SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(1), 985.
https://doi.org/10.62617/mcb985

Article

Data mining technology for monitoring and physiological and biochemical indicators of football players in different training periods

Shijie Zhao[1], Xueqin Wang[2,*]

[1] College of Sports, Shenyang Normal University, Shenyang 110034, Liaoning, China
[2] College of Sports and Health, Linyi University, Linyi 276000, Shandong, China
*** Corresponding author:** Xueqin Wang, lydx_wxq@sina.com

CITATION

Zhao S, Wang X. Data mining technology for monitoring and physiological and biochemical indicators of football players in different training periods. Molecular & Cellular Biomechanics. 2025; 22(1): 985. https://doi.org/10.62617/mcb985

ARTICLE INFO

Received: 3 December 2024
Accepted: 12 December 2024
Available online: 7 January 2025

COPYRIGHT

Copyright © 2025 by author(s). *Molecular & Cellular Biomechanics* is published by Sin-Chn Scientific Press Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license. https://creativecommons.org/licenses/by/4.0/

Abstract: In the monitoring and analysis of physiological and biochemical indicators of athletes, traditional data mining (DM) technology cannot extract compelling features and laws when processing high-dimensional and complex multivariate data, and the accuracy of the analysis results is low. The lack of real-time monitoring of the dynamically changing physiological state makes it impossible to detect athletes' overtraining or fatigue in time, which affects the training effect and the health of athletes. This paper constructs an improved XGBoost (eXtreme Gradient Boosting) model to clean and normalize the collected physiological and biochemical data, remove outliers and fill in missing values, and construct a variable set representing the characteristics of different training periods to provide high-quality input data for subsequent model analysis. This paper combines the SHAP (SHapley Additive exPlanations) method to quantify the importance of each feature, selects the variables that contribute most to the recognition of the training state to optimize the model input, reduce the model complexity, and improve the computational efficiency. Based on the original XGBoost model, the loss function can be adjusted and the adaptive learning rate mechanism can be added to enable the model better to capture the dynamic changes of physiological and biochemical indicators and improve the prediction accuracy. Combined with the prediction results of the improved model, a real-time monitoring system was designed to track the changes in the physiological state of athletes during different training periods, and to issue an alarm when abnormal trends were detected to assist coaches in adjusting training plans. The experimental results show that in the feature evaluation, three key physiological indicators, namely blood oxygen saturation, blood lactate concentration, and heart rate, are extracted, which reduces the computational complexity of the subsequent model. In the four training stages of the basic period, load period, high-intensity period and recovery period, the loss values of the XGBoost model were approximately 0.5, 0.42, 0.4 and 0.35 respectively. In the monitoring data of 4 batches of football players, with 100 players in each batch, the accuracy rate remained above 0.83 and the response time was below 2 s. The experiment proved the effectiveness of the research model in the monitoring and analysis of physiological and biochemical indicators.

Keywords: data mining technology; football player; physiological and biochemical indicators monitoring; extreme gradient boosting; shapley additive explanations; adaptive learning rate mechanism; dynamic Monitoring System

1. Introduction

The training effect of football players is closely related to health management, and physiological and biochemical indicators are important data for measuring the status of athletes. Real-time monitoring and accurate analysis of the physiological changes of athletes at different training stages [1,2] can effectively avoid sports injuries, optimize training effects, and improve competitive level [3]. How to

efficiently process and analyze these data has become a core issue in modern sports science research.

In the traditional monitoring of athletes' physiological and biochemical indicators, there are still difficulties to be solved, especially in the accuracy and efficiency of high-dimensional, multivariate data processing. Traditional methods rely on manual experience or simple statistical analysis in data feature extraction and pattern discovery [4,5], and cannot effectively deal with complex data structures [6,7]. The physiological responses of athletes at different training stages are highly dynamic and changeable. Traditional methods cannot capture these subtle changes in time [8,9], and are prone to miss key physiological change signals, resulting in delayed detection of problems such as fatigue and overtraining [10]. Due to the high dimensionality of the data, it is difficult for traditional methods to select the most representative features from a large number of variables [11,12], which causes the model to overfit or have information redundancy during training, affecting the accuracy of the prediction results [13]. The physiological and biochemical data of athletes are interfered by noise, missing values and outliers. Traditional data processing methods are difficult to effectively clean up these problems, which in turn affects the reliability of the analysis results [14,15]. Traditional analysis methods are difficult to monitor in real time the changes in physiological status during different training periods, and cannot adjust the training program in time when athletes are overly tired or unwell [16]. Athletes' physical conditions generally change significantly in a short period of time. If these changes cannot be captured in time, it is easy to make the training load inappropriate or the recovery inadequate, increasing the risk of sports injuries [17,18]. Even though some methods use time series-based analysis, they are still difficult to adapt to the frequently changing physiological states during training due to the lack of dynamic optimization mechanisms [19,20]. When processing athletes' physiological and biochemical data, traditional methods not only have problems with insufficient analysis accuracy and poor real-time performance, but also have significant deficiencies in dealing with data noise and outliers, and cannot provide efficient training monitoring and health management support. These problems have prompted the demand for more efficient and accurate analysis methods, which in turn has promoted the application research based on DM (Data Mining) technology.

This paper uses advanced DM technology to improve the monitoring and analysis accuracy of physiological and biochemical indicators of football players in different training periods, and solves the limitations of traditional methods in data processing, feature selection and dynamic change capture by constructing an improved XGBoost model. The collected physiological and biochemical data of athletes can be cleaned, normalized and processed for outliers to ensure data quality and provide reliable input for subsequent model analysis; the SHAP (SHapley Additive exPlanations) method can be combined to quantify the contribution of each feature to the prediction results, select the most influential features, optimize the model input, and improve the accuracy of the prediction. By adjusting the loss function of the XGBoost model and introducing an adaptive learning rate mechanism, this paper further enhances the adaptability of the model to the dynamic changes of physiological and biochemical indicators, allowing the model to reflect the changes in the athletes' physiological state in real time and accurately. A real-time monitoring system can be built, combined with

the improved model, to track the changes in the athletes' physiological state in real time during different training periods, detect abnormal trends in time and issue warnings, help coaches adjust training plans, and optimize athletes' training effects and health management. The research goal of this paper is not only to improve the accuracy and efficiency of data analysis, but also to provide an operational and practical technical means for the training monitoring of football players, so as to promote the long-term development of athletes and the improvement of their competitive performance.

2. Related work

Current DM technology is developing rapidly, and many studies have begun to focus on how to use machine learning and DM methods to analyze athletes' physiological and biochemical indicators, improve training effects and reduce sports injuries. Some studies have applied SVM (Support Vector Machine) and decision tree algorithms to classify and regress athletes' physiological data, attempting to identify athletes' fatigue status and adjust training loads [21,22]. Lei et al. established an athlete heart rate measurement model based on support vector machine combined with an improved algorithm, and used a multi-channel spectral matrix decomposition denoising algorithm to eliminate interference factors, thereby improving the accuracy and efficiency of athlete heart rate measurement [23]. Other studies have attempted to capture the long-term trend of physiological indicators through time series analysis and deep learning methods to provide decision support for training plans [24,25]. Song used an optimized convolutional neural network based on a deep learning model, using a self-adjusting algorithm and an autoencoding method to enhance convolution to ensure successful detection and risk assessment of sports medicine diseases [26]. These studies mostly focus on offline analysis and lack real-time monitoring capabilities. In actual applications, the physiological state of athletes changes dynamically. Traditional methods cannot effectively capture these changes and cannot provide real-time feedback. During training, the risk of fatigue or overtraining cannot be discovered in time, affecting training results and athlete health.

