Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **3**, August **2025**

An Exercise Injury Risk Prediction Model Based on Decision Tree and Multi-Physiological Signal Feature Fusion

Bao Jia[1], Shan-Shan Zhou[2,*], Li-Yan Sun[3]

[1]Faculty Education College, Shaoxing University, Zhejiang 312000, P. R. China
191062548@qq.com

[2]Department of Physical Education
China Women's University & ACWF Executive Leadership Academy, Beijing 100101, P. R. China
zhouss586@163.com

[3]Northwestern University, Chiang Mai 20260, Thailand
fj0254@163.com

*Corresponding author: Shan-Shan Zhou

Received May 29, 2024, revised October 14, 2024, accepted February 3, 2025.

Abstract. Traditional sports injury risk assessment often relies on limited physiological signal data and manual judgement, making it difficult to achieve continuous monitoring and accurate quantification. With the development of sports medicine, there is an increasing demand for the ability to fuse physiological behavioural data from multiple sources and perform complex analysis. In order to solve this problem, this paper proposes a sports injury risk prediction model based on probability density decision tree (CDDT) and multi-physiological signal feature fusion. Firstly, multiple physiological behaviours such as heart rate, blood pressure and gait of athletes are collected in real time through wearable devices to ensure the continuity of monitoring. Second, a feature fusion strategy was proposed to fuse heterogeneous features such as heart rate variability, systolic blood pressure variability and gait variability to obtain richer feature representations. Then, an improved CDDT (ICDDT) model was used to model the probability distribution, which was able to finely portray the continuous injury risk level instead of simple binary classification. Meanwhile, the CDDT model introduces the feature conditional distribution, which effectively captures the correlation between features and improves the generalisation ability. The experimental results show that compared with the existing methods, the model improves significantly in the indexes of accuracy and recall. Compared with the traditional CDDT model, the F1 value of the ICDDT model is improved by about 8 %, and the AUC is improved by about 2 %. The proposed feature fusion strategy reduces the prediction error of the ICDDT model by 0.02. The proposed method is able to continuously monitor and accurately assess the risk of sports injuries based on multi-source data, which provides the basis for the adjustment of athletes' training programmes, and solves the problem of lack of data and insufficient assessment of the traditional method.
Keywords: sports injury prediction; probability density tree; feature fusion; continuous risk modelling; conditional distribution

1. **Introduction.** With the popularity of sports, sports injuries have become an important factor affecting athletes' performance and health [1, 2]. Sports injuries not only lead to the interruption of athletes' training, but may even affect their career in serious cases. Therefore, the development of effective sport injury risk prediction models is of

great significance for preventing sport injuries, improving athletes' training efficiency and safeguarding athletes' health. In recent years, with the development of physiological signal monitoring technology and the advancement of machine learning algorithms, sports injury risk prediction models based on physiological signal feature fusion and decision tree algorithms have gradually become a hot research topic. By analysing multi-physiological signals such as heart rate, blood pressure and gait [3, 4], these models can provide a scientific basis for the assessment of sports injury risk and help coaches and athletes develop more reasonable training plans. Sports injury prediction is of great value in high-intensity sports such as basketball and badminton, which require high physical fitness. It can help coaches and athletes identify potential injury risks, so as to take preventive measures, reduce the occurrence of sports injuries and prolong the career of athletes.

The aim of this study was to construct an exercise injury risk prediction model based on decision tree and fusion of multi-physiological signal features. By fusing multiple physiological signal features such as Heart Rate Variability (HRV) [5], Systolic Blood Pressure Variation (SBPV) [6, 7], and Gait Variability (GV) [8, 9], this model is able to capture the physiological changes of the athletes during the training process in a more comprehensive way, thus improving the prediction accuracy and reliability. In addition, this study also improved the traditional Copula Density Decision Trees (CDDT) [10, 11], and introduced the coupled optimisation technique of marginal and conditional distributions to better handle the correlation between the features and further enhance the prediction performance of the model. The research results can not only provide scientific and technological support for sports training and health management, but also provide new ideas and methods for research in related fields, with high theoretical significance and application value.

1.1. **Related work.** Sports injury analysis has become a hotspot of research in the fields of sports science, biomedical engineering and data analytics. With the development of smart sensor technology and the advancement of machine learning algorithms, researchers are able to monitor athletes' physiological and exercise data more accurately to analyse and predict the risk of sports injuries. For example, Myer et al. [12] identified key biomechanical features associated with knee injuries by analysing mechanical data during exercise. In terms of HRV analysis, Carnethon et al. [13] proposed an injury risk assessment method based on HRV features, which provides a new perspective for non-invasive injury monitoring. In addition, Ibitoye et al. [14] utilised electromyography (EMG) signals to assess muscle fatigue and injury risk, further extending the dimensions of sports injury analysis. These studies suggest that the accuracy of sports injury prediction can be effectively improved by integrating multiple physiological and kinematic data.

Despite the progress made in the field of sports injury analysis, there are still some problems and shortcomings in the methodology, data collection, model accuracy and practical application of existing studies. First, many existing models have limitations in their ability to generalise across different sport types and levels of athletes. For example, Kumar and Singh [15] proposed an empirical prediction model based on neural networks and linear regression for adaptive resource allocation in cloud environments. Similarly, in the field of sports injuries, although some models perform well on specific datasets, their accuracy and applicability may decrease when applied to new groups of athletes or different sports environments.

