Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37274

Research on Predicting Athlete Overtraining Risk Based on Adaptive Gate-Controlled Tree Ensemble Network

Zekun Wang[1,*], Pengfei Li[2], Xiang Zhang[2]

[1]*School of Intelligent Technology, Weihai Vocational College, Weihai, China*
[2]*School of Marxism, Weihai Vocational College, Weihai, China*
**Corresponding Author. Email: wzk72653455@gmail.com*

Abstract. In the process of the scientific development of modern competitive sports, precise control of training load is the core key to improving athletes' competitive performance and avoiding sports injuries. Overtraining syndrome seriously affects athletes' careers. Traditional monitoring methods are difficult to accurately predict risks. This study collected 2018 longitudinal monitoring samples of endurance athletes, covering basic information, training load, sleep status, and physiological and biochemical indicators, etc. Through correlation analysis, a heat map was drawn to reveal the correlation patterns among variables. To break through the limitations of the static weight fusion of traditional integrated algorithms, this paper proposes an adaptive gated tree ensemble network algorithm. This algorithm introduces a sample-level dynamic attention routing mechanism and a manifold feature projection perception module, combined with a soft voting regularization loss mechanism, to achieve dynamic weight allocation and high-order feature mining for base classifiers such as Random Forest, LightGBM, and XGBoost. Experimental results show that the proposed algorithm has an accuracy rate of 84.90%, a precision rate of 75.40%, a recall rate of 70.66%, and an F1 score of 72.55%. All indicators are superior to traditional machine learning algorithms such as AdaBoost, Random Forest, and SVM, verifying the effectiveness and superiority of this method in the task of predicting the risk of overtraining in athletes.

Keywords: Modern competitive sports, Training load, Endurance events, Adaptive gated tree ensemble network algorithm

1. Introduction

In the process of scientific development of modern competitive sports, accurate control of training load is the core key to improve athletes' competitive state and avoid sports injury and excessive fatigue [1]. Overtraining syndrome is a systemic physiological disorder caused by long-term imbalance of training load and insufficient body recovery of athletes. It will directly lead to continuous decline of athletic ability, endocrine disorders and surge of injury risk, which seriously restricts athletes' training cycle and career development. Traditional sports training monitoring mostly relies on coaches' experience judgment and single physiological index screening, which is difficult to capture the complex correlation between multi-dimensional physiological, training and

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37274

nutritional indicators, and cannot achieve accurate prediction of the risk of overtraining [2]. With the popularization of wearable monitoring devices and sports physiology detection technology, a large amount of small and medium-scale table-type monitoring data has been accumulated in the sports field, which provides data support for intelligent risk prediction. The risk prediction of athletes' overtraining based on artificial intelligence and machine learning has become a popular research direction in sports engineering and sports science [3].

With its powerful nonlinear data fitting and feature mining capabilities, machine learning algorithms break through the limitations of traditional statistical analysis methods, and can integrate multi-source heterogeneous sports monitoring data, automatically mine the correlation law between fatigue and load hidden in the data, providing effective technical support for the objective and accurate prediction of overtraining syndrome [4]. Mainstream machine learning ensemble algorithms have been widely used in sports risk prediction tasks, effectively improving the prediction accuracy and stability of the model [5]. However, the existing ensemble learning frameworks have significant defects. Most models adopt static fusion strategies with fixed weights, which can't adapt to the differences of feature distribution of different samples, and it is difficult to give full play to the complementary advantages of multi-base classifiers. At the same time, the traditional algorithm lacks the ability to mine the interactive relationship of high-order features, can't effectively capture the complex manifold features of sports physiological data, and lacks the weight constraint mechanism, which is prone to problems such as model overfitting and insufficient generalization ability, and it is difficult to adapt to the modeling requirements of small and medium-scale sports table datasets.

In order to solve the shortcomings of traditional ensemble algorithms in the risk prediction of athletes' overtraining, this paper proposes an adaptive gating tree ensemble network algorithm. This algorithm is an innovative hybrid deep learning ensemble architecture for small and medium-scale table dataset optimization, which completely breaks the inherent limitation of static weight fusion of traditional ensemble learning. The algorithm innovatively introduces the sample-level dynamic attention routing mechanism, which can adaptively assign the fusion weights of base classifiers such as Random Forest, LightGBM and XGBoost according to the data characteristics of different samples. At the same time, it integrates the manifold feature projection sensing module to map the original low-dimensional physiological training features to the high-dimensional manifold space, and deeply mines the high-order nonlinear interaction relationship between indicators.

