ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Application of multimodal deep learning in action pattern recognition and sports injury prediction of football players

Zhigang Li[1,*]

[1] Jilin Sport University, No2476, Ziyou Road, A Changchun City

SUMMARY: *Aiming at the problems of action recognition relying on a single video feature, injury risk prediction lag and insufficient explanation of training load in soccer training, an action pattern recognition and sports injury prediction model based on multi-modal deep learning was constructed. The system fuses video skeleton key points, IMU-GPS motion sequences, heart rate, RPE, training load and previous injury records to form a dataset of 126 football players, 18420 action clips and 1260 weekly risk samples. The model extracts joint coordination features by spatio-temporal graph convolution, analyzes the changes of exercise load by sequence network, and introduces an attention mechanism to complete multimodal fusion. The experimental results show that the accuracy of action recognition of the full model is 94.6%, Macro-F1 is 93.8%, the AUC of damage prediction is 0.921, and the recall rate of high-risk damage is 88.7%. Compared with the model that removes key modes, the proposed method is more stable in scenes such as complex confrontation, emergency stop and landing buffer, and can provide data support for football training monitoring, action correction and injury warning.*

KEYWORDS: *multimodal deep learning; Football player; Action pattern recognition; Sports injury prediction*

1 Introduction

In recent years, professional football training has gradually shifted from experience-led to data-driven, and athletes' movement performance, training load changes and injury risk identification have attracted more and more attention [1-3]. Football has the characteristics of high confrontation, high speed and frequent change of direction. Actions such as sprint, emergency stop, turn, jump, landing and physical confrontation occur continuously in competition and training. If there is compensation, left-right asymmetry or excessive load accumulation in the action mode, it is easy to cause muscle strain, ligament injury and joint overload [4-6]. Traditional training monitoring mostly relies on coach observation, post-match statistics and single physical performance index, which is difficult to capture the dynamic relationship between movement details and injury risk in time, and it is also difficult to explain the differences in the execution process of athletes in different positions and different load levels.

With the development of computer vision, wearable sensors and deep learning technologies, football player action recognition and injury prediction have a more fine-grained data basis [7-10]. GPS and inertial sensors can record speed, acceleration, deceleration, running distance and changing direction intensity. Heart rate, fatigue perception,

*13321412144@163.com
https://doi.org/10.65102/is20261001
Published: 30 April, 2026

training load and previous injury information can reflect the internal state of athletes. The time synchronization, feature extraction and multi-modal fusion of these data are helpful to identify potential damage signs from the chain of "action performance-load variation-risk accumulation", and improve the intelligent analysis ability in the process of football training.

This paper focuses on the task of action pattern recognition and sports injury prediction of soccer players, and constructs a multi-modal deep learning analysis framework. The framework takes video posture features, motion trajectory features, training load features and physiological state features as input, uses spatio-temporal feature extraction network to identify typical action patterns such as sprint, emergency stop, change direction, shooting, fight for the top and confrontation, and estimates the training-related injury risk of athletes in the next week through a fusion prediction module. The attention mechanism is introduced in the model design to dynamically weight the contributions of different modalities in different time Windows, so that the system can distinguish different risk sources such as transient action abnormalities, continuous load increases and insufficient recovery.

The main work of this paper is reflected in three aspects: first, a multi-modal data processing flow for soccer training scenarios is constructed, and video, sensors, load and injury records are incorporated into a unified modeling framework. The second is to design a collaborative model of action pattern recognition and damage prediction, so that the action recognition results can further serve the risk judgment. Third, the effectiveness of the model is verified by ablation experiments, classification performance comparison and comprehensive application effect analysis. The research results can provide computer-aided basis for football players' training monitoring, technical action adjustment and injury warning, and provide reference for the application of multimodal deep learning in competitive sports.