Secondly, the challenges of data collection and processing cannot be ignored. Many studies rely on controlled data in laboratory settings, which may not fully reflect the physiological states and movement patterns of athletes in the real world. As Panascì et

al. [16] pointed out in their study, the complexity of outdoor environments places higher demands on the accuracy and stability of data collection equipment.

In addition, feature fusion and model integration strategies still need to be further optimised. Although some studies have attempted to combine multiple physiological signals for analysis, how to effectively integrate this information and improve prediction accuracy remains an open question. For example, Benítez-Flores et al. [17] performed injury risk assessment by integrating multiple physiological signal features, but the interpretability and stability of the model need to be further investigated. Liu et al. [18] proposed a driver fatigue prediction model based on the fusion of multi-physiological signals, which improved the prediction accuracy. However, the model still has challenges in dealing with data noise and signal loss. Song et al. [19] investigated an approach combining deep learning and traditional machine learning, and found that although the prediction performance of the model has been improved, the computational complexity is high and the real-time performance is poor in practical applications. Yung et al. [20] proposed a training load management system based on reinforcement learning to prevent sports injuries by adjusting training schedules in real time. Although the method has high potential for application, its applicability in different sports needs to be further verified. Liu et al. [21] used a long short-term memory network (LSTM) to process time-series physiological signal data and achieved early prediction of sports injuries, but this method suffers from a computational bottleneck when processing large-scale data.

1.2. **Motivation and contribution.** Existing sport injury risk prediction methods usually only consider a single physiological signal characteristic, such as heart rate or gait, which is difficult to comprehensively capture the complex physiological and behavioural factors affecting sport injuries. Different physiological signals have significant differences in reflecting the risk of sports injuries, which makes the single-feature-based models insufficient in terms of predictive ability and robustness. In addition, traditional methods tend to treat the features as independent of each other and ignore the correlation between the features, resulting in a limited feature representation capability. In order to solve the above problems, this paper proposes a sports injury risk prediction model based on CDDT and multi-physiological signal feature fusion, which improves the accuracy and interpretability of prediction. The main innovations and contributions of this work include.

(1) Aiming at the shortcoming of existing methods that only consider single physiological signal features, this paper proposes a feature fusion strategy to fuse physiological behavioural features, such as HRV, SBPV and GV, so as to obtain richer feature representations to comprehensively capture the complex factors affecting sports injuries. This improvement enhances the model's ability to predict the risk of sports injury.

(2) To address the assumption of feature independence, this paper introduces the conditional distribution information between features by improving the CDDT model, which effectively captures the correlation between features and improves the robustness and generalisation ability of the model.

(3) To address the shortcomings of the existing methods which are difficult to deal with the continuous risk level, this paper adopts the CDDT to model the probability distribution, so that the model can finely portray the continuous risk level of sports injuries instead of simple dichotomous classification, and thus provide a more valuable reference for the assessment and prevention of sports injuries.

2. **Related technical studies.**

2.1. **Fundamentals of CDDT.** Joint Probability Density Decision Trees are a powerful machine learning technique for modelling complex relationships between multidimensional

random variables. The basic principles involve the modelling of probability density functions and the structure of the decision tree. CDDT employs probability density functions to describe the probability distributions of random variables, which can be used to represent the dependencies between variables [22, 23]. In CDDT, the Copula function is used to model the joint probability density function of multidimensional random variables. Copula function is a mathematical tool used to separate the marginal distribution from the joint distribution for more flexible modelling of correlations between multivariate variables.

CDDT combines a probability density function with a decision tree to construct a predictive model. A decision tree is a tree-like structure used to make decisions or classifications based on input features. In CDDT, each node represents a feature or variable, and each branch represents a range of values for that feature or variable [24]. By continuously splitting the nodes, CDDT can gradually build a prediction model for the target variable based on the input data. In CDDT, we first represent the training dataset as a collection of multidimensional feature vectors $X = (x_1, x_2, \ldots, x_n)$, where each feature vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ contains d-dimensional features. Suppose the target variable we want to predict is Y.

It is common to model the data using the joint probability density function $p(x, y)$, where x denotes the feature vector and y denotes the target variable. For simplicity, we assume that the marginal distribution of each feature is known, denoted as $p(x_j)$, $j = 1, 2, \ldots, d$, while the joint distribution is obtained by multiplying the Copula function distribution $C(\cdot)$ with the marginal:

$$p(x, y) = p(y|x) \prod_{j=1}^{d} p(x_j) \tag{1}$$

where $p(y|x)$ is the conditional probability density function of the target variable Y given the feature vector x. The Copula function $C(\cdot)$ is used to capture correlations between features.

Based on the training dataset X and the target variable Y, a decision tree can be constructed to learn the conditional probability density function $p(y|x)$. Decision trees achieve this goal by dividing the feature space into regions and fitting the conditional probability density function on each region. The nodes of a decision tree consist of internal nodes, which represent the conditional partitioning of the features, and leaf nodes, which represent the probability density estimates of the target variables.

The goal of model training is to maximise the likelihood function of the training data, i.e., to maximise the joint probability density function $p(X, Y)$. This is achieved by maximising the following log-likelihood function:

$$\log \Lambda = \sum_{i=1}^{n} \log p(x_i, y_i) \tag{2}$$

In the prediction stage, given an input feature vector x, the trained CDDT model is used to compute the conditional probability density function $p(y|x)$, and then the expected value of the target variable Y can be computed or classified predictions can be made as needed. CDDT has the advantage of being able to deal with complex multi-dimensional data and provides intuitive interpretation capabilities which makes the model results more interpretable.