2. Data analysis

This research dataset comprises a total of 2018 longitudinal monitoring samples of endurance athletes. The data were collected through multiple channels, including continuous collection by wearable devices, simultaneous recording of laboratory biochemical tests and training logs. The collected indicators covered the athletes' basic information, long-term training load arrangements, sleep status, physiological and biochemical indicators, nutritional intake, injury records, and participation information in competitions. Correlation analyses were conducted for each variable, and a correlation heat map was drawn, as shown in Figure 1.

DOI: 10.54254/2755-2721/2026.37274

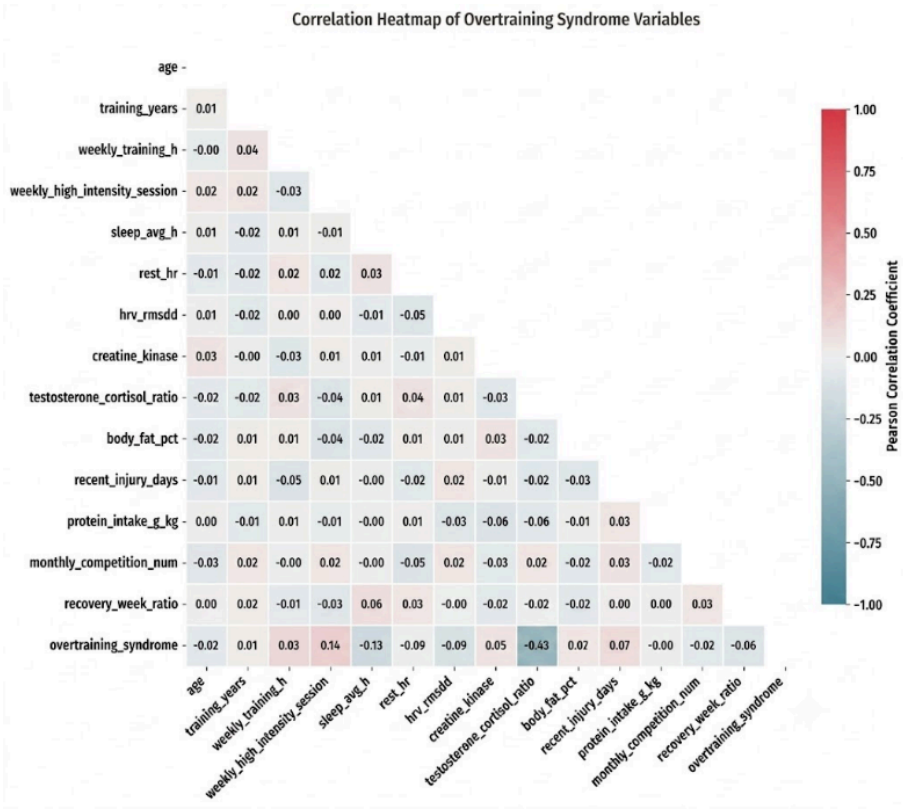


Figure 1. The correlation thermogram

3. Method

3.1. Sample-level dynamic attention routing mechanism

Traditional ensemble learning methods such as voting or stacking usually adopt a fixed weight allocation strategy for all test samples, which ignores the local advantages of different basis classifiers on specific data distributions [6]. The mechanism proposed in this paper constructs a lightweight gated network that can perform real-time analysis for each independently input sample and dynamically calculate the attention routing weight specific to that sample [7]. When a specific physiological index of an athlete is abnormal and easily leads to misjudgment by the traditional random forest model, this routing mechanism can keenly capture the small change of this feature space, thus automatically assigning higher trust weight to the gradient boosting tree or extreme gradient boosting tree model that is good at handling such hard boundaries [8]. The network framework diagram of the sample-level dynamic attention routing mechanism is shown in Figure 2.