2 Related Research

In recent years, football training monitoring and sports injury prediction have gradually shifted from single statistical analysis to machine learning modeling. Majumdar et al. [1] focused on the problem of football injury understanding and prediction, pointing out that machine learning can extract risk features from training load, match exposure, physical state and historical injuries, and provide data support for injury warning. Bullock et al. [2] systematically analyzed the methodological quality of sports injury prediction models and believed that the existing models still had shortcomings in sample size, variable selection, external validation and generalization ability. Page et al. [3] analyzed the injury changes of professional men's soccer from the perspective of schedule intensity, indicating that the interval between matches, cumulative fatigue and insufficient recovery would affect the incidence of injury. Silva et al. [4] further pay attention to the acceleration and deceleration requirements in soccer training, and emphasize that high-intensity direction change, emergency stop and repeated sprints are important contents of external load assessment. These studies provide a basis for football injury risk analysis, but most of them stay at the level of load indicators or schedule factors, and the discussion of action pattern details and multi-source data fusion is still insufficient.

In terms of football intelligent analysis, Rico-Gonzalez et al. [5] summarized the application of machine learning in football technical and tactical analysis, sports performance evaluation and risk identification, and pointed out that data quality and task scene adaptation would directly affect the model effect. Nassis et al. [6] further emphasize that the application of machine learning to injury risk in soccer requires the integration of training load, athletic performance, physical response, and previous injuries into the analysis. Robles-Palazon et al.

[7] used machine learning methods to predict the injury risk of male youth soccer players, and proved that age, training exposure, physical ability and load fluctuation have an impact on the prediction results. Haller et al. [8] analyzed the risk of injury and disease through the comprehensive monitoring data of elite youth football players for three months, indicating that continuous monitoring is more suitable for sports risk judgment than a single test. Pillitteri et al. [9] reviewed the relationship between external load, internal load and damage, and found that there were complex correlations between different load indicators. Tsilimigkras et al. [10] combined machine learning with training load analysis to improve the accuracy of football injury risk assessment. The above studies illustrate that football injury prediction cannot rely on a single variable, but should extract joint features from the training process, body state and action performance.

At the same time, football action recognition research is also developing rapidly in the field of computer vision. Giancola et al. [11] proposed the idea of active learning for action localization in soccer videos to reduce the cost of dense labeling. Mkhallati et al. [12] constructed a soccer game video description task to enable the model to understand event semantics from video clips. Held et al. [13] used multi-view video to support automatic soccer decision, indicating that multi-view information can improve the reliability of complex scene recognition. Xarles et al. [14] proposed an action localization Transformer model for soccer videos, which strengthened the modeling of long-term action relationships. Denize et al. [15] applied self-supervised learning and knowledge distillation to soccer action localization, which improved the representation ability of the model under the condition of limited labeling. These methods have promoted the development of soccer video understanding, but most of them focus on game event recognition, action location or auxiliary decision, and pay insufficient attention to the transfer relationship between individual athlete action patterns and injury risk.

Multimodal deep learning provides a new technical path to solve the above problems. Psaltis et al. [16] studied the multimodal representation method for 3D human action recognition in federated scenarios, indicating that different sensing modalities can be complementary to each other. Pajak et al. [17] applied UWB and inertial sensors to physical activity recognition and proved that the fusion of position data and inertial data could improve the effect of action classification. Mekruksavanich et al. [18] adopted the convolutional attention structure to recognize sports and daily activities, which improved the model's ability to capture key motion segments. Muller et al. [19] used IMU time series CNN to complete fitness activity recognition, demonstrating the application value of sensor sequence in action classification. Mekruksavanich and Jitpattanakul[20] combined EMG and IMU sensors to identify training actions, indicating that muscle activation information can supplement posture and motion trajectory features. It can be seen that there is a strong complementarity among video, skeletal keypoints, IMU, GPS, training load and physiological signals, which is suitable for the joint modeling of action pattern recognition and injury prediction of soccer players. In order to more clearly present the differences of related studies in terms of research objects, technical paths and shortcomings, this paper summarizes and compares the existing typical studies, as detailed in Table 1.