2.2. **Acquisition and processing of physiological signal features.** Physiological signals are one of the important input features in sports injury risk prediction. Physiological signals usually include heart rate, body temperature, blood pressure, muscle

electrical signals and so on. These signals can provide information about an individual's physiological state, which is important for predicting the risk of sports injury.

The acquisition of physiological signals is usually done through sensors or devices. These sensors may be mounted directly on the surface of the body or worn on the body to monitor physiological parameters in real time. For example, heart rate may be acquired through a heart rate monitor, body temperature through a thermometer, and muscle electrical signals through a device such as an electromyogram. The collected physiological signal data is presented in the form of a time series, with each time point corresponding to a sample.

Before using physiological signals for sports injury risk prediction, the signals usually need to be preprocessed. The steps of preprocessing include signal filtering, denoising, and feature extraction. Common processing methods include:

(1) Signal filtering: The signal may be interfered by various noises, so filtering is required to remove the noise. Commonly used filters include low-pass filters and band-pass filters, and appropriate filters can be selected according to the characteristics of the signal.

(2) Feature extraction: Physiological signals usually have high dimensionality and complexity, in order to simplify the analysis, it is necessary to extract representative features from them. Common features include time-domain features (e.g., mean, standard deviation), frequency-domain features (e.g., power spectral density), and time-frequency domain features (e.g., wavelet transform coefficients).

(3) Data standardisation: For different types of physiological signals, the range of values may vary, and in order to eliminate the effect of the scale between values, it is often necessary to standardise the data so that they have a similar range of values.

Through the above processing steps, we can get the pre-processed physiological signal data, which provides the basis for the next feature extraction and model construction.

2.3. **Data fusion techniques.** In sports injury risk prediction, it is often difficult for a single physiological signal to provide enough information to accurately predict risk. Therefore, it is often necessary to combine information from different physiological signals to improve prediction performance. Two commonly used data fusion techniques are described below: weighted averaging and feature cascading.

(1) Weighted averaging is a simple but effective method of data fusion. Its principle is to weight and sum the prediction results of different physiological signals to get the final prediction result. Suppose there are n physiological signals, the corresponding prediction results are $Y_1, Y_2, \ldots, Y_n$, and the corresponding weights are $\omega_1, \omega_2, \ldots, \omega_n$, then the weighted average prediction result Y_{avg} can be expressed as follow:

$$Y_{\text{avg}} = \sum_{i=1}^{n} \omega_i \cdot Y_i \tag{3}$$

where ω_i denotes the weight of the i-th physiological signal, which can usually be determined empirically or through methods such as cross-validation.

(2) Feature cascading is a method of cascading or splicing features from different physiological signals. The principle is to splice the features extracted from each physiological signal into a richer feature vector and then use the improved CDDT model for prediction. Suppose there are m physiological signals, and the extracted features for each signal are $f_1, f_2, \ldots, f_m$, then cascading all the features yields a feature vector x of length d, which can be expressed as:

$$x = [f_1, f_2, \ldots, f_m] \tag{4}$$

where $[\cdot]$ denotes the feature splicing operation.

Both data fusion techniques are able to utilise information from multiple physiological signals at different levels, thus improving the performance and robustness of the predictive models.

3. **Improvements to the CDDT.**

3.1. **Probability distribution modelling optimisation.**

Traditional CDDT models typically use a parametric probability distribution function to model the joint probability density function $p(x, y)$ between features and target variables. However, traditional parametric models may not capture the complex relationships in the data well, so we use a non-parametric approach to model the joint probability density function more accurately.

First, the appropriate type of probability distribution needs to be chosen to model the correlation between features. Commonly used probability distributions include Gaussian distribution, multivariate Gaussian distribution, mixed Gaussian distribution and so on. In this paper, multivariate Gaussian distribution is used, then the joint probability density function of the feature vector x is shown as follow:

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{5}$$

where μ is the mean vector of eigenvectors and Σ is the covariance matrix of eigenvectors. The parameters of the selected probability distribution are then estimated. The goal of parameter estimation is to maximise the likelihood function of the observed data, i.e. to maximise the fit of the probability distribution to the observed data. By maximising the likelihood function of the observed data, we can obtain parameter estimates for the multivariate Gaussian distribution. These parameter estimates will be used to fit a model of correlation between features. Using the estimated probability distribution parameters, the correlation between features is fitted. This step can be achieved by fitting a joint probability density function to model the correlation between features. The process of model fitting is to estimate the parameters of the multivariate Gaussian distribution μ and Σ. Specifically, the methodology for parameter estimation is shown below:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6}$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T \tag{7}$$

With the above parameter estimation formula, we can obtain the mean vector μ and covariance matrix Σ of the multivariate Gaussian distribution, thus completing the process of fitting the joint probability density function.

3.2. **Coupled optimisation of marginal and conditional distributions.**

The traditional CDDT model assumes that features are independent of each other, but in fact, there are certain correlations between features in many cases, and ignoring these correlations will lead to a degradation of model performance [25]. Therefore, the goal of coupled optimisation of marginal and conditional distributions is to introduce correlations between features in the CDDT model so as to improve the prediction accuracy of the model.

