# The Investigation of Athlete Injuries Prediction Based on Machine Learning Models

Junliang Lv

Department of International Business, Jinan University, Guangzhou, 510000, China
email:linton99@hhu.edu.cn

**Abstract.** This article provides a literature review on the application of machine learning in athlete pain detection. This review can help future researchers and learners gain a general understanding of the current research status and future expectations of machine learning in the field of athlete pain detection. This article also explains the commonly used machine learning methods in athlete pain detection, such as random forest, regression analysis, artificial neural networks, etc. It briefly describes the steps of machine learning to establish a pain detection model, including data collection, model building, and model detection. In the conclusion section of this article, the advantages of machine learning in accurately predicting pain through data and the difficulty in collecting large amounts of data, as well as the lack of high generalization in the constructed models, are mentioned. Based on these shortcomings, a prospect is made for inventing better data collection methods and building higher quality machine learning methods.

**Keywords:** Machine learning, athlete injury, artificial intelligence

## 1 Introduction

In the development of modern sports, the results of competitions not only depend on the efforts of individual athletes, but are also affected by the good external environment, high-quality equipment and scientific management provided by the team. In various external environments, athletes' health management is a crucial part, because it not only determines the athlete's performance on the field, but also relates to the athlete's possible injury and his career lifespan. As competition in modern sports becomes increasingly fierce, how to scientifically and effectively manage athlete health and prevent injuries has become an important task for sports organizations and coaching teams. However, traditional sports health management methods may have considerable subjective factors, such as the experience of medical staff, athletes' unclear understanding of potential risks or the need to hide existing injuries [1]. At the same time, as machine learning technology becomes more mature, people can extract more information from massive training and competition data to help athletes and the coaching team to identify potential health risks, allowing people to abandon the original overly subjective health management method and use more objective, scientific and effective methods to optimize athletes' health management


Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_63

and monitor their physical status, so that the possibility of injury is reduced to the lowest possible level. and extend their career life expectancy [2].

Machine learning has been used to a considerable extent in many sports fields. It can analyze athletes' historical performance training data to obtain abstract data such as the athlete's technical level, recent performance, and training effects [3], and use these abstract data to evaluate upcoming games. Predictions, such as analyzing National Basketball Association (NBA) player performance [4]. Machine learning can also monitor the possibility of athlete injury by tracking the athlete's sports data [5]. Various machine learning methods are used to monitor athletes' health management and injuries. There are also many methods for pain risk prediction, such as using logistic regression to predict common injuries among athletes [6]. Through these studies, the coaching team can more accurately help athletes adjust their physical condition and prevent injuries. Some articles also highlight that machine learning can efficiently analyze large amounts of data and output prediction data with high accuracy. These innovative statistical techniques have the potential to transform the practice of sports medicine and represent a significant future direction in the field [7].

The remainder of this article is organized as follows. First, in Section 2, this article will review solutions of machine learning in health management and injury prevention. Then, in the third section, this article will discuss and evaluate the effectiveness, universality, etc. of different methods, and explore the challenges and prospects of machine learning in the field of athlete health management. Finally, in Section 4, this paper will summarize the above and present conclusions.

## 2 Methods

### 2.1 Basic Process of Machine Learning Analysis Methods

The general machine learning methods are divided into the following steps: data collection and processing, feature selection, model selection, and evaluation of model results.

In the data collection and processing part, it is necessary to collect as much raw data as possible because this will provide more learning opportunities for the model. After collecting the data, the original data should be processed accordingly, such as normalization or standardization, so that the model can analyze them more accurately and quickly, and make the results more intuitive.

The collected data cannot be fully input into the model for processing, as meaningless information can interfere with the model and result in poor performance. So before training the model, the data should be selected first. The selection of data can adopt a manual filtering mode, but this mode is too subjective and may exaggerate or ignore the importance of certain information, leading to model inaccuracy. Data-driven methods are used to solve this problem, such as using principal component analysis to identify important information. The mainstream view now is that subjective analysis should be combined with data analysis, as data analysis cannot find all potential key information.