Table 1: Comparison of key contents of existing studies

Research Source	Research Content	Main Advantage	Limitation	Relationship to This Study
Majumdar et al. [1]	Machine learning-based football injury prediction	Demonstrates that machine learning is suitable for injury risk modeling	Insufficient use of action details	Provides a basis for the injury prediction task
Bullock et al. [2]	Methodological evaluation of sports injury prediction models	Identifies issues in model validation and generalization	Does not construct a specific football multimodal model	Provides constraints for model evaluation
Page et al. [3]	Fixture congestion and injuries in professional football	Reveals the relationship between fatigue accumulation and injury	Lacks computer vision-based action analysis	Supports the design of load-related risk variables
Rico-González et al. [5]	Review of machine learning applications in football	Covers tactical, performance, and risk analysis	Insufficient discussion of multimodal deep fusion	Clarifies the direction of intelligent football analysis
Robles-Palazón et al. [7]	Injury risk prediction in youth football	Combines individual characteristics and training exposure	Limited scenarios and sample scope	Provides a basis for selecting risk prediction indicators
Giancola et al. [11]	Football video action localization	Reduces the cost of action annotation	Focuses more on match events than individual injury	Supports the video action recognition module
Xarles et al. [14]	Transformer-based football action localization	Strengthens long-sequence relationship modeling	Does not integrate physiological and load data	Supports temporal action feature extraction

However, there are still three shortcomings in the existing research. First, action recognition and injury prediction are often treated separately, and it is difficult for the model to explain why a certain type of action pattern increases the risk of injury. Second, some studies focus on GPS, heart rate or training load indicators, and do not make full use of video pose and bone motion information. Third, multi-modal fusion mostly stays at the feature splicing level, lacking dynamic modeling of different time Windows and different modal contributions. Therefore, this paper intends to construct a multi-modal deep learning framework, which integrates football video posture, motion trajectory, training load, physiological state and injury records into a unified model, and realizes the collaborative analysis of action pattern recognition and sports injury prediction through spatio-temporal feature extraction, attention fusion and risk prediction modules.

3 Research Methods

3.1 Multi-modal data acquisition and preprocessing related to action and injury of soccer players

This paper aims to solve the problems of insufficient expression of single video features in soccer player action pattern recognition, unclear connection between training load indicators and injury risk, and difficulty in synchronous modeling of different modal data. The research objectives include: (1) To construct a multi-modal data set containing video pose, IMU inertial signal, GPS motion trajectory, training load and injury records, so that the model can simultaneously describe the action execution process and body load changes of soccer players; (2) Complete the unified labeling of action category, action quality and injury risk labels, which provides stable input for subsequent action pattern recognition and injury prediction. (3) Through filtering, normalization, missing completion and time synchronization processing, the influence of sensor noise, occlusion error and sampling frequency difference on model training is reduced.

In this paper, 126 male soccer players were selected as the research objects. The data collection period was 10 weeks, and each athlete completed 4-6 training or competition records per week. The collection content covers typical football movements such as sprinting, stopping, changing direction, shooting, passing, fighting for the top, physical confrontation and landing buffer. The video data was collected by a field-side multi-view camera, and the frame rate was set to 60 fps to extract the trajectory of key points such as shoulder, hip, knee and ankle. The IMU sensor is fixed on the lateral side of the waist and calf to record the acceleration, angular velocity and body rotation amplitude. GPS devices recorded instantaneous speed, running distance, acceleration times, and high-intensity running distance. Training load was composed of RPE, heart rate interval, training duration and acute-chronic load ratio. Injury information was recorded by the team doctor, including injury location, injury type, missing training days and return training status. All data were anonymously coded before collection, and names, numbers and personal identifying information were removed.