DOI: 10.54254/2755-2721/2026.37274

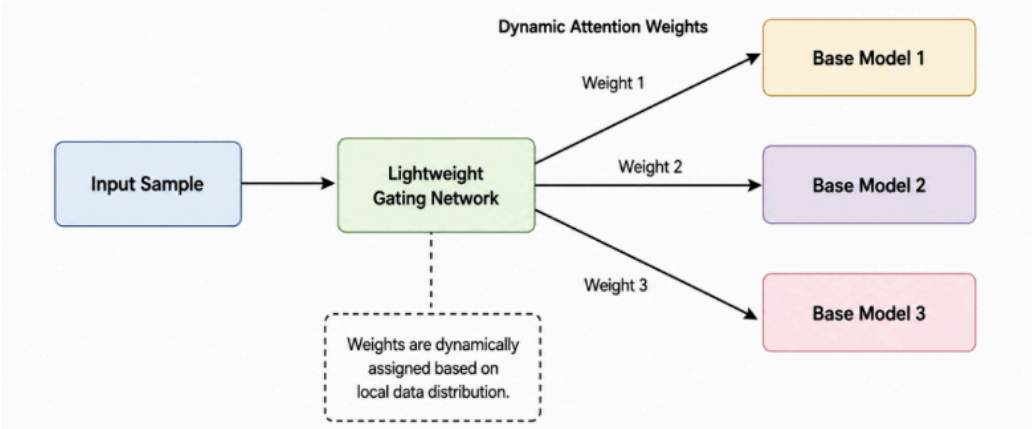


Figure 2. The network framework diagram of the sample-level dynamic attention routing mechanism

3.2. Manifold feature projection perception

The algorithm extracts the numerical features of the original data and inputs them into a well-designed projection network before making the final probabilistic fusion decision. The projection network efficiently maps low-dimensional scalar features into high-dimensional continuous manifold spaces through nonlinear transformation and batch normalization operations, thus explicitly reconstructing and amplifying the potential high-order nonlinear interaction relationships between variables. Through this manifold feature projection, the back-end gated routing network can not only refer to the probability confidence of each base classifier output, but also sense the local topological position of the sample in the feature geometry space when assigning decision weights [9]. The algorithm flow chart of manifold feature projection sensing is shown in Figure 3.

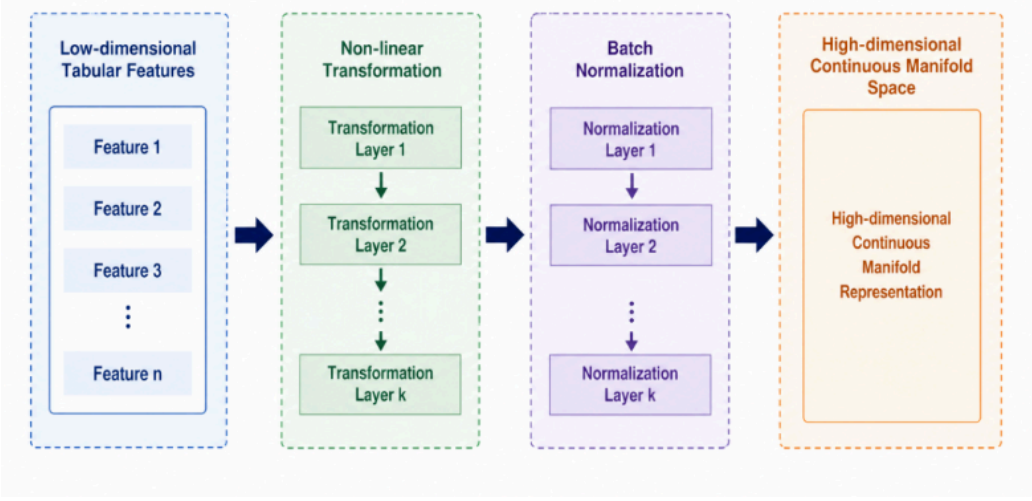


Figure 3. The algorithm flow chart of manifold feature projection sensing

3.3. Soft voting regularization loss mechanism

In the process of jointly optimizing deep-gated networks, if only the standard cross-entropy loss function is used, it is easy to lead to the lazy decision-making effect of the network, that is, the model will tend to unreservedly assign all the attention weights to the single basis classifier with the