In recent years, some studies have proposed improved DM methods to enhance the adaptability of the model to dynamic physiological changes. Some studies have combined XGBoost's ensemble learning method to try to capture the changes in athletes' physiological state in real time during training [27,28]. Zhao et al. proposed an integrated framework that uses Spark-based big data analysis and the XGBoost algorithm. The framework provides powerful sports medical services, including real-time health monitoring and data-driven insights. It can skillfully manage the large amount of sports data generated during training and activities, promote instant health assessment, and combine XGBoost algorithm for DM to enhance health prediction and recommendation capabilities [29]. Hou and Xue used the spatiotemporal graph convolutional network as the main algorithm, introduced the adaptive graph convolution module and the residual channel attention module, and combined with XGBoost to form the final physical training injury risk assessment model, which improved the accuracy of processing physical training injury risks [30]. These methods improve the accuracy of the model and reduce unnecessary complexity by

effectively extracting and optimizing features of high-dimensional data; however, existing models lack real-time data processing capabilities and are still limited when faced with problems such as noisy data and missing values. Although these methods can improve analysis accuracy to a certain extent, they still have obvious deficiencies in real-time monitoring and feedback.

3. Improve the construction and application of XGBoost model

3.1. Data preprocessing and feature construction

3.1.1. Data processing

In order to ensure the quality of the data, the collected physiological and biochemical data are cleaned. The goal of data cleaning is to remove noise data caused by sensor errors or interference factors in the acquisition process. Here, outlier detection based on the *Z*-score [31,32] method is used to identify extreme values in the data. The *Z*-score method can effectively identify extreme data that deviates far from the mean by calculating the degree of deviation between the data points and the mean. It is particularly suitable for outlier detection in high-dimensional data. This method can eliminate abnormal data points, reduce interference with model training, and improve the accuracy of analysis results. For each variable x_i, its mean μ and standard deviation σ are calculated, and then its Z score is calculated using the following formula:

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

When $|Z_i| > 3$ is used, the data point is considered an outlier and is removed. This method can effectively remove noise data that is far from the overall distribution and ensure the validity of subsequent training data.

In addition to outlier processing, data normalization is also an essential step. Physiological and biochemical indicators of different dimensions need to have the same scale. The Min-Max normalization [33,34] method is used to linearly transform the value of each feature to the [0, 1] interval. The transformation formula is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2}$$

$\min(x)$ is the minimum value of the feature, $\max(x)$ is the maximum value of the feature, and x' is the normalized representation of the feature. This method can eliminate the dimensional differences between different features, so that each feature has the same weight during the model training process, and avoid some features dominating the model training process due to excessive numerical range.

During the physiological data collection process of athletes in different training periods, data loss may occur. The processing of missing values is an important part of data preprocessing. In order to avoid incomplete training data due to missing values, the KNN (K-Nearest Neighbor) [35,36] interpolation method is used to fill in the missing data points. The KNN method calculates the similarity between the row where the missing value is located and other rows, selects the K most similar neighbor values, and uses the average value of these neighbors to fill in the missing data points. In this

process, for a feature x_j in the jth row where the missing value is located, the K most similar samples are selected based on the Euclidean distance between other feature values and the jth row data, and then the filling value is calculated by weighted average:

$$x_i' = \frac{\sum_{j=1}^{K} w_j \times x_j}{\sum_{j=1}^{K} w_j} \tag{3}$$

w_j is the weight of the jth neighbor, calculated by the inverse of the Euclidean distance. In data processing, the KNN algorithm is used to effectively fill missing values and reduce the impact of outliers. The algorithm is based on the distance measurement principle and predicts missing data through the weighted average of neighboring samples, thereby improving the completeness and accuracy of the data.

3.1.2. Feature construction

In physiological and biochemical data, the physiological responses of different training periods vary significantly, so it is necessary to construct a feature set that can effectively characterize these differences. Common physiological and biochemical indicators include cardiovascular system-related indicators such as heart rate, blood pressure, blood oxygen saturation and heart rate variability, and metabolic and energy system-related indicators such as blood lactate concentration, blood glucose concentration, liver function indicators and muscle glycogen reserves. In each training cycle, changes in cardiovascular system-related indicators can reflect the intensity and fatigue of football players. Metabolic and energy system-related indicators are important indicators for evaluating the physical load of football players. Indicators related to muscle damage and recovery help evaluate whether the training load is appropriate and promptly detect whether football players are at risk of overtraining. Endocrine and hormone levels reflect the physical functions of football players.

These characteristic variables not only include data from a single time point, but also introduce dynamic change characteristics during the training period, such as the heart rate change rate, the growth rate of lactate concentration, and the fluctuation range of muscle damage indicators. After analyzing the time series data, the trend of the athlete's physiological state changing with the training intensity is captured. The data of each variable during the training period is processed with time difference, and the data change rate $\Delta x(t)$ at each moment is calculated:

$$\Delta x(t) = x(t) - x(t-1) \tag{4}$$

The construction of this feature can capture the dynamic changes in physiological data and provide richer input features for subsequent model training.

Table 1 shows the frequency of collecting physiological and biochemical characteristics in different training periods: the frequency of collecting is low during the basal period and recovery period, and the frequency of collecting is increased during the load period and high-intensity period, focusing mainly on cardiovascular, metabolic and muscle damage indicators.

Table 1. Data collection frequency.

Training period	Cardiovascular system(time/week)	Metabolism and energy system(time/week)	Muscle damage and recovery(time/week)	Endocrine and hormonal levels(time/week)
Base period	1	1	1	1
Load period	3	2–3	2	2
High intensity period	4	3–4	3	3
Recovery period	2	1–2	2	1

The data processing and feature construction process in **Figure 1** shows the complete process from data collection to feature construction. Data collection enters the data cleaning stage, including using Z-score to remove outliers and KNN to handle missing values. The cleaned data is normalized to ensure the uniform scale of the features. Feature construction introduces dynamic change characteristics during the training period, and performs time difference processing on the data of each variable during the training period to obtain $\Delta x(t)$. This series of processes are interconnected and jointly ensure the high quality of data, which is particularly suitable for real-time monitoring and health management of physiological and biochemical data.

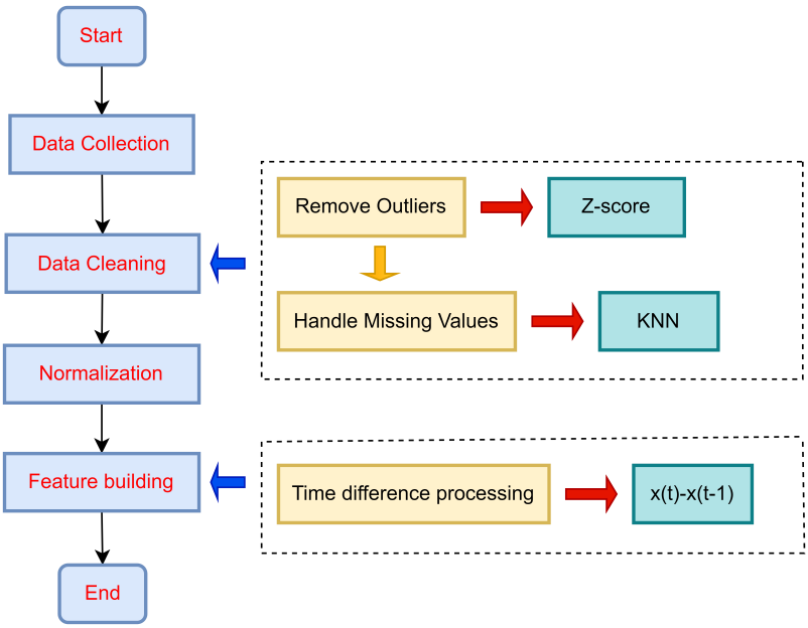

Figure 1. Data processing and feature construction process.