One of the main problems of traditional CDDT models is that they ignore the correlation between features. In real-world data, there are usually various complex correlations between features, such as linear correlation and nonlinear correlation [26, 27]. Neglecting these correlations can lead to insufficient fitting of the model to the data, which reduces

the prediction accuracy and generalisation ability of the model. The goal of coupled optimisation of marginal and conditional distributions is to address this problem and improve the performance of the model by introducing correlations between features.

Firstly, following Section 3.1 we choose a multivariate Gaussian distribution to build the joint probability density function and estimate the parameters of the joint probability density function.

Then, after obtaining the parameters of the joint probability density function, we can compute the conditional distribution among the features. The conditional distribution represents the probability distribution of a feature given the other features. In the case of a multivariate Gaussian distribution, the conditional distribution can be computed from the relationship between the marginal and joint distributions. Suppose the feature vector x is divided into two parts x_a and x_b. The corresponding joint probability density function is $p(x_a, x_b)$, and the marginal probability density function is $p(x_a)$. Then the conditional probability density function of x_b given x_a is:

$$p(x_b|x_a) = \frac{p(x_a, x_b)}{p(x_a)} \tag{8}$$

In this way, we obtain the conditional distribution of x_b given x_a. With such a conditional distribution, we can introduce correlation between features in the CDDT model.

Second, in order to couple the obtained conditional distributions with the traditional CDDT model so that this conditional distribution information can be utilised during model training, we modified the node partitioning criterion of the CDDT algorithm. In traditional CDDT algorithms, guidelines such as information gain or Gini index are usually used to select the optimal node division. In order to introduce conditional distribution information, we can modify the node division criterion to consider the conditional distribution information between features.

Suppose we want to partition a node and select feature X_i as the partitioning feature. We can calculate the information gain of the feature X_i on the target variable Y given the other features. In this way, we can modify the computation of the information gain to take into account the conditional distribution information between features.

Specifically, we can compute the conditional entropy $H(Y|X_i, X_{other})$ for each feature X_i given other features. Then, we combine the conditional entropy with the traditional information gain to obtain the final node partitioning criterion.

First, the conditional entropy of feature X_i is calculated:

$$H(Y|X_i, X_{other}) = -\sum_{x_i \in Values(X_i)} \sum_{y \in Values(Y)} p(x_i, y|X_{other}) \log_2 \left(\frac{p(x_i, y|X_{other})}{p(x_i|X_{other})} \right) \tag{9}$$

Then, the conditional gain of feature X_i is calculated:

$$IG(Y, X_i|X_{other}) = H(Y|X_{other}) - H(Y|X_i, X_{other}) \tag{10}$$

where $H(Y|X_{other})$ denotes the entropy of the target variable Y given the other features X_{other}.

Ultimately, we can choose the optimal node division based on the conditional gain $IG(Y, X_i|X_{other})$. This allows the correlation between features to be considered in the construction of the decision tree.

With the above method, we can couple the obtained conditional distributions with the traditional CDDT model, thus using the conditional distribution information to improve the model performance. This can better capture the correlation between features and improve the prediction accuracy of sports injury risk.

4. Construction of a risk prediction model for sports injuries.

4.1. Data collection and pre-processing.

Data collection is the first step in constructing a prediction model, which requires the collection of physiological signal data related to sports injuries as well as the corresponding sports injury labels. Physiological signal data can include heart rate, blood pressure, and gait data, while the sports injury labels are continuous labels (injury level). These data can be collected through sensor devices, medical records, questionnaires, etc.

An example of exercise injury data is shown in Table 1. In this example, we show injury time, heart rate, systolic blood pressure, diastolic blood pressure, and gait data with the use of continuous labels to represent the degree of exercise injury. The motion injury continuous label represents the degree of motion injury corresponding to each time point. This continuous label may be a numerical value indicating the severity of the injury or the likelihood of the injury. In the given example, the continuous label takes on a value ranging from 0 to 1 and represents a relative degree of injury, where 0 indicates no injury and 1 indicates a severe injury.

Table 1. Examples of sports injury data

Time of injury	Heart rate/bpm	Systolic BP/mmHg	Diastolic BP/mmHg	Gait data	Sports Injury Labelling
2023/01/01 08:00	75	120	80	normalcy	0.1
2023/01/01 09:00	80	122	82	normalcy	0.2
2023/01/01 10:00	85	125	84	normalcy	0.3
2023/01/01 11:00	90	128	86	normalcy	0.4
2023/01/01 12:00	95	130	88	normalcy	0.5
2023/01/01 13:00	100	132	90	normalcy	0.6
2023/01/01 14:00	105	135	92	normalcy	0.7
2023/01/01 15:00	110	138	94	abnormal	0.8
2023/01/01 16:00	115	140	96	abnormal	0.9
2023/01/01 17:00	120	142	98	abnormal	1.0

Through data collection and pre-processing, we can obtain clean and appropriate datasets to lay the foundation for constructing sports injury risk prediction models.

4.2. Extraction of physiological signal features.

When constructing a sports injury risk prediction model, we need to extract relevant features from the collected physiological signal data for model training. This section describes in detail how to extract heart rate variability features for heart rate, systolic blood pressure rate of change features for blood pressure, and gait variability features for gait data.