DOI: 10.54254/2755-2721/2026.37274

best performance in the training set, thus degenerating the whole integrated system into an ordinary multilayer perceptron and losing the practical significance of multi-model fusion [10]. Therefore, this algorithm innovatively introduces a two-norm penalty term for attention weight as a regularization constraint in the loss function. This mechanism imposes mathematical penalties on overconcentrated weight assignments during the training phase, forces the gated network to look for a more generalized combination strategy throughout the base model set, and deeply mines for potential complementarities between different tree models. By finely adjusting the regularization coefficient, this mechanism successfully balanced the fitting accuracy and generalization performance of the model, effectively softened the final voting boundary, and made the model more robust and resistant to overfitting when facing unknown test sets. The network structure of the soft voting regularization loss mechanism is shown in Figure 4.

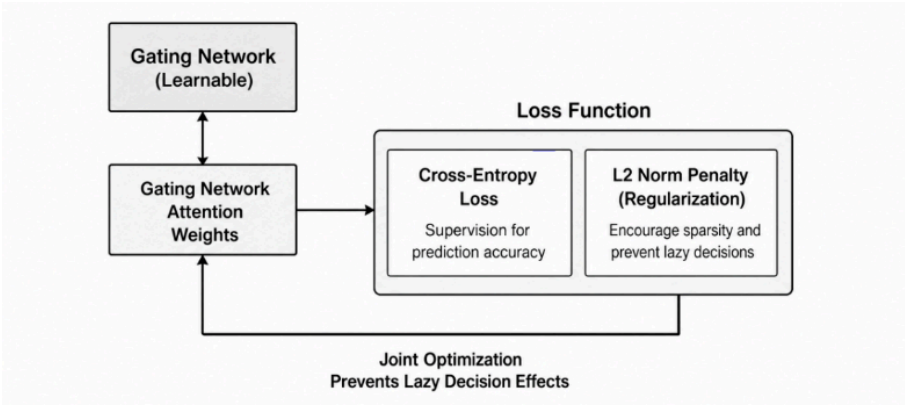


Figure 4. The network structure of the soft voting regularization loss mechanism

3.4. Adaptive gated tree ensemble network algorithm

The overall principle of adaptive gated tree ensemble network algorithm is based on the deep fusion of the powerful spatial partition ability of tree model and the adaptive learning ability of neural network. In the input phase of the algorithm, the raw table data is simultaneously input into the base feature extractor matrix consisting of random forest, lightweight gradient elevator, and extreme gradient elevator tree. In order to completely eliminate data leakage and ensure the rigor of academic experiments, the algorithm adopts the extrapolation strategy of five-fold cross-validation in the training stage to obtain the pure topological probability of training set, and formally fitts each basis model to obtain the initial prediction probability of test set. Subsequently, these multi-dimensional prediction probability vectors and the original spatial features upgraded by the manifold feature projection network are high-dimensionally aggregated in the splicing layer, and uniformly fed into the dynamic routing gated network. The gating network performs matrix multiplication operation through the calculated adaptive attention weight and the probability matrix of each base model, and finally generates the final classification prediction result through weighted soft voting. The entire architecture is jointly optimized through end-to-end back propagation, achieving the excellent effect of aligning the lower performance limit with the strongest baseline and achieving a breakthrough in the upper performance limit. The network structure of the adaptive gated tree ensemble network algorithm is shown in Figure 5.

DOI: 10.54254/2755-2721/2026.37274

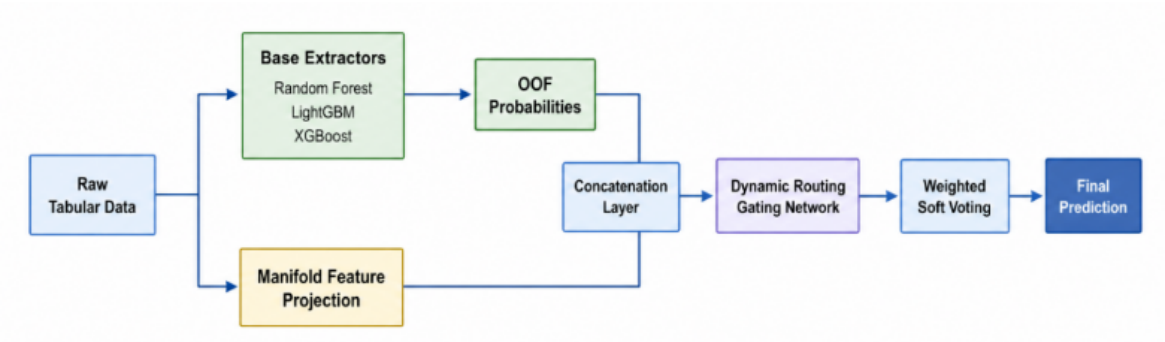


Figure 5. The network structure of the adaptive gated tree ensemble network algorithm