3.2. Feature importance evaluation and screening

In order to optimize the input features of the training state recognition model and reduce the computational complexity, the SHAP [37,38] method is used here to evaluate and screen the importance of features. The SHAP value provides a way to quantify the contribution of each feature to the model output. By analyzing the contribution of each feature, the features that have a greater impact on the prediction of the target variable are identified, and on this basis, they are screened to optimize the model input and remove redundant and irrelevant features.

The core of feature importance assessment is to calculate the marginal

contribution of each feature in different feature combinations through SHAP value. The SHAP method quantifies the importance of each feature based on SHapley value theory. For a specific data point, the SHAP value of its feature represents the contribution of the feature to the model prediction result. The calculation formula is:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{5}$$

N is the feature set, S is the subset of features, $f(S)$ is the predicted value output after training the model on the feature set S, and $f(S \cup \{i\})$ is the output value after adding the features. After calculating the impact of all possible feature subsets on the prediction results, the SHAP value can accurately quantify the contribution of the features to the model output.

Based on the SHAP value of each feature, all features can be sorted and the features with the greatest contribution can be selected for model training. In this process, the average absolute SHAP value of each feature is calculated and sorted to obtain the importance ranking of the features. The quantitative indicator of feature importance is its average absolute SHAP value:

$$\mathrm{SHAP}_{\mathrm{avg}}(x_i) = \frac{1}{N} \sum_{j=1}^{N} \left| \phi_i^j \right| \tag{6}$$

ϕ_i^j is the SHAP value of feature x_i in the jth sample. By calculating the average SHAP value of all features, it can identify the features that contribute more to the model prediction. These features are usually strongly associated with the target variable, and retaining these features can ensure that the model can still maintain a high prediction accuracy in a lower dimension.

When screening features, in addition to relying on the SHAP value sorting, correlation analysis is also required to eliminate redundant features. Highly correlated features can introduce collinearity problems, affecting the stability and performance of the model. The correlation coefficient between features can be calculated to identify highly correlated features and eliminate them, which is a key step in further optimizing input features. Feature correlation is measured by the Pearson correlation coefficient, which is calculated as follows:

$$\rho(x_i, x_j) = \frac{\sum_{n=1}^{N} (x_i^n - \bar{x_i}) \left(x_j^n - \bar{x_j}\right)}{\sqrt{\sum_{n=1}^{N} \left(x_i^n - \bar{x_i}\right)^2 \sum_{n=1}^{N} \left(x_j^n - \bar{x_j}\right)^2}} \tag{7}$$

According to the preset correlation threshold, the redundancy between features is determined, and unnecessary features are eliminated to ensure that the model training is not disturbed by redundant information. The SHAP method eliminates redundant features by quantifying the contribution of features to the model output, while the Pearson correlation coefficient measures the linear relationship between features and helps identify highly correlated features. When the two are used in combination, SHAP helps to screen out important features, and the Pearson correlation coefficient further ensures that there is no multicollinearity between features, thereby improving the stability and interpretability of the model.

3.3. Improved training and optimization of the XGBoost model

3.3.1. Adjusting the loss function to adapt to the dynamic changes of physiological and biochemical indicators

When predicting physiological and biochemical indicators, the traditional XGBoost model uses a fixed loss function, usually a logarithmic loss function. Since physiological and biochemical indicators show complex dynamic trends, a simple fixed loss function may not be able to accurately capture these changes. In order to make the model better adapt to this dynamic change, the loss function of XGBoost is improved here.

Using the weighted loss function, the paper introduces a time weighting mechanism based on the traditional loss function to adjust the contribution of different time points to model training. At a certain moment, the predicted value of t is $\widehat{y_t}$, and the actual value is y_t. In a certain period of time, the accuracy of the prediction has a greater impact on the model. The paper uses the introduced time weight w_t to weight it. The new loss function is expressed as:

$$L(\theta) = \sum_{t=1}^{T} w_t \times (\widehat{y_t} - y_t)^2 \tag{8}$$

w_t is the weight of time t, θ is the model parameter, and T is the total number of time steps. The weight w_t is set according to the dynamic change law of physiological and biochemical indicators. For those indicators with drastic changes, the corresponding time step can be given a larger weight, while for those indicators with gentle changes, the corresponding time step can be given a smaller weight.

The study also introduces the adaptability of adaptive loss function to physiological and biochemical indicators changes; based on the principle of gradient lifting tree model, the loss value of the model can be adaptively adjusted according to the results of the previous round of training; for samples with large errors in the previous iteration, its loss weight in the current iteration is increased. This process can be expressed by the following adaptive loss function formula:

$$L(\theta) = \sum_{t=1}^{T} \alpha_t \times (\widehat{y_t} - y_t)^2 \tag{9}$$

In the formula, α_t is the adaptive adjustment coefficient. As the model training progresses, α_t can automatically increase or decrease according to the error size, allowing the model to focus more on samples that were previously misjudged. By dynamically adjusting the loss function, the model can more effectively learn the changing patterns of physiological and biochemical indicators during the training process and improve prediction accuracy.

3.3.2. Introducing an adaptive learning rate mechanism to improve model stability

In the traditional XGBoost model, the learning rate is a fixed parameter, which is generally small to prevent overfitting. A fixed learning rate may cause the training process to converge too slowly when facing complex data sets, or fail to adapt to the speed of data changes in certain training stages. Based on this situation, the study

improved the original XGBoost model and added an adaptive learning rate mechanism, allowing the model to dynamically adjust the learning rate according to data changes at different training stages.

In this process, a gradient-based adaptive learning rate strategy is introduced. In each tree construction process, the current gradient information is calculated, and the learning rate is dynamically adjusted according to the gradient amplitude. Assuming that in the k round iteration, the current gradient is g_k, the adjustment formula of the adaptive learning rate η_k is:

$$\eta_k = \frac{\eta_0}{1 + \beta \times |g_k|} \tag{10}$$

η_0 is the initial learning rate, β is the adjustment factor, and $|g_k|$ is the absolute value of the current gradient.

The study also introduced a dynamic learning rate adjustment mechanism based on training error to further enhance the effect of adaptive learning rate. After each round of training, the training error is calculated and the learning rate of the next round is adjusted according to the size of the error. The formula is as follows:

$$\eta_k = \eta_0 \times \left(1 - \frac{e_k}{\max(e)}\right) \tag{11}$$

e_k is the training error of the current round, and $\max(e)$ is the maximum error during the training process. Using this method, the model can dynamically adjust the learning rate according to the changes in the training error, more accurately control the update amplitude of each step during the training process, avoid over-adjustment when close to convergence, and ensure that the model is gradually refined on the basis of global optimization to improve the final prediction accuracy.

Table 2 shows the parameter settings of the loss function adjustment and adaptive learning rate mechanism during the XGBoost model training process. As the training progresses, the loss function weight gradually increases, from 0.5 in the early stage to 2.0 in the dynamic adjustment stage, emphasizing the importance of key features. The adaptive learning rate factor β also gradually increases with the progress of training, from 0.1 to 0.4, ensuring that the model can be flexibly adjusted according to error feedback at different training stages, and the initial learning rate remains unchanged to ensure the stability of training. The error threshold gradually increased from 0.02 to 0.15, reflecting the increasing accuracy requirements of the model. The adaptive learning rate η_k was adjusted from 0.045 to 0.052 as the training progressed, optimizing the convergence process of the model.