Table 2: Content of multimodal data collection for soccer players

Data Type	Collection Device or Source	Main Variables	Sampling Setting	Modeling Purpose
Video posture data	Multi-view HD cameras	Joint coordinates, limb angles, body center of gravity, action intervals	60 fps	Action pattern recognition and action quality analysis
IMU inertial data	Waist and lower-leg inertial sensors	Triaxial acceleration, angular velocity, turning amplitude, impact peak	100 Hz	Detection of change-of-direction, landing, and sudden-stop actions
GPS trajectory data	Athlete vest-type GPS device	Speed, running distance, number of accelerations and decelerations, high-intensity running distance	10 Hz	External training load calculation
Physiological load data	Heart-rate belt and training record sheet	Average heart rate, maximum heart rate, RPE, training duration	Training-session level	Representation of fatigue state and load intensity
Injury record data	Team doctor and rehabilitation files	Injury site, injury grade, days absent from training, return-to-training status	Periodic summary	Construction of injury risk labels

The core variables of the collected data are shown in Table 2. It can be seen from Table 2 that different modes correspond to action structure, exercise intensity, physiological load and injury results, respectively, which can describe the training status of soccer players from different aspects. IMU and GPS data can reflect external load fluctuations, heart rate and RPE can supplement internal fatigue state, and injury records are used to construct risk labels. The relationship between each mode is not simply parallel, but jointly serves the modeling chain of "action performance-load cumulation-damage risk".

Figure 1 shows the multimodal data acquisition and preprocessing process. The process takes the training scenario as the entry point, and converts video, sensor, load and injury records into structured features respectively, and then forms a unified input with sample slices through time alignment. Each module in Figure 1 does not directly give the model prediction results, but provides a clean, synchronous, and computable data basis for subsequent deep learning networks.

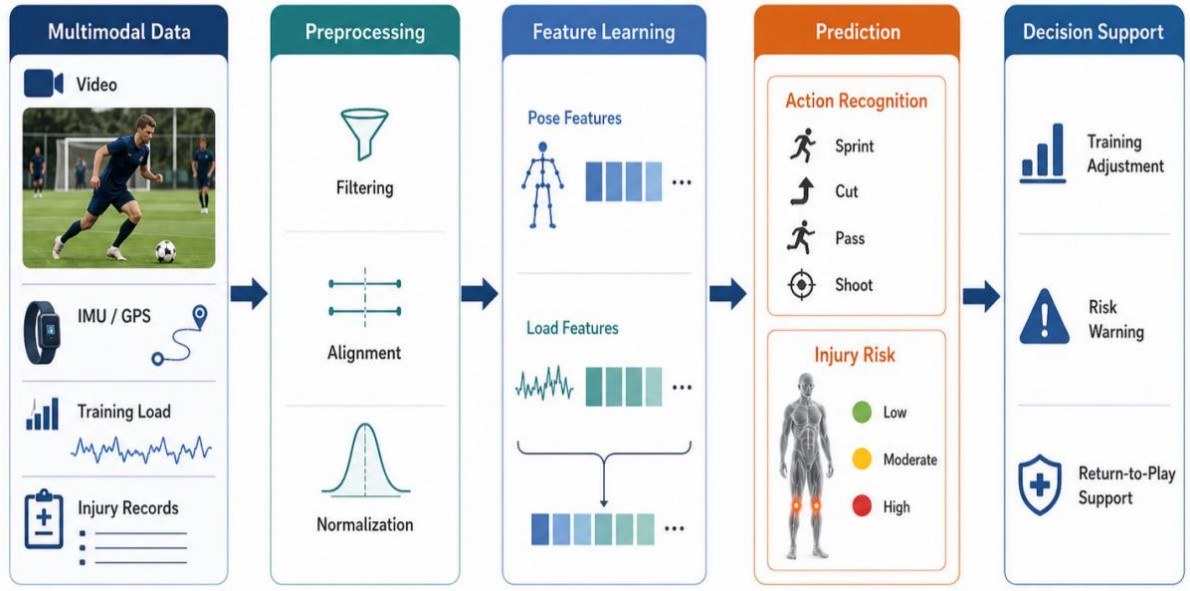

Figure 1: Flow of multimodal data acquisition and preprocessing related to action and injury of soccer players