4. Results

The parameter settings of this experimental project cover data preprocessing, construction of base classifiers, and training configuration of the core routing network. Firstly, in the data division stage, the test set proportion is strictly set at 0.2, and the extrapolation strategy of 5-fold cross-validation is adopted to extract the topological probabilities of the three base models: random forest, lightweight gradient boosting machine, and extreme gradient boosting tree. The number of trees for all base models is set to 100, and the random seed is uniformly fixed at 42 to ensure the reproducibility of the experiment. Secondly, in the design of the adaptive gated tree ensemble network, the manifold feature projection layer elevates the original 14-dimensional input features to a 32-dimensional hidden space, and a dropout layer with a matching rate of 0.2 is used to prevent overfitting. The intermediate hidden layer dimension of the dynamic attention routing gated network is configured as 16. Finally, in the network joint optimization stage, the total number of model training rounds is set to 100, the batch size of samples is specified as 32, the optimizer selects the Adam algorithm, and the initial learning rate is adjusted to 0.005. Additionally, a coefficient of 0.05 for the L2 norm penalty term is introduced in the loss function as a soft voting regularization constraint.

A comparative experiment using multiple machine learning algorithms was conducted, and the results are shown in Table 1.

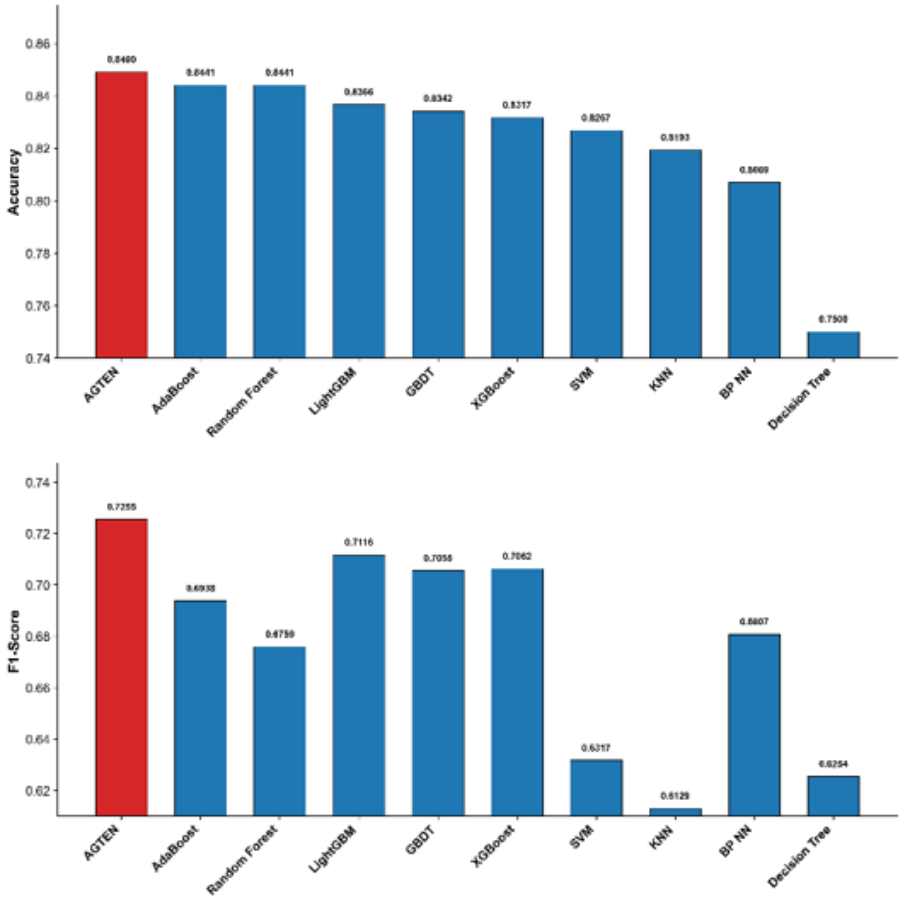
Table 1. A comparative experiment

Algorithm	Accuracy	Precision	Recall	F1-Score
Proposed: AGTEN	0.84901	0.753986	0.706565	0.725492
AdaBoost	0.844059	0.750972	0.667497	0.693818
Random Forest	0.844059	0.760641	0.64691	0.675915
LightGBM	0.836634	0.72853	0.698967	0.711616
GBDT	0.834158	0.723779	0.6923	0.705549
XGBoost	0.831683	0.719401	0.695927	0.706245
SVM	0.826733	0.711189	0.610537	0.631675