Table 2. Parameter settings for loss function adjustment and adaptive learning rate mechanism.

Stage/condition	Loss function weighting parameter (w_t)	Adaptive learning rate factor (β)	Initial learning rate (η_0)	Error threshold (e_k)	Adaptive learning rate (η_k)
Early training	0.5	0.1	0.05	0.02	0.045
Mid training	1	0.2	0.05	0.05	0.048
Late training	1.5	0.3	0.05	0.1	0.05
Dynamic adjustment	2	0.4	0.05	0.15	0.052

Figure 2 is the improved XGBoost architecture, which shows the whole process from input to prediction results, divided into four main modules: feature selection, tree model training, model optimization and prediction. The input data is passed to SHAP to quantify the importance of each feature, select key features, and improve the quality of model input. The filtered features are used to train multiple decision trees. Each tree is optimized through residual iteration, combined with an improvement mechanism, a weighted loss function to adjust the weight of rare data, an adaptive learning rate to improve training efficiency, and dynamic regularization to prevent overfitting. The output of each tree is weighted and summarized as the final prediction result. This architecture balances feature extraction, model generalization ability, and prediction accuracy to meet the needs of dynamic physiological and biochemical data prediction.

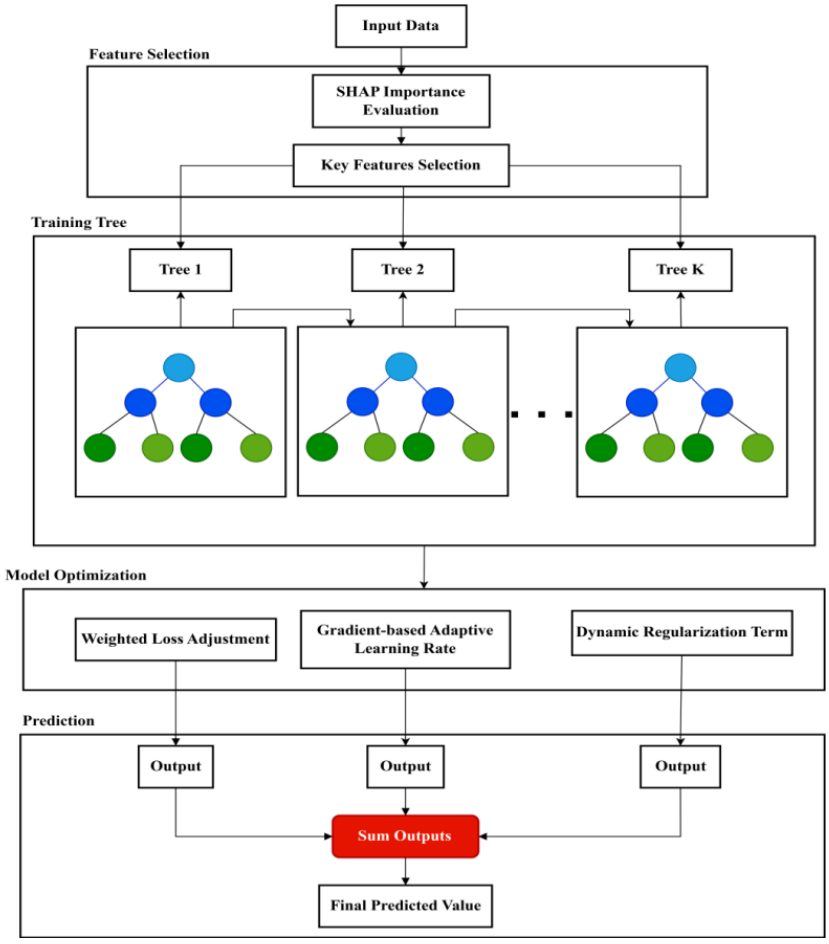

Figure 2. Improved XGBoost architecture diagram.

3.4. Construction of dynamic monitoring system

3.4.1. Real-time data processing and analysis based on prediction results

In order to realize the dynamic monitoring of the physiological state of football players, combined with the prediction results of the improved XGBoost model, a real-time data processing module is designed to analyze the changing trends of physiological indicators in different training periods. The module is built based on time series data flow, and its core includes three parts: data collection, feature update and trend analysis.

In data collection, wearable devices are used to obtain athletes' physiological data in real time. These data are sampled at fixed time intervals and transmitted wirelessly to the central processing unit. In order to reduce transmission and storage costs, a dynamic feature update method based on sliding windows is designed. The sampling time window is T_w, and the feature vector update formula at time t is:

$$\mathrm{X}_t = \frac{1}{T_w} \sum_{i=t-T_w+1}^{t} \mathrm{x}_i \tag{12}$$

X_t is the average eigenvalue in the window, and x_i is the eigenvalue of i at the time point. The sliding window method is used to ensure the real-time nature of the data, reduce the interference of random fluctuations on the model input, and improve the stability of the prediction results.

In the trend analysis stage, the output of the improved XGBoost model not only provides the current prediction value, but also calculates the change rate and abnormal trend indicators in combination with the historical prediction value. The current prediction value of a physiological indicator is $\hat{y}_t$, and the prediction value at the previous time point is $\hat{y}_{t-1}$, then the change rate is defined as:

$$r_t = \frac{\hat{y}_t - \hat{y}_{t-1}}{\hat{y}_{t-1}} \tag{13}$$

The study adopts a systematic anomaly detection method to build an adaptive alarm threshold based on the change rate of the time series; the mean μ_r and standard deviation σ_r of the change rate are calculated using a sliding window, and the abnormal alarm conditions are:

$$r_t > \mu_r + k \times \sigma_r \tag{14}$$

k is the adjustment coefficient, which is determined by experimental tuning. Through the above method, the system can timely identify abnormal fluctuations of physiological indicators during real-time monitoring and provide efficient early warning support for coaches.

3.4.2. Design of real-time alarm and auxiliary decision module

In the monitoring system, the real-time alarm module is a key part to ensure the safety of training. In order to improve the accuracy and response speed of the alarm, the study constructed an alarm mechanism based on multi-indicator fusion, which combines the abnormal trends of multiple physiological indicators to avoid false alarms or missed alarms caused by abnormal single indicators.

A separate alarm weight ω_i is defined for each monitoring indicator, and the weight is determined based on the sensitivity of the indicator to the training state. The calculation formula for the comprehensive alarm score S_t is:

$$S_t = \sum_{i=1}^{n} \omega_i \times \mathbb{I}\left(r_{t,i} > \mu_{r,i} + k \cdot \sigma_{r,i}\right) \tag{15}$$

$\mathbb{I}$ is an indicator function. When the change rate of the ith indicator exceeds the abnormal threshold, the value is 1, otherwise it is 0. According to the comprehensive alarm score S_t, the system sets multi-level alarm response rules.

After the alarm is triggered, the auxiliary decision module automatically generates adjustment plan suggestions. Based on the feature importance analysis of the improved XGBoost model, the system is able to identify the key physiological indicators that cause abnormalities and provide optimization suggestions in combination with the current training stage. When abnormal fluctuations in lactate concentration are detected, the system may recommend reducing the duration of high-intensity training or increasing the rest time between intervals.