The preprocessing stage mainly includes signal denoising, outlier processing, normalization and time synchronization. For IMU and GPS sequences, low-pass filtering is used to remove high-frequency jitter and retain the main trends in action changes. Let the original sensing sequence be x(t), the filter kernel be h(t), and the processed signal be y(t), which can be calculated as follows.

$$y(t) = \int_{-\infty}^{+\infty} x(\tau)h(t-\tau)d\tau \tag{1}$$

Equation (1) is used to weaken the influence of device jitter and transient abnormal peak on action recognition. For variables with different dimensions, such as velocity, heart rate, acceleration, and RPE, Z-score normalization is used:

$$\hat{x}_i = \frac{x_i - \mu}{\sigma + \varepsilon} \tag{2}$$

Here, x_i is the ith original observation, μ and σ are the mean and standard deviation of the variable in the training set, and ε is a stable term to prevent the denominator from being zero. After standardization, all kinds of features are mapped to similar numerical ranges to avoid excessive weighting of high-dimensional variables in model training.

Due to the different sampling frequencies of video, IMU, GPS, and training records, this paper uses 0.2 s as a uniform time interval for resampling, and uses linear interpolation to fill in short-time missing values. For video clips with occlusion exceeding 1.5 s or samples with continuous sensor packet loss exceeding 3 s, they are directly eliminated to avoid error samples interfering with model convergence. Action clips were sliced according to four stages: onset, development, peak, and recovery, each sample window length was set to 6 s, and 1 s context was retained before and after the action, so that the model could recognize the load change after action preparation and action end. After the above processing, this paper obtained 18 420 valid action clips and 1 260 weekly injury risk samples, which provided data support for subsequent action pattern recognition and sports injury prediction.

3.2 Action pattern recognition of soccer players and design of sports injury prediction model

This paper constructs a multimodal deep learning model for action pattern recognition and sports injury prediction of soccer players. The model inputs include video skeletal keypoint sequences, IMU inertial sequences, GPS exercise load sequences, physiological load sequences, and historical injury features. The video branch is used to capture the spatial attitude changes of the actions such as sprinting, stopping, changing direction, shooting, fighting for the top and landing buffer. The sensor branch is used to describe the acceleration, angular velocity, impact peak and body steering amplitude. The load branch is used to characterize fatigue-related factors such as high-intensity running distance, acceleration and deceleration times, heart rate interval, RPE and training duration. The model does not deal with action recognition and damage prediction separately, but takes action pattern recognition results as an important intermediate representation of risk prediction, so that damage judgment can be traced back to specific actions and load changes.

Let the video skeleton features of the ith athlete in time window t be $P_{i,t}$, the sensor features $S_{i,t}$, the training load features $L_{i,t}$, and the historical injury features M_i. For the sequence of skeletal keypoints, this paper uses spatio-temporal graph convolution to extract the structural relationship between joints, which is calculated as follows.

$$H^{(l+1)} = \sigma\left(\sum_{k=1}^{K} \tilde{A}_k H^{(l)} W_k^{(l)} + b^{(l)}\right) \tag{3}$$

where, $H^{(l)}$ represents the l-th layer bone feature, $\tilde{A}_k$ represents the normalized joint connectivity matrix, $W_k^{(l)}$ is the learnable weight, and σ is the nonlinear activation function. The proposed structure can extract motion information closely related to football injury, such as knee internal buckle, abnormal hip-knee-ankle linkage, and landing impact excursion.