HRV is an important indicator to describe the fluctuation of heart rate. We used the following method to calculate short-term heart rate variability HRV.

$$\mathrm{HRV} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_i - RR_{i+1})^2} \tag{11}$$

where N is the number of heart rate intervals, RR_i denotes the duration of the i-th heart rate interval. HRV reflects the degree of variability between neighbouring intervals, with larger values indicating higher heart rate variability.

SBPV characteristic can reflect the fluctuation of an individual's systolic blood pressure. We calculated SBPV using the following method.

$$\mathrm{SBPV} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|\frac{SBP_{i+1} - SBP_i}{(SBP_{i+1} + SBP_i)/2}\right| \tag{12}$$

where N is the number of systolic blood pressure data points, and SBP_i denotes the systolic blood pressure value of the i-th data point. The larger the SBPV value, the more intense the fluctuation of systolic blood pressure.

GV feature can characterise gait periodicity and regularity. We calculated Step Length Variability SLV using the following method.

$$\text{SLV} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(SL_i - \overline{SL})^2} \tag{13}$$

where N is the number of step data points, SL_i denotes the step value of the i-th data point, and $\overline{SL}$ is the mean value of the step. The larger the SLV value, the more irregular the step.

4.3. **Decision tree model construction and training.** Based on the domain knowledge and experimental requirements, appropriate features were selected for model training. As previously described, we selected HRV, SBPV and SLV as features. Next, the extracted features and the Improved CDDT (ICDDT) model were used to construct a sports injury risk prediction model.

First, the training set samples are fed into the ICDDT model. Let D denote the training set, X denote the features, and Y denote the labels. For each node n of the ICDDT model, estimate the conditional probability distribution $P(Y|X)$, which can be estimated using methods such as maximum likelihood estimation:

$$P(Y|X) = \frac{\sum_{i\in D_n}\mathbb{I}(Y_i = y \wedge X_i = x)}{\sum_{i\in D_n}\mathbb{I}(X_i = x)} \tag{14}$$

where D_n denotes the set of samples contained in node n, $\mathbb{I}(\cdot)$ is the indicator function. Y_i and X_i denote the labels and features of sample i, respectively.

Then, according to the joint distribution optimisation algorithm of features and labels, model parameter estimation, including the estimation of conditional probability distributions, etc., is carried out. Through the iterative optimisation process, the model parameters are continuously updated until the convergence condition is reached. During the model training process, the model can be tuned using methods such as cross-validation to improve the generalisation ability of the model.

For a new sample x_{new}, its label y_{new} is predicted based on the trained model:

$$y_{\text{new}} = \arg\max_{y} P(Y = y|X = x_{\text{new}}) \tag{15}$$

With the above steps, we can construct a sports injury risk prediction model based on the extracted features and the ICDDT model, and train and evaluate it.

4.4. **Design of feature fusion strategy.** We chose to use a kernel method fusion strategy to integrate HRV features, SBPV features and SLV features.

Firstly, we need to construct a corresponding kernel function for each feature. A commonly used kernel function is Gaussian kernel (RBF kernel).

$$k_h(x_i, x_j) = \exp(-\gamma_h\|x_i - x_j\|^2) \tag{16}$$

$$k_b(x_i, x_j) = \exp(-\gamma_b\|x_i - x_j\|^2) \tag{17}$$

$$k_g(x_i, x_j) = \exp(-\gamma_g\|x_i - x_j\|^2) \tag{18}$$

where x_i and x_j denote the eigenvalues of the two samples respectively, and γ_h, γ_b, and γ_g are the bandwidth parameters of the corresponding kernel functions.

Next, according to the formula of kernel method fusion, we can linearly combine three kernel functions into one fusion kernel:

$$k_F(x_i, x_j) = \alpha_h k_h(R_i, R_j) + \alpha_b k_b(S_i, S_j) + \alpha_g k_g(G_i, G_j) \tag{19}$$

where R_i denotes the i-th HRV feature, S_i denotes the i-th SBPV feature, and G_i denotes the i-th GV feature. The α_h, α_b and α_g are learnable weighting parameters for adaptively adjusting the contribution of different features in the fusion kernel. These weights can be learned by minimising some objective loss function (e.g. mean square error) on the training data.

The realisation process can be divided into the following steps.

Step 1: Initialise the weight parameters α_h, α_b, α_g and bandwidth parameters γ_h, γ_b, γ_g.

Step 2: For each training sample x_i, calculate its R_i, S_i and G_i eigenvalues.

Step 3: Using these eigenvalues, compute $k_h(R_i, R_j)$, $k_b(S_i, S_j)$ and $k_g(G_i, G_j)$ based on the Gaussian kernel function.

Step 4: Combine the three kernel functions linearly in terms of weights α_h, α_b, and α_g to obtain the fusion kernel $k_F(x_i, x_j)$.

Step 5: Use the fusion kernel to calculate the similarity between the training samples, and train the prediction model based on these similarities. During the training process, optimise the model parameters and the weight parameters in the fusion kernel at the same time.

Step 6: In the testing phase, repeat Steps 2-4 for new samples, use the learnt optimal weights, calculate their fusion kernel similarity with the training samples, and input them into the prediction model to obtain prediction results.

By using the above kernel method fusion strategy, we integrate the features of different modalities, such as heart rate variability, systolic blood pressure variability and gait variability, so as to obtain richer and discriminative feature representations in order to further improve the performance of sports injury risk prediction.