The study also designed a graphical interface to enhance the practicality of the system, showing the dynamic curves of athletes' physiological indicators and alarm records in real time. Coaches can intuitively understand the training effect through the interface and quickly adjust the training strategy according to the system's suggestions. This graphical visualization method improves the user-friendliness of the system and provides data-based decision support for coaches.

This paper combines real-time data processing and analysis, alarm mechanism and auxiliary decision-making module. The dynamic monitoring system can track the changes in the athlete's physiological state in multiple dimensions, providing important guarantees for scientific and refined training management. The real-time data processing and analysis module provides an accurate abnormal judgment basis for the alarm mechanism by monitoring the trend of changes in the athletes' physiological data. After identifying the abnormal trend, the alarm mechanism combines with the auxiliary decision-making module to generate personalized adjustment suggestions, forming a closed-loop feedback loop to optimize the training plan and protect the health of athletes.

Table 3. System alarm rules and decision table.

Alert level	Composite score S_t	Trigger condition	Response strategy	Decision advice
Low alert	$0.5 \leq S_t < 1.0$	Slight deviations in rate	Observe, adjust training intensity.	Increase rest time, adjust training cycle.
Medium alert	$1.0 \leq S_t < 1.5$	Significant deviation in rate	Reduce training intensity, increase recovery time.	Optimize training plan, reduce high-intensity training.
High alert	$S_t \geq 1.5$	Major deviation across multiple indicators	Immediately pause training, conduct medical assessment.	Stop current training, adjust entire training plan.

Table 3 evaluates the athlete's physiological state through comprehensive scoring S_t. Three different alarm levels, low, medium and high, can be set according to different ranges of scores, and specific response strategies and auxiliary decision-making suggestions are provided for each alarm level. Low-level alarms correspond to minor abnormalities, which can be solved by adjusting training intensity and rest cycles; medium-level alarms involve more significant physiological abnormalities, which require reducing training intensity and appropriately extending recovery time. Advanced alarms are serious abnormalities in physiological status, and training must be suspended immediately and medical evaluation must be performed to protect the health of athletes. This numerical alarm mechanism can effectively help coaches understand the status of athletes in a timely manner and make corresponding adjustments to ensure the scientificity and safety of training.

4. Method effect evaluation

4.1. Feature importance evaluation

Figure 3 shows the feature importance ranking calculated based on the SHAP value, reflecting the contribution of each physiological and biochemical indicators to the prediction results of the training state recognition model. Each horizontal bar represents the SHAP value of a feature. The higher the SHAP value, the greater the impact of the feature on the model prediction. Features include heart rate, blood oxygen saturation, blood lactate concentration, etc. These physiological indicators usually have a higher weight in athlete training monitoring, so their SHAP values are larger. The SHAP values of indicators such as liver function, blood pressure, and creatine kinase are low, reflecting the relatively low importance of these features in the current model. Through such a visualization, the key indicators with the most predictive power in the training state evaluation can be intuitively identified, helping researchers to optimize feature selection, reduce redundant features, and improve the computational efficiency and prediction accuracy of the model. **Figure 3** clearly shows the contribution of different features in the model, providing a strong basis for further feature screening and model optimization.

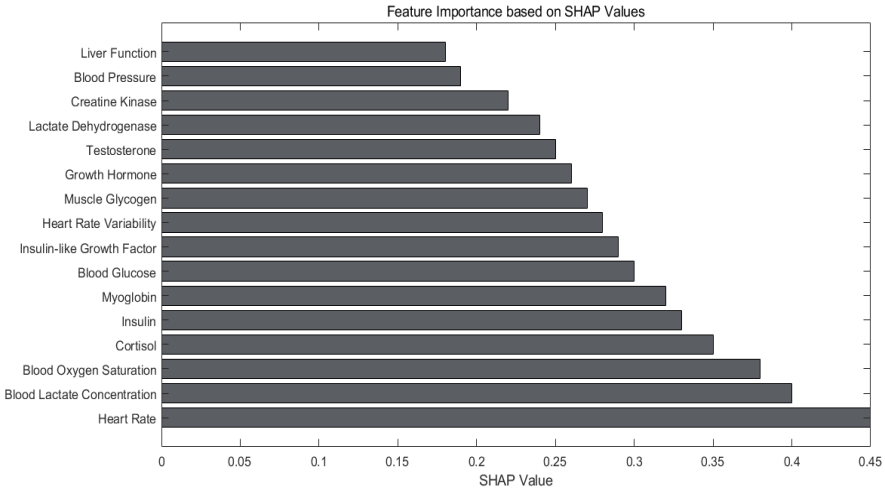

Figure 3. SHAP value ranking diagram.

Figure 4 shows the correlation matrix between six physiological indicators, including myoglobin, insulin, cortisol, blood oxygen saturation, blood lactate concentration, and heart rate. Each value in the matrix represents the Pearson correlation coefficient between the corresponding features, ranging from −1 to +1, reflecting the linear relationship between the variables. The closer the value is to 1, the stronger the positive correlation is. The closer the value is to −1, the stronger the negative correlation is. The closer the value is to 0, the weaker the linear relationship is. The matrix reveals the mutual influence between various physiological indicators. For example, there is a strong positive correlation of 0.85 between myoglobin and insulin, indicating that they may be regulated by similar mechanisms in physiological processes. There is a strong negative correlation between blood lactate concentration and blood oxygen saturation, which is −0.8. The correlation of the features is weak, indicating that they may be independent or affected by different factors in

physiological responses. Finally, according to **Figure 4**, three weakly correlated physiological indicators of blood oxygen saturation, blood lactate concentration and heart rate were selected. **Figure 4** helps to understand the relationship between the features in the modeling process and provides an important reference for feature selection and model optimization.

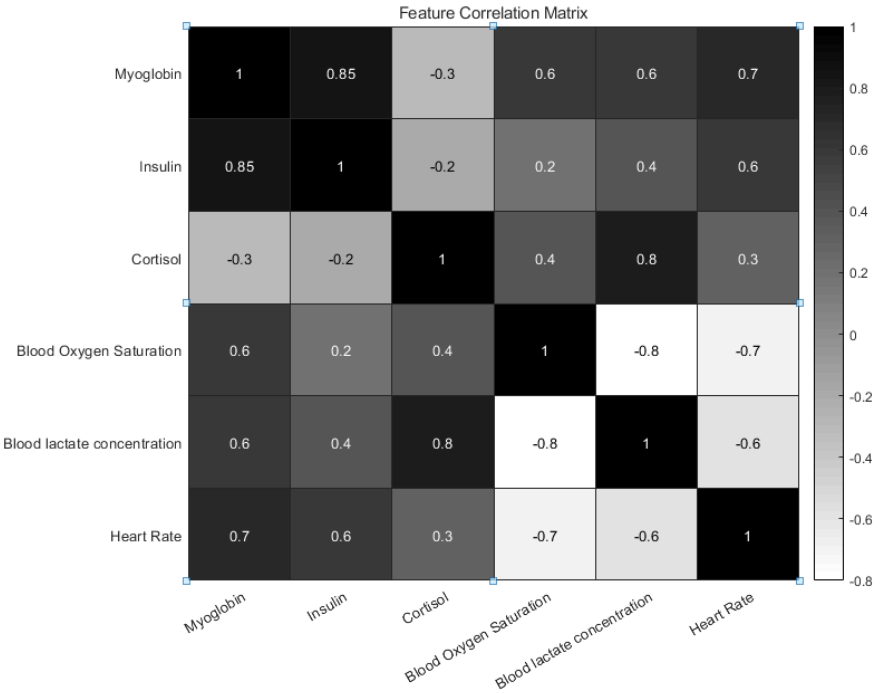

Figure 4. Feature correlation matrix.

