed diagnosis of damage is much higher than the misdiagnosis of damage, and the recall and F2 indexes of XGBoost are better than RF, so the XGBoost model is considered to have better performance and internal validity.

TABLE III. Analysis of variance of model performance evaluation results

MODEL	PRECISION	RECALL	F2	AUC
LR	0.1646 ± 0.0276	0.6248 ± 0.1403	0.4003 ± 0.0786	0.6642 ± 0.0655
DT	0.6097 ± 0.1043	0.7395 ± 0.1539	0.7071 ± 0.1368	0.8473 ± 0.0798
RF	0.7229 ± 0.0933	0.8381 ± 0.0838	0.8099 ± 0.7192	0.9033 ± 0.0428
XGBOOST	0.6662 ± 0.0686	0.8943 ± 0.0580	0.8349 ± 0.0440	0.9253 ± 0.0284
p	0.000***	0.000***	0.000***	0.000***

*$p<0.1$, **$p<0.05$, ***$p<0.01$

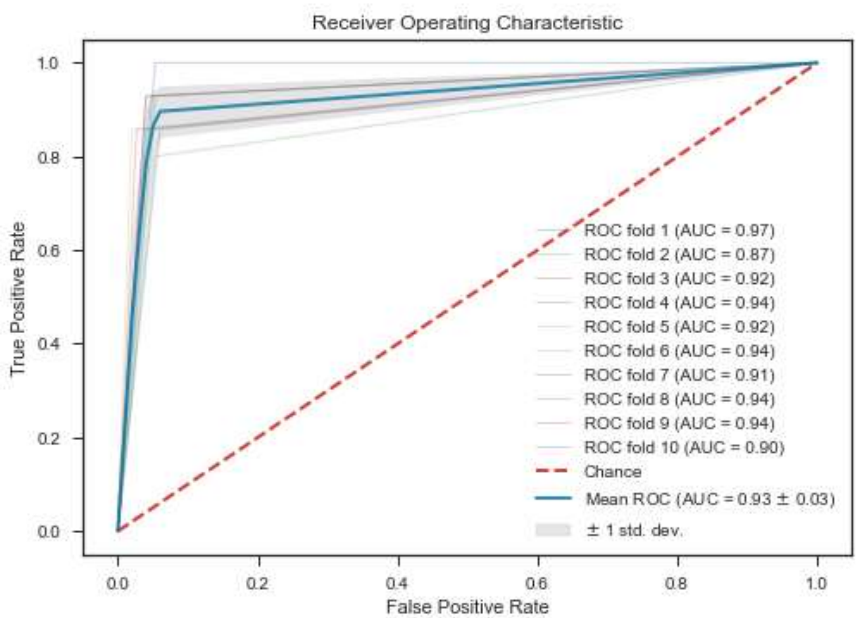

Fig 1: ROC curve of XGBoost.

3.2.2 External validity assessment

The test set was used to evaluate the external validity of XGBoost. The results showed that the precision, recall, F2, and AUC of XGBoost in the test set were 0.6522, 0.9375, 0.8621, and

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

0.9447, respectively, which were highly consistent with the stratified 10-fold cross-validation results. The Pearson $\chi 2$ goodness-of-fit test showed that the predicted value of the XGBoost model is not significantly different from the true value ($\chi 2$=2.4386, p=0.1184), so the XGBoost model is considered to have good external validity.

The above results show that in the LR, DT, RF and XGBoost models, XGBoost has good internal validity, external validity and fit, and can effectively identify the NCI risk of teenage female basketball players. The model can detect about 89.4% of the NCI risk with a prediction accuracy of about 66.6%, which is better than Rossi et al. [6]. At the same time, the authors found that DT, RF and XGBoost are significantly better than LR models in prediction performance. The results are consistent with previous research reports, indicating that the damage prediction problem may not be linearly separable, which proves that the non-parametric model can be used to predict the injury and effectively identify the NCI risk of players.

3.3 Analysis of important features

In this study, the global imputation analysis method in the Shapley Additive exPlanations (SHAP) explanatory framework was used to perform imputation analysis on the XGBoost model to assess the impact of features on NCI [21]. The distribution of the variable values and the corresponding SHAP values are shown in Fig 2A. The variable weight clustering analysis is shown in Fig 2B. When SHAP value > 0 indicates the possibility that the variable value increases the risk of NCI, and vice versa the possibility that decreases the risk of NCI.