After selecting features, one should choose a machine learning method and use it as a model to fit the data. Due to the diverse methods of machine learning, including regression models, random forest models, neural network models, etc., different machine learning methods can lead to different prediction results. Multiple machine learning methods can be used to fit the data and select the appropriate model based on their evaluation of the results. This paper will introduce several commonly used models for predicting athlete injuries in the following text.

After using the model to fit the data, it is necessary to evaluate the fitting results. The commonly used model evaluation methods include cross validation and hyperparameter optimization, which can be used to obtain the relevant performance of the model, such as accuracy and robustness. In the process of model selection, it is necessary to select evaluation data based on needs. If a more universal model is needed, a more robust model should be selected. If only analyzing a single situation, a more accurate model can be used.

## 2.2 Common Machine Learning Models

### 2.2.1 Regression Model

Regression models have many sub classifications, including linear regression, polynomial regression, logistic regression, etc. These regression methods are often used to study the relationship between independent and dependent variables and are commonly used in predictive analysis. When encountering different situations, different regression models can be used. For example, if there is a clear linear relationship between the independent variable and the dependent variable, linear regression or polynomial regression can be used. If classification problems need to be studied, logistic regression can be used.

Yining used regression models to reveal the relationship between athlete strain and their behavior [8]. They use data from public data websites as datasets, define independent and dependent variables based on the obtained data and actual needs, analyze the variables using logistic regression methods, evaluate and interpret the model, and finally put the digital model into practical use.

Regression models can not only reveal significant relationships between independent and dependent variables, but also indicate the strength relationship between multiple independent variables and a single dependent variable.

### 2.2.2 Random Forest Model

Random forest is an idea of ensemble learning [9], which obtains data through random sampling, inputs numerous decision trees, and votes to ultimately obtain the final output result. Simply put, the establishment of a random forest model can be divided into three steps. Firstly, samples are selected as the training set, and n random samples with replacement are selected and repeated k times; Then extract x features from X total features, where x is much smaller than X; Finally, all the decision trees obtained will be democratically voted on and used for prediction or classification.

Yining also used regression models to reveal the relationship between athlete strain and their behavior [8]. They use data from public data websites as datasets, define independent and dependent variables based on the obtained data and actual needs,

analyze the variables using logistic regression methods, evaluate and interpret the model, and finally put the digital model into practical use.

The principle of the random forest model is simple, with high accuracy, and can be used for various problems including regression, classification, etc.

#### 2.2.3 Artificial Neural Network Model

A neural network is a complex network system formed by a large number of simple processing units (called neurons) that are widely interconnected. It reflects many basic characteristics of human brain function and is a highly complex nonlinear dynamic learning system. Neurons are biological models based on the neural cells of the biological nervous system. When people study the biological nervous system to explore the mechanisms of artificial intelligence, they mathematize neurons, resulting in the development of neural mathematical models. A large number of neurons with the same form are connected together to form a neural network. Neural networks are highly nonlinear dynamic systems. Although the structure and function of each neuron are not complex, the dynamic behavior of neural networks is very complex; Therefore, neural networks can express various phenomena in the actual physical world.

Milad Keshtkar Langaroudi mentioned artificial neural network models that obtain data from multiple men's tournaments, then analyze the data using ANN software and multiply the data into multiple datasets to obtain appropriate noise, resulting in a more general model [10].

## 3 Discussion

### 3.1 Limitations and Challenges

Although the current consistency and calibration level of machine learning has reached an excellent level, in order to simplify the steps of data collection, most research samples are mostly students or professional athletes. The data obtained from this dataset will make the model unsuitable for the elderly, children, sick and weak. Therefore, these models can only serve the purpose of education and inspiration and are difficult to be widely used in daily life.