4.2. Evaluation of high-dimensional data processing capabilities

In the evaluation of high-dimensional data processing capabilities, the loss function is used as the main evaluation indicator to measure the fitting effect of different models when processing high-dimensional physiological and biochemical data. The evaluation process can cover four training stages: basic period, load period, high-intensity period and recovery period, to comprehensively evaluate the performance of the model under different physiological conditions.

High-dimensional physiological and biochemical data sets from four training stages were selected and processed to form high-dimensional feature vectors as the input of the model. The improved XGBoost model, the traditional XGBoost model and the CatBoost (Categorical Boosting) model were trained to ensure the reliability of the evaluation results. CatBoost is a popular algorithm based on gradient boosting trees, which is specially optimized for classification features. CatBoost was chosen as the experimental comparison object because it performs well in processing high-dimensional data and categorical features, and can effectively reduce data bias and improve the prediction accuracy of the model; comparison with XGBoost is necessary because they are both tree models based on gradient boosting, but CatBoost has unique advantages in feature preprocessing and avoiding overfitting, and can provide more robust analysis of athletes' physiological and biochemical data.

During the evaluation process, the loss value of each model in each training period is calculated to compare its performance in processing high-dimensional data.

The loss value reflects the gap between the model's predicted value and the true value. The smaller the value, the stronger the model's fitting ability. For each training stage, the loss value of each model is evaluated to analyze its ability to process high-dimensional data under different training conditions. Using this series of evaluations, it can comprehensively compare the performance of the improved XGBoost model with other models in high-dimensional data processing and reveal the advantages and disadvantages of each model at different training stages.

Figure 5 shows the performance of the loss values of the three models, namely the improved XGBoost model, the traditional XGBoost model and the CatBoost model, in 100 training rounds in four different training periods: the basic period, the load period, the high-intensity period and the recovery period.

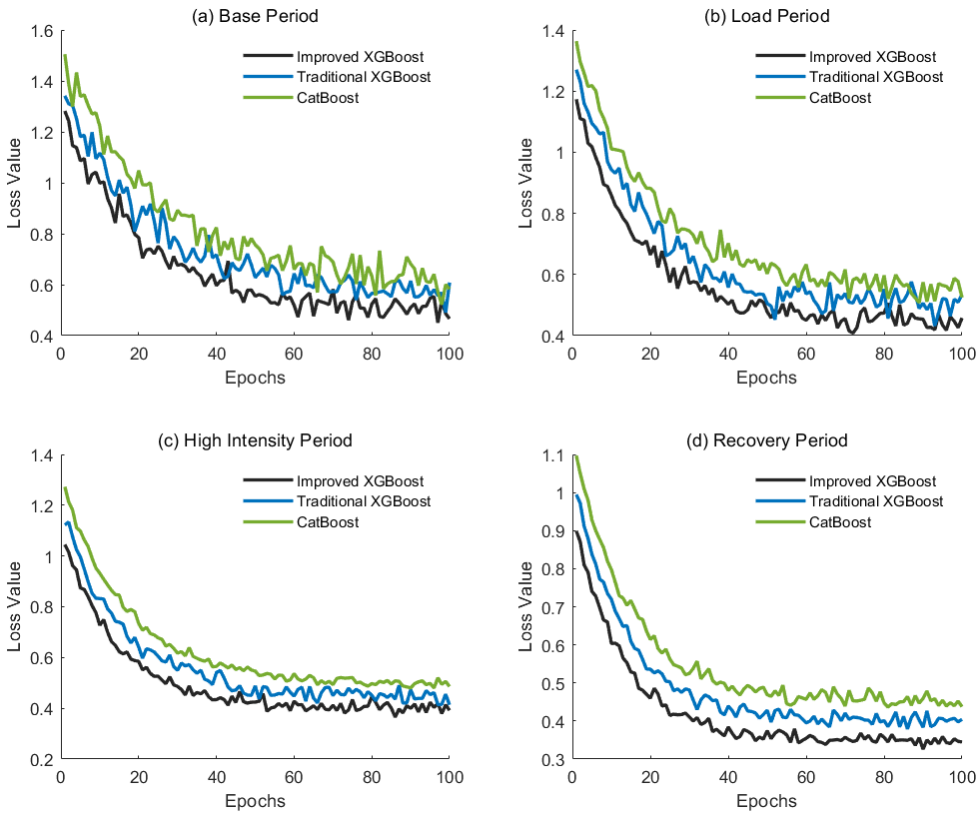

Figure 5. The loss values of different models in different training periods.

In the basic period, the loss values of all three models dropped rapidly from high at the beginning. The loss value of the improved XGBoost model dropped the fastest and finally fluctuated around 0.5, indicating that the model was able to fit the data quickly in the initial training. Although the loss values of the traditional XGBoost model and the CatBoost model also showed a downward trend, the final values were higher than those of the improved XGBoost model. This shows that the improved XGBoost has obvious advantages in data fitting and convergence speed in the early stage of training, and can better capture the patterns of physiological and biochemical data; during the load period, the loss value of the improved XGBoost model is still lower than that of the other two models, and finally tends to 0.42. The model can also adapt well to data changes during this stage; after entering the high-intensity period, the improved XGBoost model still maintains a low loss value, and the downward trend

gradually slows down, and finally tends to 0.4, reflecting its adaptability to high-intensity data during training. The loss values of traditional XGBoost and CatBoost decrease less during this period, and the curve changes are not as stable as the improved XGBoost model. This phenomenon reflects that the improved XGBoost can better capture the data patterns and show higher stability and accuracy when facing the fluctuations of physiological and biochemical indicators during high-intensity training. During the recovery period, the loss value tends to be stable, and the loss values of the three models are close to a stable value. The improved XGBoost model maintains the optimal loss value with the smallest change, and finally tends to 0.35. It is proved that it can stably maintain a high prediction accuracy during the recovery period. Although the loss values of traditional XGBoost and CatBoost tend to be stable, the fluctuation range is relatively large, and it fails to achieve the stability of improved XGBoost. The improved XGBoost model shows strong adaptability during the recovery period and can effectively adapt to the recovery process of physiological state.

The improved XGBoost model always maintains a low loss value during each training period. Compared with traditional XGBoost and CatBoost, it has faster convergence speed and stronger stability, and shows better adaptability and accuracy when processing high-dimensional data.

4.3. Dynamic monitoring accuracy and real-time analysis

In the accuracy and real-time evaluation of the dynamic monitoring system, the experiment selected 4 batches, with 100 athletes in each batch participating in the monitoring. Accuracy and response time were used as core indicators to measure the system's performance in capturing changes in the athletes' physiological state, and the average value was calculated for each batch of athletes. The evaluation covers four training phases: basic period, load period, high intensity period and recovery period, verifying the applicability and stability of the system under different physiological states.

This paper combines the athlete's physiological state records and model prediction results to analyze the system's ability to detect key state changes. The accuracy rate is used as an evaluation indicator to measure the consistency between the system's predicted state and the actual state. In verifying the real-time performance of the system, the response time of the system between processing input data and outputting prediction results is recorded, and high-frequency data input is used to simulate real-time dynamic changes. The response efficiency of the system under different training times is evaluated. The measurement of response time is combined with the characteristics of each training stage to ensure that the evaluation results can fully reflect the real-time processing capabilities of the system. This evaluation process provides data support for the performance optimization of the dynamic monitoring system in practical applications, while revealing its potential for improvement under different training conditions.