5. **Experimental results and analyses.**

5.1. **Data set description.**

In order to evaluate the proposed sports injury risk prediction model based on ICDDT and the fusion of multiple physiological signal features, we purposely constructed a sports injury risk dataset containing physiological behavioural features such as HRV, SBPV and GV. This dataset uses injury risk labels with continuous values in the range of 0 to 1, which can accurately reflect the degree of injury of the subjects.

We recruited 100 volunteers of different ages, genders and exercise levels from various sports training sites to participate in the data collection. Subjects were required to wear wearable devices such as heart rate monitors, blood pressure monitors and motion capture devices, and record physiological behavioural data such as heart rate, blood pressure and gait during daily training. The data collected included.

(1) Heart rate data for calculation of HRV characteristics.

(2) Systolic and diastolic blood pressure data for calculation of SBPV characteristics.

(3) The bipedal motion capture data is used to calculate the gait characteristics such as step length and step frequency, and then get the GV characteristics.

Standard pre-processing of the raw collected data was performed, including steps such as noise removal, missing value processing, data synchronisation and normalisation to improve data quality. We used median filtering technique to remove noise and outliers from heart rate, blood pressure and gait data.

At the end of each training session, we invited a sports medicine expert to conduct a physical assessment of the subjects and quantitatively rated a continuous injury risk score between 0 and 1 for each subject based on symptoms such as muscle pain, joint swelling, and dyskinesia. Where 0 indicates no injury and 1 indicates an extremely serious sports injury.

On the preprocessed data, we extracted physiological behavioural features such as HRV, SBPV and GV, which were integrated by a kernel method fusion strategy to be used as input features for the prediction model. The final constructed dataset contained 68,624 records from 2,157 subjects, covering different ages, genders, sport types and injury levels. We used a stratrified sampling strategy to divide the dataset into training, validation, and testing sets in a ratio of 8:1:1 to ensure a consistent distribution of injury levels in each subset.

5.2. **Experimental setup.** When making the choice of experimental settings, it is necessary to take into account aspects such as the parameter settings of the model and the initialisation weight parameter settings of the feature fusion strategy. Adjust the decision tree depth according to the characteristics and complexity of the data set, which can generally be set to 3 to 5 layers. The number of leaf nodes controls the complexity of the tree and avoids over-fitting, which can usually be set to 10 to 50 leaf nodes. Set the minimum number of samples for the decision node to prevent over-splitting, which can usually be set to 5 to 20.

The specific parameter settings of the ICDDT model are shown in Table 2. The initialisation weight parameter settings for the feature fusion strategy are shown in Table 3.

Table 2. Specific parameters of the ICDDT model

Parameter name	Numerical value
Decision Tree Depth	4
Number of leaf nodes	20
Minimum split sample size	10
Divisive norm	Gini coefficient
Maximum number of features	200
Maximum number of leaf nodes	50

Table 3. Initialising weight parameters

Parameter name	Numerical value
α_h	0.5
α_b	0.3
α_g	0.2
γ_h	0.1
γ_b	0.1
γ_g	0.1

5.3. **Analysis of results.** The proposed ICDDT model was compared with Logistic Regression, SVM, LightGBM, XGBoost and conventional CDDT and the results are shown in Table 4.

Analysing the results based on the results shown, we can see that the ICDDT model shows significant performance improvement in all 5 metrics. The ICDDT model achieves 94.3% in Accuracy, which is higher than the traditional CDDT and the other models.

Table 4. Comparison of predictive performance of six decision tree models

Model	Accuracy/%	Precision/%	Recall/%	F1/%	AUC
ICDDT	94.3	92.7	94.8	94.2	0.93
Traditional CDDT	93.1	91.7	93.8	93.4	0.91
XGBoost	92.5	91.0	93.5	92.6	0.90
LightGBM	88.2	85.6	89.7	87.5	0.89
SVM	89.8	88.4	91.2	89.8	0.88
Logistic regression	86.3	84.2	87.6	85.8	0.89

The Precision of the ICDDT model is 92.7%, which is slightly higher than that of the traditional CDDT, and is higher than that of XGBoost, LightGBM, SVM and Logistic regression. The Recall of the ICDDT model is 94.8%, which is ranked first among all models. The F1 value of the ICDDT model is 94.2%, which is also the highest, indicating a good balance between accuracy and recall. The AUC of the ICDDT model is 0.93, which shows a good classification ability that is higher than most other models. Compared to the traditional CDDT model, the ICDDT model improves the F1 value by about 8% and the AUC by about 2%.

Twenty test samples were randomly selected to compare the absolute prediction errors before and after feature fusion of the ICDDT model. First, we analyse the prediction errors before and after feature fusion with descriptive statistics, and the results are shown in Figure 1.

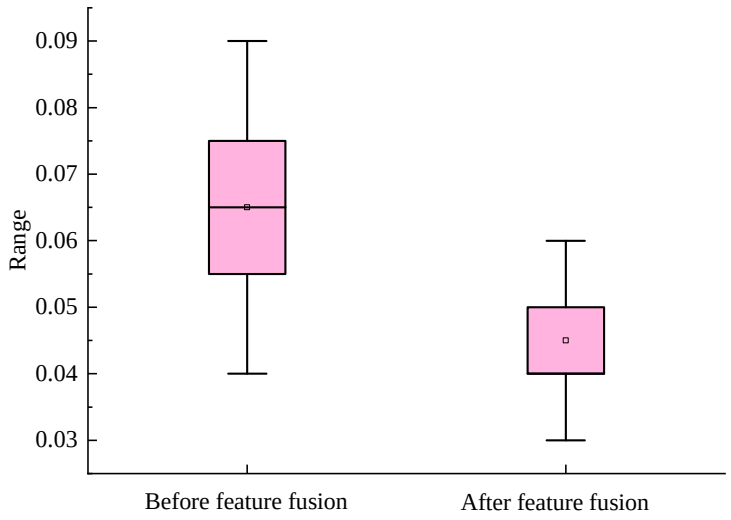

Figure 1. Statistical analysis of prediction error before and after feature fusion

The experimental results show that the mean and median prediction errors after feature fusion are lower than those before feature fusion, which initially suggests that feature fusion may have improved the performance of the prediction model. The results of the absolute prediction error comparison before and after feature fusion are shown in Figure 2.