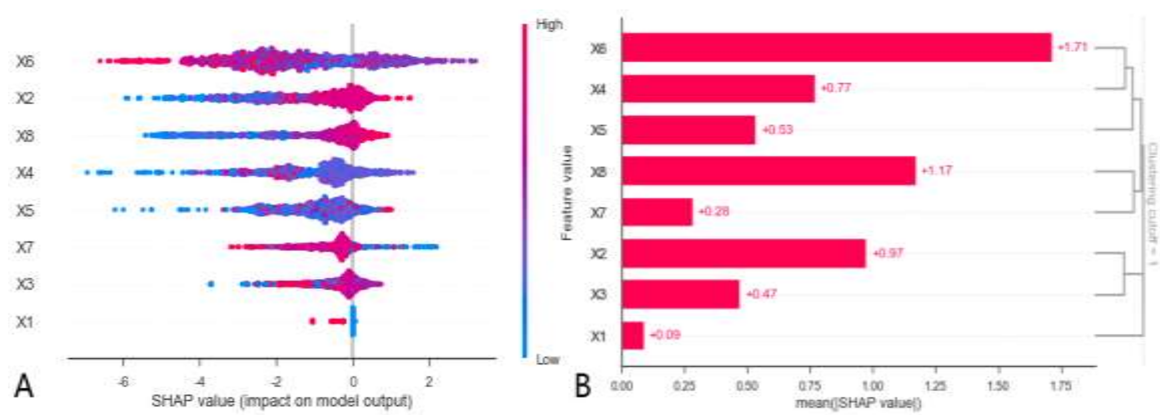

Fig 2: Characteristic attribution analysis based on SHAP method. (A) Scatter diagram of variable numerical distribution; (B) Hierarchical clustering of variable weights.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

The results showed that the distribution of SHAP values corresponding to most of the characteristic variables showed a clear distinction, and the distribution of SHAP values corresponding to some of the characteristic variables did not show a clear distinction (Fig 2A). This indicates that the injury prediction problem is not linearly differentiable, and this result is consistent with Talukder's view [5].

The variable weights of the observed indicators were analyzed by hierarchical cluster analysis, and the variables with higher weights in each category were considered as important variables. The results showed that TM, 4WFatigue, 4WStress, and MVJ had the highest weights of 0.97, 0.77, 1.71, and 1.17, respectively, indicating that TM, 4WFatigue, 4WStress, and MVJ are important factors influencing the risk of NCI in young female basketball players in Fujian Province. The results are highly consistent with those reported in sports science studies [22]. It confirmed the good interpretability and reliability of the model constructed in this study.

3.4 Impact of important feature on NCI risk

The marginal effects of TM, 4WFatigue, 4WStress, and MVJ on NCI risk showed a nonlinear variation relationship, as shown in Fig 3. $E\left[f\left(x\right)\right]$ in the figure is the model prediction threshold for NCI risk, and when the model prediction is higher than this value, i.e., the variation increases the likelihood of NCI risk, and vice versa, it decreases the likelihood of NCI risk. Using the marginal effects plot, the authors determined that the threshold values of TM, 4WFatigue, 4WStress, and MVJ for youth female basketball players in Fujian Province were 1.55 AU, 2.98 AU, 2.99 AU, and 107.92 cm. Relative risk was estimated when important characteristics were high or below the critical values. The results showed that when TM was higher than 1.55 AU (Fig 3A), the risk of NCI was approximately 1.727 times higher (RR=1.727, 95% CI=1.238-2.409, p=0.001); when MVJ was higher than 107.92 cm (Fig 3B), the risk of NCI was approximately 5.942 times higher (RR=5.942, 95% CI=3.325 ~ 10.619, p<0.001); when 4WFatigue was lower than 2.98 AU (Fig 3C), the risk of NCI was about 1.828 times (RR=1.828, 95% CI=1.357-2.461, p<0.001); when 4WStress was lower than 2.99 AU (Fig 3D), the risk of NCI was about 3.531 times (RR=3.531, 95% CI=2.605 ~ 4.785, p<0.001). The coach can use the above calculated cut-off point and relative risk to gain a qualitative understanding of the player's NCI risk and make timely adjustments to the player's training intensity and training volume.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