Figure 6 shows the accuracy of the dynamic monitoring system in four different training stages. The accuracy of the monitoring of the four groups of athletes participating in the experiment varies, ranging from 0.8 to 1.0. The accuracy rate in

the high-intensity period reached the highest value in all groups of athletes, ranging from 0.91 to 0.94, which shows that the dynamic monitoring system is more accurate in detecting drastic physiological changes. The accuracy rate in the basic period was slightly lower, ranging from 0.83 to 0.86, which may be due to the small fluctuation of physiological state, resulting in unclear data characteristics. The accuracy of the load period and the recovery period was relatively close among the groups, ranging from 0.87 to 0.90 during the load period and 0.86 to 0.89 during the recovery period, indicating that the system also has stable performance when processing physiological data of moderate intensity and gradual recovery. The fluctuation of data accuracy reflects the sensitivity and adaptability of the system to the data characteristics of different training stages. **Figure 6** clearly shows the performance advantage of the dynamic monitoring system under complex physiological conditions, providing data support for subsequent optimization.

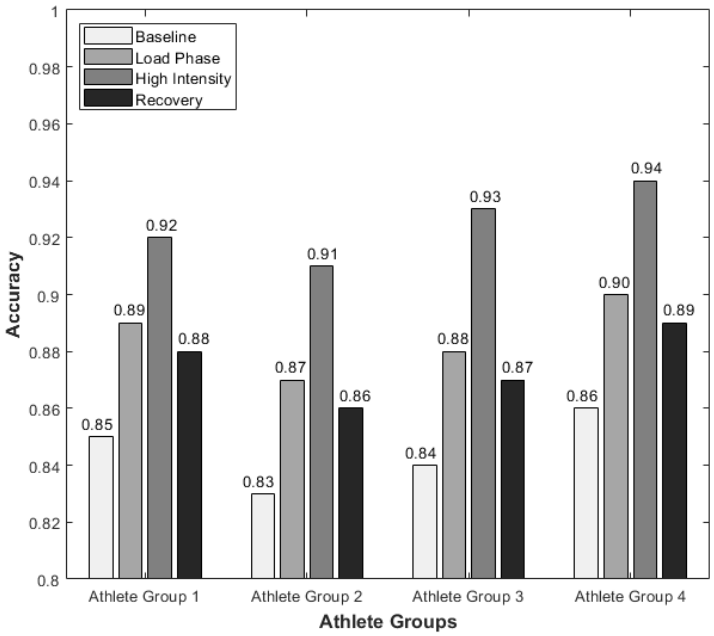

Figure 6. Accuracy of the dynamic monitoring system in different training stages.

Figure 7 shows the response time performance of the dynamic monitoring system at four different training stages, with the response time ranging from 1.0 s to 2.0 s. During the high-intensity period, the response time of athletes in each group reached a peak of 1.7 s to 1.9 s, indicating that the system had a high computational complexity when dealing with drastic changes, and the response time was extended, but within a reasonable range. During the basic period and recovery period, the response time was relatively short, ranging from 1.1 s to 1.3 s and 1.2 s to 1.4 s, respectively, indicating that the data characteristics changed less during this stage and the system's computational burden was lighter. The response time during the load period was between 1.4 s and 1.6 s. The overall trend shows that the response time increases with the increase of physiological state complexity during the training phase, but it still remains within a reasonable range. This fully demonstrates the real-time and stability of the system in complex data processing.

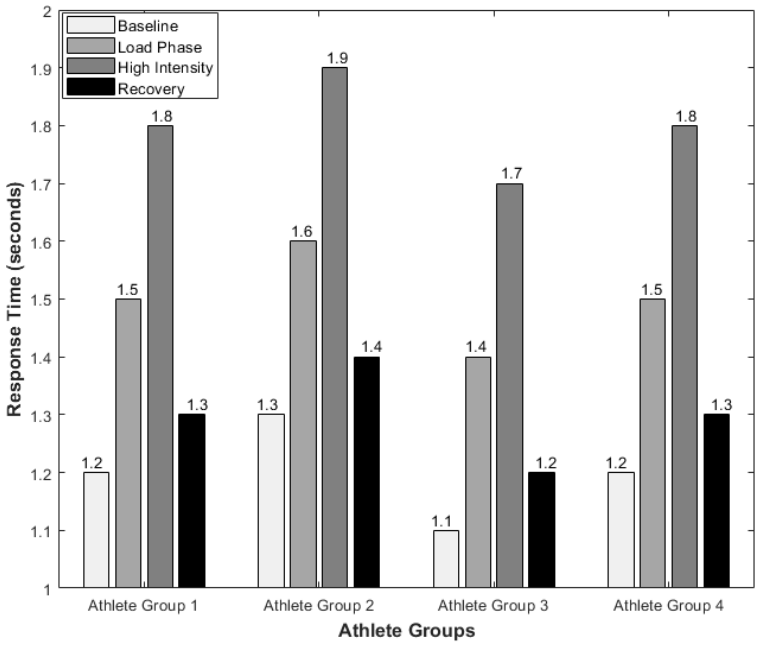

Figure 7. Response time of the dynamic monitoring system at different training stages.

5. Conclusions

Aiming at the monitoring needs of dynamic changes in physiological and biochemical data of football players during training, an improved XGBoost model was constructed, and the system performance was deeply analyzed from three aspects: feature importance evaluation, high-dimensional data processing capability, and dynamic monitoring accuracy. The research results show that the proposed system introduces the SHAP method to quantitatively screen the input variables, which reduces the model complexity and significantly improves the computational efficiency. The improved XGBoost model more accurately obtains the changes in the physiological state of athletes at different training stages by adjusting the loss function and introducing an adaptive learning rate mechanism, providing reliable prediction support for the monitoring system. In the evaluation of high-dimensional data processing capabilities, the improved model showed lower loss values than traditional XGBoost and CatBoost, In the four training stages of base, load, high-intensity, and recovery, the loss values of the XGBoost model were approximately 0.5, 0.42, 0.4 and 0.35, respectively, verified that optimizing algorithms and adjusting input features can effectively improve the model's adaptability to complex physiological changes. In the monitoring data of four batches of football players, with 100 players in each batch, the accuracy remained above 0.83 and the response time was below 2 s. The dynamic monitoring system provides a scientific basis for coaches to formulate and adjust training plans by obtaining abnormal trends in real time during training. The response speed and monitoring accuracy showed significant advantages in the experiment.

The dynamic monitoring system proposed in this paper can efficiently and accurately realize real-time monitoring and prediction of athletes' training status, providing a new technical path for the field of sports science and a useful reference for dynamic monitoring research in other similar fields. Future work can further expand the application scope of the system and combine more biological data sources

to further enhance the generalization ability and applicability of the model.