Some sports are too niche and have limited information to collect. Due to people's limited understanding of these sports, the selection of data before model establishment often relies on the nature of the data itself, which cannot be combined with subjective screening by humans. This may lead to inaccurate variable selection in the model. At the same time, the causes of human injury caused by sports are diverse and full of randomness. When constructing models for these sports, various factors need to be considered, such as scores, morale, venue, etc., which further deepens the difficulty of constructing models for niche sports. Many models adopt a small sample size, limited to a few matches or a league, which may be influenced by some special factors, resulting in significant uncertainty and contingency in the results, making it difficult to use the data obtained in other matches. It is unlikely to conduct machine learning analysis for every competition.

### 3.2 Future Prospects

Machine learning methods can be used to identify athletes who may be injured during exercise and help identify risk factors. Although many models correctly use machine learning methods to predict injuries, the quality of their research methods and model building is relatively low. Machine learning is a constantly evolving field and given the enormous potential for development between machine learning and artificial intelligence, developing better model building or research methods in the future can help facilitate better development.

In the future, increasing the sample size of participating models or using better machine learning methods to eliminate accidental factors in small amounts of data will help sports teams obtain better models. Inventing an efficient machine or method for collecting samples can also help solve this problem. And using this method allows more nonprofessional athletes and enthusiasts to monitor their data and fit models that are more suitable for them.

In the next decade, machine learning can help humans improve the objectivity of injury prediction and sports science decision-making, but there are still significant challenges to overcome, such as avoiding traps in machine learning model creation and popularizing the application of machine learning methods in the field of sports.

## 4 Conclusion

This article provides a review of research on machine learning in predicting sports pain. In the second part, the general steps of machine learning are described, including data collection, feature selection, model construction and evaluation, as well as three common machine learning models: regression model, random forest model, and artificial neural network model, and their specific operational steps in practical use are provided. In the third part, this article analyzes the application of machine learning in sports pain prediction, pointing out the limitations of certain machine learning data that are not universal and some niche sports that are difficult to collect enough data to build models. It also looks forward to more scientific and effective machine learning or artificial intelligence methods and more efficient data collection instruments in the future and holds a positive attitude towards machine learning helping humans to scientifically and effectively predict athlete pain.

## References

1. Makos, S., Thompson, C. M.: Core and Catalyst Criteria Motivating CrossFit Athletes to Reveal or Conceal Their Non-Visible Health Conditions. Communication & Sport (2024).
2. Pareek, A., Karlsson, J., Martin, R. K.: Machine Learning/Artificial Intelligence in Sports Medicine: State of the Art and Future Directions. Journal of ISAKOS (2024).
3. Fearnhead, P., Taylor, B. M.: On estimating the ability of NBA players. Journal of Quantitative Analysis in Sports 7(3) (2011).
4. Wheeler, K.: Predicting NBA Player Performance. Wheeler-PredictingNBAPlayerPerformance.pdf (2012).

5. Li, J.: An investigation of an athlete injury likelihood monitoring system using the random forest algorithm and DWT. Technology and Health Care Preprint (2024): 1-15.
6. Ayala, R. E. D., et al.: Novel Study for the Early Identification of Injury Risks in Athletes Using Machine Learning Techniques. Applied Sciences 14(2), 570 (2024).
7. Pareek, A., Karlsson, J., Martin, R. K.: Machine Learning/Artificial Intelligence in Sports Medicine: State of the Art and Future Directions. Journal of ISAKOS (2024).
8. Lu, Y., et al.: Machine learning for predicting lower extremity muscle strain in national basketball association athletes. Orthopaedic Journal of Sports Medicine 10(7): 23259671221111742 (2022).
9. Rigatti, S. J.: Random forest. Journal of Insurance Medicine 47(1), 31-39 (2017).
10. Keshtkar Langaroudi, M., Yamaghani, M.: Sports result prediction based on machine learning and computational intelligence approaches: A survey. Journal of Advances in Computer Engineering and Technology 5(1), 27-36 (2019).