For IMU, GPS, and physiological load sequences, in this paper, a bidirectional gated recurrent network is used to extract continuous temporal dependencies and an attention mechanism is used to assign weights to different modalities. The multi-modal fusion features are expressed as follows.

$$Z_{i,t} = \sum_{m=1}^{M} \alpha_m F_m(X_{i,t}^m) \tag{4}$$

Here, $F_m(\cdot)$ represents the encoder of the MTH modality, $X_{i,t}^m$ represents the input of this modality, and α_m represents the attention weight. In order to avoid one mode dominating the model due to its large numerical magnitude, the attention weights are obtained through a normalization function:

$$\alpha_m = \frac{\exp(e_m)}{\sum_{j=1}^{M} \exp(e_j)} \tag{5}$$

On the output side, the model sets up two task branches. The action recognition branch uses Softmax to output the probability of action categories, and the damage prediction branch uses Sigmoid to output the probability of damage risk in the next week:

$$\hat{y}_a = \text{Softmax}(W_a Z_{i,t} + b_a) \tag{6}$$

$$\hat{y}_r = \text{Sigmoid}(W_r [Z_{i,t}; M_i] + b_r) \tag{7}$$

where, $\hat{y}_a$ is the result of action pattern recognition, $\hat{y}_r$ is the result of injury risk prediction, and $[Z_{i,t}; M_i]$ represents the concatenation of fusion features and historical injury features. A multi-task loss function is used in the model training, so that the action recognition error and the damage prediction error jointly participate in the parameter update:

$$\mathcal{L} = \mathcal{L}_a + \lambda \mathcal{L}_r \tag{8}$$

where $\mathcal{L}_a$ is the action classification cross-entropy loss, $\mathcal{L}_r$ is the damage risk binary classification loss, and λ is the task balance coefficient. This design enables the model to learn the characteristics of soccer specific movements and enhance the sensitivity to injury risk factors. Figure 2 shows the overall structure of the model. Figure 2 shows the complete path from multi-source data input, feature encoding, attention fusion to dual-task output, which can clearly reflect the shared representation relationship between action recognition and damage prediction.

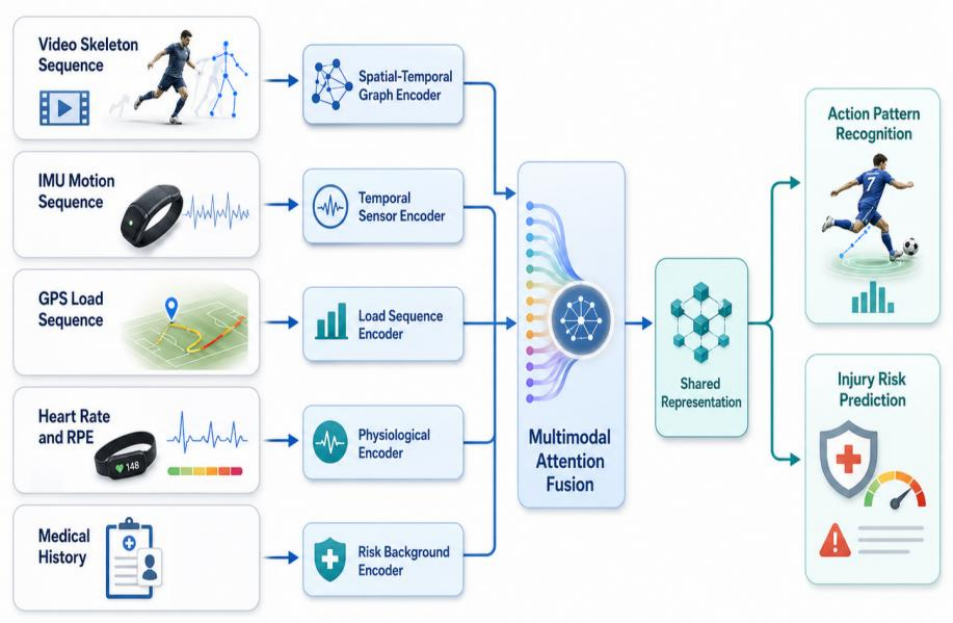

Figure 2: Structure of multimodal action recognition and damage prediction model

In the specific training process, the model uses the AdamW optimizer, the initial learning rate is set to 2×10^{-4}, the batch size is 32, and the maximum training rounds are 80. The early stopping strategy is started when the validation set loss does not decrease for 8 consecutive rounds. In order to reduce the influence of class imbalance on damage prediction, the class weight is introduced in the training stage of damage samples, and random time cropping and pose perturbation enhancement are used in the action recognition branch, so that the model can adapt to occlusion, speed change and action differences in different training scenarios.