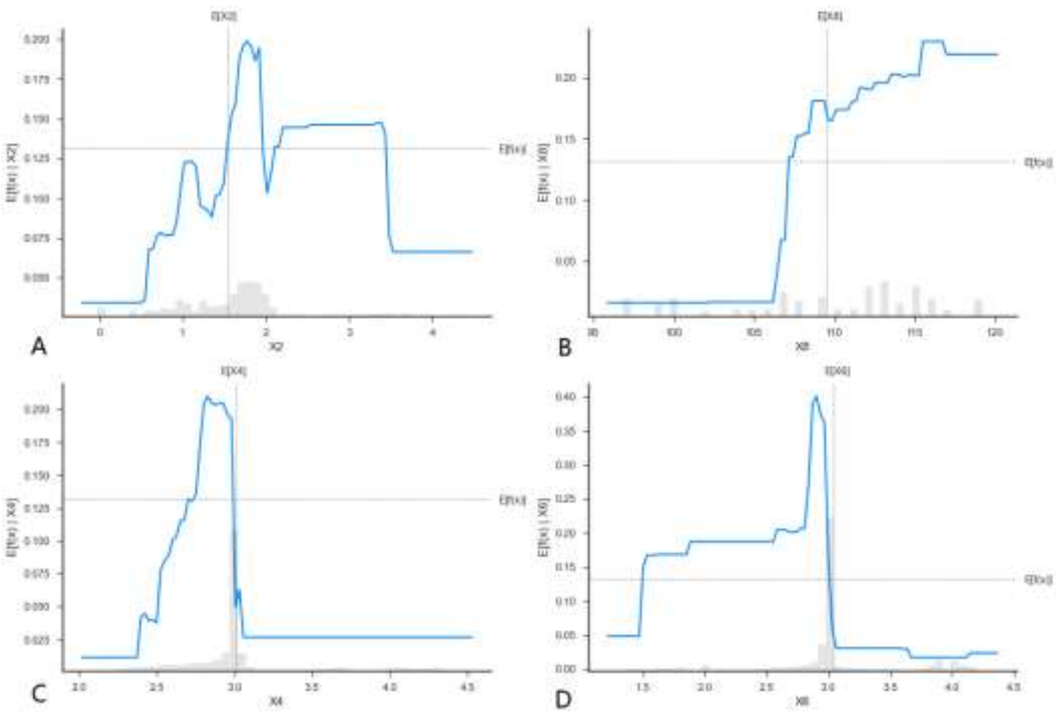

Fig 3: Partial dependence plots of important variables. (A) Partial dependence plot of TM versus NCI risk; (B) Partial dependence plot of MVJ versus NCI risk; (C) Partial dependence plot of 4WFatigue versus NCI risk; (D) Partial dependence plot of 4WStress versus NCI risk.

3.5 Impact of interactions between important characteristics on NCI risk

Although feature attribution analysis provides an understanding of the marginal effects of feature weights and features on the predicted outcomes of the model, the use of changes in single variable metrics does not predict the occurrence of athletic injuries. Through the two-dimensional interaction partial dependence plot between TM, 4WFatigue, 4WStress, and MVJ (Fig 4), the authors found that there was a significant interaction between TM, 4WFatigue, and MVJ, and the phenomenon was largely consistent with Bittencourt et al. [10]. On the one hand, it confirms that the occurrence of sports injuries is not the result of linear combination of multiple risk factors, but originates from linear and nonlinear interactions among multiple factors, which cannot be accurately predicted by using only a single indicator. On the other hand, the injury risk prediction models constructed by applying machine learning models can tap into potential patterns and help promote the development of sports science and sports medicine research.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