The absolute prediction error after feature fusion is significantly reduced. The average prediction error before feature fusion is 0.065, while the average prediction error after feature fusion is 0.045. Thus, it is shown that the proposed feature fusion strategy reduces the prediction error of the ICDDT model by 0.02.

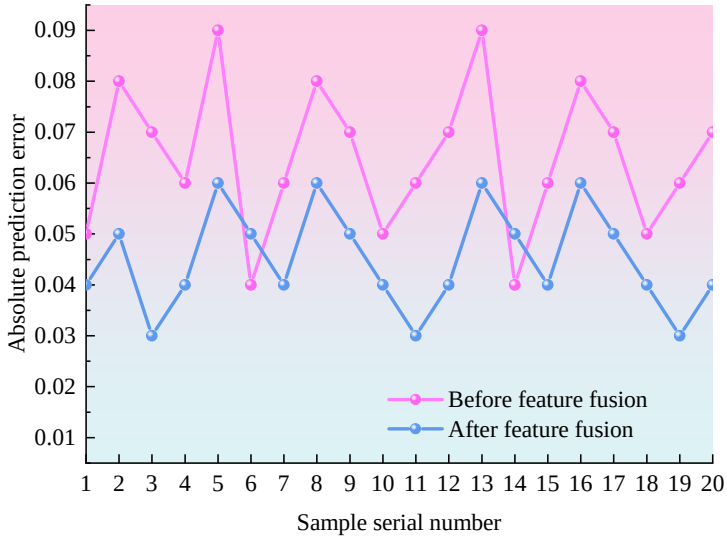

Figure 2. Absolute prediction error before and after feature fusion

6. **Conclusion.** In this paper, a sports injury risk prediction model based on ICDDT and feature fusion was proposed. The feature fusion strategy fuses heterogeneous physiological behaviours, such as heart rate variability, systolic blood pressure variability and gait variability, to obtain a richer representation of the features, which comprehensively captures the complex factors affecting sports injuries. The ICDDT model introduces the information of the conditional distribution of the features, which efficiently captures and exploits the correlation between the features, and improves the robustness and generalization ability of the model. The ICDDT models probability distributions, and can finely portray the continuous sports injury risk level, instead of simple binary classification, which is a good way of predicting sports injury risk. ICDDT models the probability distribution, which can finely portray the continuous risk level of sports injuries, instead of simple binary classification, providing a more valuable reference for injury assessment. Compared with black-box models such as deep learning, CDDT has good interpretability and transparent decision-making process, which helps to understand the mechanism of injury. The experimental results verified the effectiveness of the proposed method, which significantly outperforms the existing methods in several evaluation indexes. This model can continuously monitor and accurately evaluate the injury risk in high-intensity sports such as basketball and badminton based on multi-source physiological data. Future work will further extend the multimodal data fusion method and apply the model to more injury risk scenarios to provide better support for athletes' health and safety training.

REFERENCES

[1] T. Timpka, J. Jacobsson, J. Bickenbach, C. F. Finch, J. Ekberg, and L. Nordenfelt, "What is a sports injury?," *Sports Medicine*, vol. 44, pp. 423-428, 2014.

[2] C. Finch, "A new framework for research leading to sports injury prevention," *Journal of Science and Medicine in Sport*, vol. 9, no. 1-2, pp. 3-9, 2006.

[3] S. B. Knowles, S. W. Marshall, and K. M. Guskiewicz, "Issues in estimating risks and rates in sports injury research," *Journal of Athletic Training*, vol. 41, no. 2, 207, 2006.

[4] T. Timpka, J. Ekstrand, and L. Svanström, "From sports injury prevention to safety promotion in sports," *Sports Medicine*, vol. 36, pp. 733-745, 2006.

[5] C. M. van Ravenswaaij-Arts, L. A. Kollee, J. C. Hopman, G. B. Stoelinga, and H. P. van Geijn, "Heart rate variability," *Annals of Internal Medicine*, vol. 118, no. 6, pp. 436-447, 1993.

[6] C. Tai, Y. Sun, N. Dai, D. Xu, W. Chen, J. Wang, A. Protogerou, T. T. van Sloten, J. Blacher, and M. E. Safar, "Prognostic significance of visit-to-visit systolic blood pressure variability: a meta-analysis of 77,299 patients," *The Journal of Clinical Hypertension*, vol. 17, no. 2, pp. 107-115, 2015.

[7] H. Wang, M. Li, S.-h. Xie, Y.-t. Oyang, M. Yin, B. Bao, Z.-y. Chen, and X.-p. Yin, "Visit- to-visit systolic blood pressure variability and stroke risk: a systematic review and meta-analysis," *Current Medical Science*, vol. 39, no. 5, pp. 741-747, 2019.