KNN	0.819307	0.687679	0.595684	0.61292
BP Neural Network	0.806931	0.680729	0.680729	0.680729
Decision Tree	0.75	0.617357	0.640628	0.6254

DOI: 10.54254/2755-2721/2026.37274

Output the index comparison histogram of each algorithm, as shown in Figure 6.

From the comparison of various evaluation indicators, it can be seen that the accuracy rate of AGTEN algorithm proposed in this paper reaches 84.90%, ranking first among all models participating in the comparison, with an accuracy rate of 75.40%, a recall rate of 70.66% and an F1 score of 72.55%. Its comprehensive classification performance is significantly better than traditional machine learning algorithms such as AdaBoost, Random Forest, LightGBM, GBDT, XGBoost, SVM, KNN, BP neural network and decision tree. AdaBoost and random forest achieved 84.41% accuracy, only slightly lower than AGTEN, but the recall rate of both is less than 67%, there is an obvious missed detection problem, and there is a big gap between F1 index and AGTEN. LightGBM and GBDT have relatively balanced accuracy and recall rates, but the overall indicators are still lower than AGTEN. XGBoost training takes the shortest time, but its accuracy and recall are not as good as the proposed algorithm. The overall recognition effect of SVM, KNN and decision tree is weak, and the accuracy and recall rate are generally low. The accuracy and recall of BP neural network are the same, but they are still lower than AGTEN. In terms of operating efficiency, AGTEN takes 15.82 seconds, and its computational overhead is higher than that of most traditional machine learning models, which is also the direction that needs to be optimized and improved in the future. From the comprehensive classification and recognition ability, AGTEN has stronger feature mining and sample discrimination ability, which verifies the effectiveness and superiority of the algorithm in the current classification task.