4 Action pattern recognition and sports injury prediction of soccer players

4.1 Action Pattern Recognition of soccer Players based on multimodal Deep Learning

The goal of football player action pattern recognition is to accurately judge the type of action that the athlete is performing from continuous training or competition data, and further extract information such as action quality, action amplitude and load change. In this paper, video skeleton key points, IMU inertial signals and GPS motion trajectories are used as the main input of action recognition, where video data is used to describe the relationship between body posture and joint coordination, IMU data is used to capture the instantaneous changes in emergency stop, turn, landing and body collision, and GPS data is used to supplement external motion features such as speed, running distance, acceleration and acceleration and acceleration intensity. Through multi-modal information fusion, the model can avoid the occlusion of a single video or the lack of spatial structure information of a single sensor, so as to improve the stability of soccer specific action recognition. In the action segment input stage, this paper represents each training window as follows.

$$X_t = \{P_t, S_t, G_t\} \tag{9}$$

where, P_t represents the video skeleton keypoint sequence in the TTH time window, S_t represents the IMU inertial sensor sequence, and G_t represents the GPS motion trajectory and velocity load sequence. The model encodes the three types of data respectively, and then maps them into a unified feature space. The video skeleton branch focuses on extracting the spatio-temporal changes of joints such as hip, knee, ankle and shoulder. The IMU branch extracts the peak acceleration, angular velocity change and impact strength. The GPS branch extracts features such as high intensity running, sharp acceleration, sharp deceleration, and variable direction load. Figure 3 shows the multimodal action recognition process.

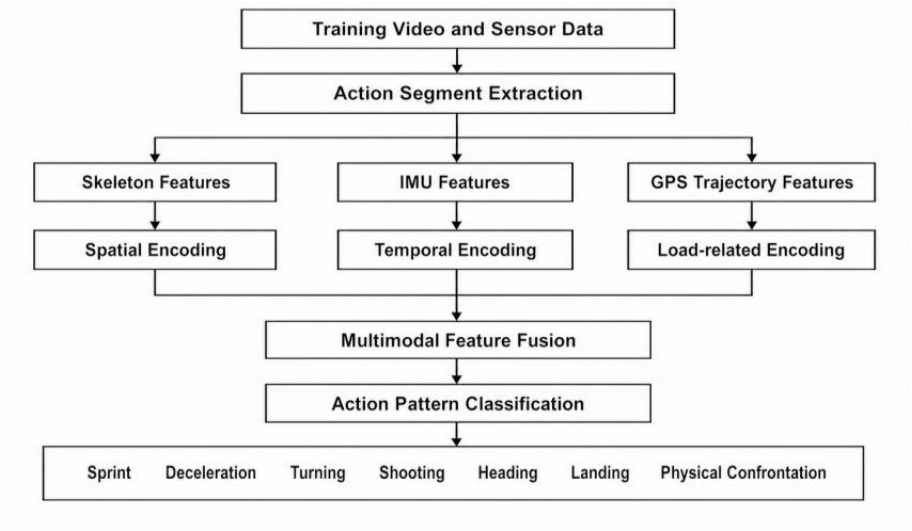

Figure 3: Process of multimodal action pattern recognition for soccer players

In the feature fusion stage, the attention weight is used to adjust the different modal contributions. Let the feature encoded by the m-th modality be Fm, and its corresponding weight be β_m. The fused action feature is expressed as follows.