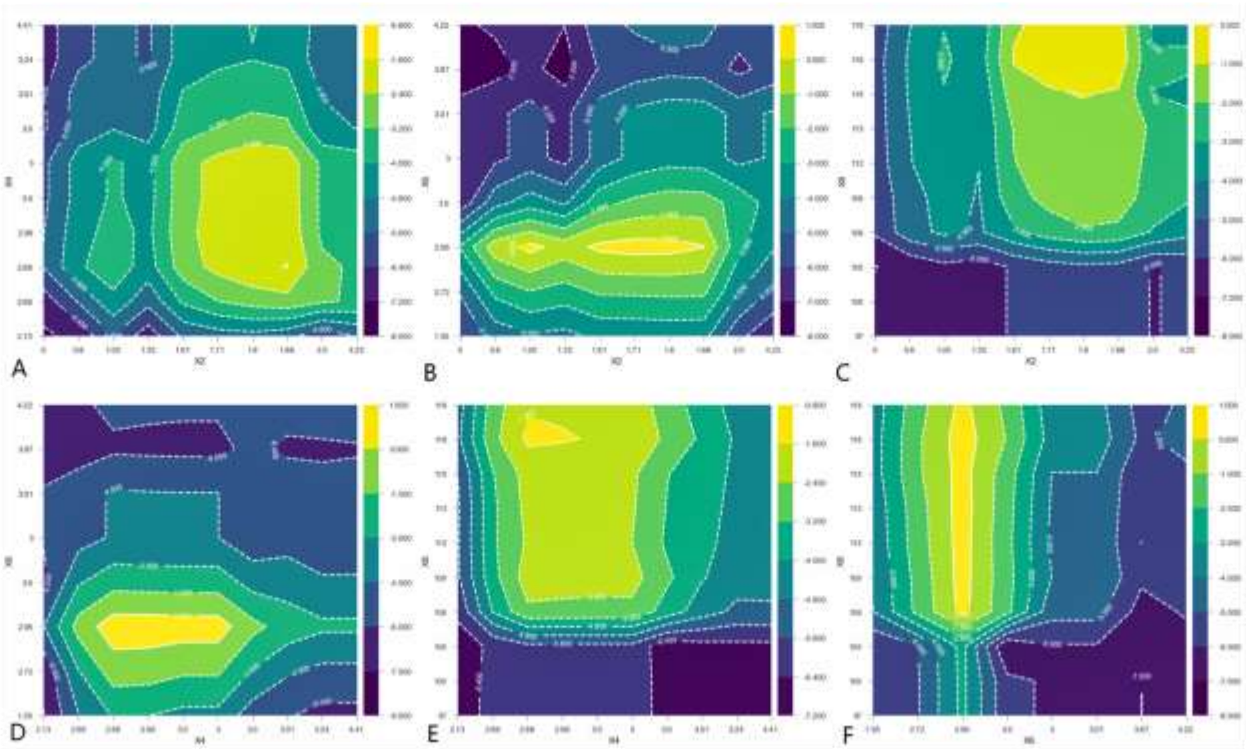

Fig 4: Two-dimensional interaction partial dependence plots between important features. (A) 2D interaction partial dependence diagrams of TM and 4WFatigue; (B) 2D interaction partial dependence diagrams of TM and 4WStress; (C) 2D interaction partial dependence diagrams of TM and MVJ; (D) 2D interaction partial dependence diagrams of 4WFatigue and 4WStress; (E) 2D interaction partial dependence diagrams of 4WFatigue and MVJ; (F) Partial dependence diagram of 2D interaction between 4WStress and MVJ.

Ⅳ. CONCLUSION

The NCI risk prediction model for adolescent female basketball players constructed based on XGBoost algorithm in this study has high performance and can better identify and predict the NCI risk. Unlike previous studies, the model constructed in this study has high explanatory and reliability. The authors identified four important variables that influence the risk of NCI in adolescent female basketball players through the model. Also the partial dependence plot of the model indicated that there was a potential interaction between the important variables. This not only helps coaches and researchers understand injury causation from the model, but also provides evidence for research in sports science. It should be noted that, considering the generalizability of the study results on NCI prevention in youth female basketball players, this paper did not conduct specific analysis for individual athlete specificity, and was limited by various factors such as research time, research conditions, research funding, and coaches'

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

cooperation, the sample size in the study was small and still needs further expansion. In the future, the authors will conduct research on sample specificity to promote the in-deep integration and development of sports science, sports medicine and data science.

REFERENCES

[1] Agel, J., Olson, D.E., Dick, R.W., Arendt, E., Marshall, S.W., & Sikka, R. (2007). Descriptive epidemiology of collegiate women's basketball injuries: National Collegiate Athletic Association Injury Surveillance System, 1988-1989 through 2003-2004. Journal of athletic training, 42 2, 202-10 .

[2] Meeuwisse, W.H., Sellmer, R., & Hagel, B.E. (2003). Rates and Risks of Injury during Intercollegiate Basketball. The American Journal of Sports Medicine, 31, 379 - 385.

[3] Gabbett, T.J. (2016). The training—injury prevention paradox: should athletes be training smarter and harder? British Journal of Sports Medicine, 50, 273 - 280.

[4] Claudino, J.G., Capanema, D.D., de Souza, T.V., Serrão, J., Machado Pereira, A.C., & Nassis, G.P. (2019). Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. Sports Medicine - Open, 5.