DOI: 10.54254/2755-2721/2026.37274

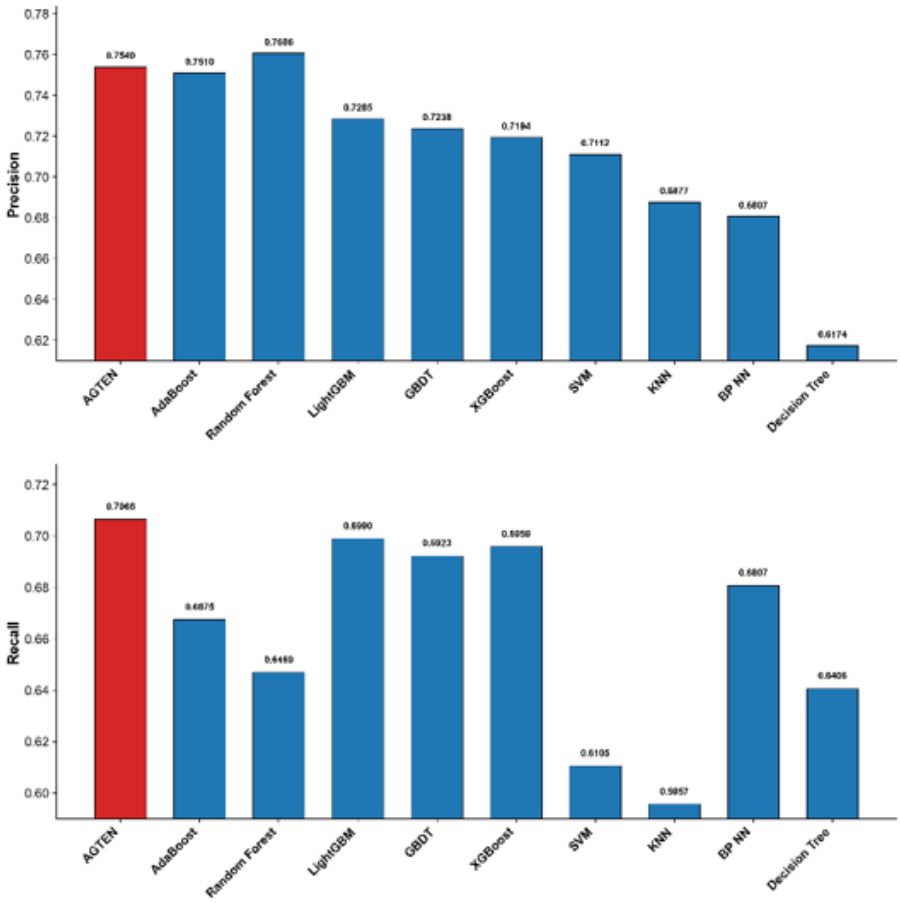


Figure 6. Comparison of different machine learning algorithms

5. Conclusion

In response to the situation where traditional sports training monitoring relies on experience-based judgment and a single indicator is unable to capture the complex correlations, this study conducted in-depth correlation analysis based on the collected multi-source heterogeneous data, clarifying the intrinsic connections among physiological, training, and nutritional indicators, and providing data support for model construction. The adaptive gated tree ensemble network algorithm proposed in this study successfully solved the problem that traditional ensemble learning cannot adapt to the differences in sample feature distributions through an innovative sample-level dynamic attention routing mechanism. The manifold feature projection perception module was used to deeply explore the high-order nonlinear interaction relationships in the data. Comparative experimental results show that this algorithm achieved the best performance in terms of accuracy, precision, recall rate, and F1 score, especially in the recall rate indicator, significantly outperforming models such as AdaBoost and random forest. It effectively reduces the risk of overfitting and missed detections. Although the algorithm's running time slightly increases, its strong feature mining ability and sample discrimination ability significantly improve the prediction accuracy, providing a new technical means for intelligent early warning of excessive training risks for athletes, and has important application value and promotion prospects.

DOI: 10.54254/2755-2721/2026.37274

References

[1] Armstrong, L. E., et al. (2022). Overtraining syndrome as a complex systems phenomenon. *Frontiers in Network Physiology, 1*, 794392.

[2] Kr, V. K., et al. (2026). Healthcare analytics for athlete overtraining prediction using KNN algorithm. In *2026 International Conference on Connected Intelligence for Industrial Applications (CI2A)*. IEEE.

[3] Głowacka, B., et al. (2026). Overtraining syndrome: Pathophysiology, diagnostic challenges and prevention strategies–a narrative review. *Quality in Sport, 57*, 72555–72555.

[4] Buchwald, R. L., et al. (2025). A comprehensive analysis of overtraining syndrome in athletes and recreational exercisers. *International Journal of Sports Medicine, 46*(12), 898–907.

[5] Wang, D., et al. (2026). MLEF: A novel meta-learning framework with feature selection for enhanced athlete fatigue risk prediction. *Journal of King Saud University Computer and Information Sciences, 38*(3), 121.

[6] Abasi, A., et al. (2025). Machine learning models for reinjury risk prediction using cardiopulmonary exercise testing (CPET) data: Optimizing athlete recovery. *BioData Mining, 18*(1), 16.

[7] Cao, J., et al. (2026). Local protectionism in the Chinese construction industry: A data analysis of construction contractors' credit scores. *Journal of Civil Engineering and Management, 32*(6), 800–817.

[8] Li, M., et al. (2026). A Kolmogorov-Arnold network enhanced tabular parameter-efficient ensembling model with SHAP analysis for athlete overtraining binary classification. *Engineering Research Express*.

[9] McGuigan, H. E. (2022). *Training load monitoring to manage overtraining syndrome and athlete burnout in applied settings* [Dissertation, Southern Cross University].

[10] Amawi, A. T., Grivas, G. V., & Alkasasbeh, W. J. (2026). Digital twin for Taekwondo athletes: Integrating sports nutrition and psychological readiness using artificial intelligence. *Frontiers in Public Health, 14*, 1822194.