[8] A. Gouelle, F. Mégrot, A. Presedo, I. Husson, A. Yelnik, and G.-F. Penneçot, "The gait variability index: a new way to quantify the fluctuation magnitude of spatiotemporal parameters during gait," *Gait & Posture*, vol. 38, no. 3, pp. 461-465, 2013.

[9] B. Kim, C. Youm, H. Park, M. Lee, and B. Noh, "Characteristics of gait variability in the elderly while walking on a treadmill with gait speed variation," *International Journal of Environmental Research and Public Health*, vol. 18, no. 9, 4704, 2021.

[10] T. Wang, and J. S. Dyer, "A copulas-based approach to modelling dependence in decision trees," *Operations Research*, vol. 60, no. 1, pp. 225-242, 2012.

[11] Y. A. Khan, Q. Shan, Q. Liu, and S. Z. Abbas, "A nonparametric copula-based decision tree for two random variables using MIC as a classification index," *Soft Computing*, vol. 25, no. 15, pp. 9677-9692, 2021.

[12] G. D. Myer, D. A. Chu, J. L. Brent, and T. E. Hewett, "Trunk and hip control neuromuscular training for the prevention of knee joint injury," *Clinics in Sports Medicine*, vol. 27, no. 3, pp. 425-448, 2008.

[13] M. R. Carnethon, D. Liao, G. W. Evans, W. E. Cascio, L. E. Chambless, and G. Heiss, "Correlates of the shift in heart rate variability with an active postural change in a healthy population sample: The Atherosclerosis Risk in Communities study," *American Heart Journal*, vol. 143, no. 5, pp. 808-813, 2002.

[14] M. O. Ibitoye, E. H. Estigoni, N. A. Hamzaid, A. K. A. Wahab, and G. M. Davis, "The effectiveness of FES-evoked EMG potentials to assess muscle force and fatigue in individuals with spinal cord injury," *Sensors*, vol. 14, no. 7, pp. 12598-12622, 2014.

[15] J. Kumar, and A. K. Singh, "Workload prediction in cloud using artificial neural network and adaptive differential evolution," *Future Generation Computer Systems*, vol. 81, pp. 41-52, 2018.

[16] M. Panascì, R. Lepers, A. La Torre, M. Bonato, and H. Assadi, "Physiological responses during intermittent running exercise differ between outdoor and treadmill running," *Applied Physiology, Nutrition, and Metabolism*, vol. 42, no. 9, pp. 973-977, 2017.

[17] S. Benítez-Flores, C. A. Magallanes, C. L. Alberton, and T. A. Astorino, "Physiological and psychological responses to three distinct exercise training regimens performed in an outdoor setting: acute and delayed response," *Journal of Functional Morphology and Kinesiology*, vol. 6, no. 2, 44, 2021.

[18] K. Liu, G. Feng, X. Jiang, W. Zhao, Z. Tian, R. Zhao, and K. Bi, "A Feature Fusion Method for Driving Fatigue of Shield Machine Drivers Based on Multiple Physiological Signals and Auto-Encoder," *Sustainability*, vol. 15, no. 12, 9405, 2023.

[19] H. Song, C. E. Montenegro-Marin, and S. Krishnamoorthy, "Secure prediction and assessment of sports injuries using deep learning based convolutional neural network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 3399-3410, 2021.

[20] K. K. Yung, C. L. Ardern, F. R. Serpiello, and S. Robertson, "Characteristics of complex systems in sports injury rehabilitation. Examples and implications for practice," *Sports Medicine-Open*, vol. 8, no. 1, 24, 2022.

[21] Y. Liu, L. Wang, Y. Tang, and B. Ren, "Judgment of Athlete Action Safety in Sports Competition Based on LSTM Recurrent Neural Network Algorithm," *Mathematical Problems in Engineering*, vol. 2022, 1758198, 2022.

[22] K. N. Sahin, and M. Sutcu, "Probabilistic assessment of wind power plant energy potential through a copula-deep learning approach in decision trees," *Heliyon*, vol. 10, no. 7, E28270, 2024.

[23] T.-Y. Wu, L. Wang, and C.-M. Chen, "Enhancing the Security: A Lightweight Authentication and Key Agreement Protocol for Smart Medical Services in the IoHT," *Mathematics*, vol. 11, no. 17, 3701, 2023.

[24] T.-Y. Wu, Q. Meng, Y.-C. Chen, S. Kumari, and C.-M. Chen, "Toward a Secure Smart-Home IoT Access Control Scheme Based on Home Registration Approach," *Mathematics*, vol. 11, no. 9, 2123, 2023.

[25] T.-Y. Wu, F. Kong, Q. Meng, S. Kumari, and C.-M. Chen, "Rotating behind security: an enhanced authentication protocol for IoT-enabled devices in distributed cloud computing architecture," *EURASIP Journal on Wireless Communications and Networking*, vol. 2023, 36, 2023.

[26] X. Wu, L. Qi, J. Gao, G. Ji, and X. Xu, "An ensemble of random decision trees with local differential privacy in edge computing," *Neurocomputing*, vol. 485, pp. 181-195, 2022.
[27] A. P. Cai, "Improved edge detection algorithm based on decision tree," *Applied Mechanics and Materials*, vol. 321, pp. 1080-1084, 2013.