Author contributions: Data curation, SZ; writing—original draft preparation, SZ; writing—review and editing, XW. 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. Vinnichuk Y D, Polischchuk A O, Goshovska Y V, et al. Changes in biochemical parameters and mitochondrial factor in blood of amateur athletes under influence of marathon running. Fiziol Zh, 2019, 65(5): 20-27.
2. Teleglow A, Marchewka J, Tota L, et al. Changes in the morphological, rheological, and biochemical blood indicators in triathletes. Folia Biologica (Kraków), 2020, 68(3): 107-120.
3. Alina S, Vişinescu A, Caramoci A, et al. Tracking performance in elite athletes. Medicina Sportiva: Journal of Romanian Sports Medicine Society, 2021, 17(1): 3300-3307.
4. Dorofeikov V V, Smirnov M S, Nevzorova T G, et al. Automated biochemical methods to assess muscle and myocardial damage in athletes. Theory and Practice of Physical Culture, 2021 (10): 49-51.
5. Tan X, Song M. Characteristics of physiological changes in athlete training based on the data mining algorithm. Revista Brasileira de Medicina do Esporte, 2022, 28(5): 386-389.
6. Heidari J, Beckmann J, Bertollo M, et al. Multidimensional monitoring of recovery status and implications for performance. International journal of sports physiology and performance, 2019, 14(1): 2-8.
7. Cadegiani F A, Kater C E. Basal hormones and biochemical markers as predictors of overtraining syndrome in male athletes: the EROS-BASAL study. Journal of athletic training, 2019, 54(8): 906-914.
8. Wlodarczyk M, Kusy K, Slominska E, et al. Change in lactate, ammonia, and hypoxanthine concentrations in a 1-year training cycle in highly trained athletes: applying biomarkers as tools to assess training status. The Journal of Strength & Conditioning Research, 2020, 34(2): 355-364.
9. Wlodarczyk M, Kusy K, Slominska E, et al. Changes in blood concentration of adenosine triphosphate metabolism biomarkers during incremental exercise in highly trained athletes of different sport specializations. The Journal of Strength & Conditioning Research, 2019, 33(5): 1192-1200.
10. Isacco L, Degoutte F, Ennequin G, et al. Rapid weight loss influences the physical, psychological and biological responses during a simulated competition in national judo athletes. European journal of sport science, 2020, 20(5): 580-591.
11. Moss S L, Randell R K, Burgess D, et al. Assessment of energy availability and associated risk factors in professional female soccer players. European Journal of Sport Science, 2021, 21(6): 861-870.
12. Nicolas M, Vacher P, Martinent G, et al. Monitoring stress and recovery states: Structural and external stages of the short version of the RESTQ sport in elite swimmers before championships. Journal of Sport and Health Science, 2019, 8(1): 77-88.
13. Skorski S, Mujika I, Bosquet L, et al. The temporal relationship between exercise, recovery processes, and changes in performance. International Journal of Sports Physiology and Performance, 2019, 14(8): 1015-1021.
14. Chamari K, Roussi M, Bragazzi N L, et al. Optimizing training and competition during the month of Ramadan: Recommendations for a holistic and personalized approach for the fasting athletes. Tunis Med, 2019, 97(10): 1095-1103.
15. Podrigalo L, Iermakov S, Romanenko V, et al. Psychophysiological features of athletes practicing different styles of martial arts-the comparative analysis. International Journal of Applied Exercise Physiology, 2019, 8(1): 84-91.
16. Arede J, Ferreira A P, Gonzalo-Skok O, et al. Maturational development as a key aspect in physiological performance and national-team selection in elite male basketball players. International Journal of Sports Physiology and Performance, 2019, 14(7): 902-910.
17. Souza R A, Beltran O A B, Zapata D M, et al. Heart rate variability, salivary cortisol and competitive state anxiety responses during pre-competition and pre-training moments. Biology of sport, 2019, 36(1): 39-46.
18. Walker A J, McFadden B A, Sanders D J, et al. Biomarker response to a competitive season in division I female soccer

players. The Journal of Strength & Conditioning Research, 2019, 33(10): 2622-2628.
19. Kruger K, Reichel T, Zeilinger C. Role of heat shock proteins 70/90 in exercise physiology and exercise immunology and their diagnostic potential in sports. Journal of Applied Physiology, 2019, 126(4): 916-927.
20. Horta T A G, Bara Filho M G, Coimbra D R, et al. Training load, physical performance, biochemical markers, and psychological stress during a short preparatory period in Brazilian elite male volleyball players. The Journal of Strength & Conditioning Research, 2019, 33(12): 3392-3399.
21. Simsek M, Kesilmis I. Predicting athletic performance from physiological parameters using machine learning: Example of bocce ball. Journal of Sports Analytics, 2022, 8(4): 299-307.
22. Marynowicz J, Lango M, Horna D, et al. Predicting ratings of perceived exertion in youth soccer using decision tree models. Biology of sport, 2022, 39(2): 245-252.
23. Lei T, Cai Z, Hua L. Training prediction and athlete heart rate measurement based on multi-channel PPG signal and SVM algorithm. Journal of Intelligent & Fuzzy Systems, 2021, 40(4): 7497-7508.
24. Ding Y. Analyzing Athletes' Physical Performance and Trends in Athletics Competitions Using Time Series Data Mining Algorithms. Journal of Electrical Systems, 2024, 20(9s): 736-746.
25. Liu Y, Ji Y. Target recognition of sport athletes based on deep learning and convolutional neural network. Journal of Intelligent & Fuzzy Systems, 2021, 40(2): 2253-2263.
26. Song H, Montenegro-Marin C E. Secure prediction and assessment of sports injuries using deep learning based convolutional neural network. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(3): 3399-3410.
27. Yigit A T, Samak B, Kaya T. An XGBoost-lasso ensemble modeling approach to football player value assessment. Journal of Intelligent & Fuzzy Systems, 2020, 39(5): 6303-6314.
28. Kim J Y, Kim J H, Kang E W, et al. The Prediction of Dry Weight for Chronic Hemodialysis Athletes Using a Machine Learning Approach: Sports Health Implications. Revista de Psicología del Deporte (Journal of Sport Psychology), 2024, 33(1): 68-82.
29. Zhao Y, Ramos M F, Li B. Integrated framework to integrate Spark-based big data analytics and for health monitoring and recommendation in sports using XGBoost algorithm. Soft Computing, 2024, 28(2): 1585-1608.
30. Hou Z, Xue Y. Sports training injury risk assessment combined with dynamic analysis algorithm. Molecular & Cellular Biomechanics, 2024, 21(3): 484-484.
31. Schober P, Mascha E J, Vetter T R. Statistics from A (agreement) to Z (z score): a guide to interpreting common measures of association, agreement, diagnostic accuracy, effect size, heterogeneity, and reliability in medical research. Anesthesia & Analgesia, 2021, 133(6): 1633-1641.
32. Henderi H, Wahyuningsih T, Rahwanto E. Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. International Journal of Informatics and Information Systems, 2021, 4(1): 13-20.
33. Kappal S. Data normalization using median median absolute deviation MMAD based Z-score for robust predictions vs. min–max normalization. London Journal of Research in Science: Natural and Formal, 2019, 19(4): 39-44.
34. Gokhan A, Guzeller C O, Eser M T. The effect of the normalization method used in different sample sizes on the success of artificial neural network model. International journal of assessment tools in education, 2019, 6(2): 170-192.
35. Abu Alfeilat H A, Hassanat A B A, Lasassmeh O, et al. Effects of distance measure choice on k-nearest neighbor classifier performance: a review. Big data, 2019, 7(4): 221-248.
36. Cunningham P, Delany S J. K-nearest neighbour classifiers-a tutorial. ACM computing surveys (CSUR), 2021, 54(6): 1-25.
37. Meng Y, Yang N, Qian Z, et al. What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values. Journal of Theoretical and Applied Electronic Commerce Research, 2020, 16(3): 466-490.
38. Hamilton R I, Papadopoulos P N. Using SHAP values and machine learning to understand trends in the transient stability limit. IEEE Transactions on Power Systems, 2023, 39(1): 1384-1397.