[5] Talukder, H, Foster G, Hu C, Huerta J, Kumar A,et al. (2016). Preventing in-game injuries for NBA players.MIT Sloan Analytics Conference.

[6] Rossi, A., Pappalardo, L., Cintia, P., Iaia, F.M., Fernández, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. PLoS ONE, 13.

[7] Luu, B.C., Wright, A.L., Haeberle, H.S., Karnuta, J.M., Schickendantz, M., Makhni, E.C., Nwachukwu, B.U., Williams, R.J., & Ramkumar, P.N. (2020). Machine Learning Outperforms Logistic Regression Analysis to Predict Next-Season NHL Player Injury: An Analysis of 2322 Players From 2007 to 2017. Orthopaedic Journal of Sports Medicine, 8.

[8] Rommers, N., Rössler, R., Verhagen, E.A., Vandecasteele, F., Verstockt, S., Vaeyens, R., Lenoir, M., D'Hondt, E., & Witvrouw, E. (2020). A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players. Medicine & Science in Sports & Exercise, 52, 1745 - 1751.

[9] Ruddy, J.D., Cormack, S.J., Whiteley, R., Williams, M., Timmins, R.G., & Opar, D.A. (2019). Modeling the Risk of Team Sport Injuries: A Narrative Review of Different Statistical Approaches. Frontiers in Physiology, 10.

[10] Bittencourt, N.F., Meeuwisse, W.H., Mendonça, L.D., Nettel-Aguirre, A., Ocarino, J.D., & Fonseca, S.T. (2016). Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition—narrative review and new concept. British Journal of Sports Medicine, 50, 1309 - 1314.

[11] Fiscutean, A. (2021). Data scientists are predicting sports injuries with an algorithm. Nature, 592 7852, S10-S11 .

[12] Foster, C., Hector, L., Welsh, R., Schrager, M., Green, M.A., & Snyder, A.C. (2004). Effects of specific versus cross-training on running performance. European Journal of Applied Physiology and Occupational Physiology, 70, 367-372.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1884-1899
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

[13] Chen, M.J., Fan, X., & Moe, S.T. (2002). Criterion-related validity of the Borg ratings of perceived exertion scale in healthy individuals: a meta-analysis. Journal of Sports Sciences, 20, 873 - 899.

[14] Williams, S., West, S.W., Cross, M., & Stokes, K.A. (2016). Better way to determine the acute:chronic workload ratio? British Journal of Sports Medicine, 51, 209 - 210.

[15] Bourdon, P.C., Cardinale, M., Murray, A.M., Gastin, P.B., Kellmann, M., Varley, M.C., Gabbett, T.J., Coutts, A.J., Burgess, D.J., Gregson, W., & Cable, N.T. (2017). Monitoring Athlete Training Loads: Consensus Statement. International journal of sports physiology and performance, 12 Suppl 2, S2161-S2170 .

[16] Hooper, S. L., & Mackinnon, L. T. (1995). Monitoring overtraining in athletes. Recommendations. Sports medicine , 20(5), 321–327.

[17] McLean, B.D., Coutts, A.J., Kelly, V.G., McGuigan, M.R., & Cormack, S.J. (2010). Neuromuscular, endocrine, and perceptual fatigue responses during different length between-match microcycles in professional rugby league players. International journal of sports physiology and performance, 5 3, 367-83 .

[18] Fuller, C.W., Ekstrand, J., Junge, A., Andersen, T.E., Bahr, R., Dvořák, J., Hägglund, M., McCrory, P., & Meeuwisse, W.H. (2006). Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. British Journal of Sports Medicine, 40, 193 - 201.

[19] Hulin, B.T., Gabbett, T.J., Lawson, D.W., Caputi, P., & Sampson, J.A. (2015). The acute:chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players. British Journal of Sports Medicine, 50, 231 - 236.

[20] Chawla, N., Bowyer, K., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res., 16, 321-357.

[21] Lundberg, S.M., Erion, G.G., Chen, H., DeGrave, A.J., Prutkin, J.M., Nair, B.G., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2, 56-67.

[22] Curtis, R.M., Huggins, R.A., Benjamin, C.L., Sekiguchi, Y., Lepley, L.K., Huedo-Medina, T.B., & Casa, D.J. (2021). Factors Associated With Noncontact Injury in Collegiate Soccer: A 12-Team Prospective Study of NCAA Division I Men's and Women's Soccer. The American Journal of Sports Medicine, 49, 3076 - 3087.