$$F = \sum_{m=1}^{3} \beta_m F_m \tag{10}$$

Here, β_m is automatically learned by the model based on the current action segment. When the athlete is in a high-speed sprint state, the weight of GPS speed change and IMU acceleration features will increase relatively. When the athlete finishes shooting, fighting for the top or landing buffer action, the hip, knee and ankle Angle changes in the key points of the skeleton will become more important basis for recognition. This mechanism enables the model to adaptively select the source of information according to the action type, rather than mechanically concatenating all features. The action classification layer uses the Softmax function to output the probability distribution of various football actions, and the calculation formula is as follows.

$$p_c = \frac{\exp(W_c F + b_c)}{\sum_{j=1}^{C} \exp(W_j F + b_j)} \tag{11}$$

where, p_c represents the probability that the sample belongs to the action of class c, C represents the total number of action categories, and W_c and b_c are the classification layer parameters. The model predicts the class with the highest probability:

$$\hat{y} = \arg\max_c p_c \tag{12}$$

In order to improve the model's ability to distinguish similar actions, this paper uses the cross-entropy loss function in the training:

$$\mathcal{L}_{cls} = -\sum_{c=1}^{C} y_c \log(p_c) \tag{13}$$

where, y_c represents the true action label. When the real category is high injury related

actions such as changing direction, emergency stop, landing or body confrontation, the loss function will promote the model to pay more attention to the action boundary, joint Angle mutation and impact peak, so as to reduce the misjudgment between similar actions.

When evaluating the effect of action recognition, this paper uses accuracy, recall and Macro-F1 as the core indicators. Accuracy is used to measure the overall classification accuracy, and Macro-F1 is used to reflect the balanced recognition ability between each action category. It is calculated as follows.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

Here, TP, FP, and FN represent the correctly recognized samples, the samples incorrectly recognized as this type of action, and the samples of this type of action that were not recognized, respectively. For soccer training scenarios, it is not enough to rely on the overall accuracy alone to explain the effect of the model, because actions such as emergency stop, top fight, and landing buffer occur infrequently, but are closely related to the risk of injury. Therefore, this paper pays more attention to the recall ability and Macro-F1 performance of various actions, so that the action recognition results can provide reliable input for subsequent damage prediction.

4.2 Sports injury prediction based on fusion of training load and action features

The key of football injury prediction is to analyze the variation of training load and the quality of action execution in the same time series. The only use of load indicators such as running distance, heart rate or RPE can only reflect how much training stimulation the athlete has undergone, but it is difficult to explain the type of movement from which the injury risk comes. Using video pose or IMU signals alone, it is easy to ignore the hidden effects caused by fatigue accumulation and insufficient recovery. Therefore, based on the action pattern recognition results, this paper fuses the training load, action impact characteristics, physiological fatigue state and previous injury records to construct a prediction method for injury risk in the next week. In this paper, the risk input of an athlete in the t-th training window is defined as follows.

$$\text{R}_\text{t} = \{\text{A}_\text{t}, \text{Q}_\text{t}, \text{E}_\text{t}, \text{I}_\text{t}, \text{H}_\text{t}\} \tag{17}$$

where, A_t represents action pattern recognition results, Q_t represents action quality features, E_t represents external training load, I_t represents internal physiological load, and H_t represents previous injury and recovery status. The motion quality characteristics mainly include the range of knee buckle, the stability of landing buffer, the body anteversion Angle during emergency stop, and the degree of hip-knee-ankle coordination during changing direction. The external training load consisted of high intensity running distance, rapid acceleration times, rapid deceleration times and sprint times. Internal load consists of heart rate interval, RPE, training duration, and fatigue score. The input structure is able to simultaneously express "whether the action is dangerous" and "whether the body is in a state

of high load". Training load accumulation was described by acute-chronic load ratio, which was calculated as follows.

$$ACWR_t = \frac{\overline{Load}_{t-7:t}}{\overline{Load}_{t-28:t}} \tag{18}$$

Here, $\overline{Load}_{t-7:t}$ represents the average training load in the last 7 days, and $\overline{Load}_{t-28:t}$ represents the average training load in the last 28 days. When the ratio increases significantly, it indicates that the recent load growth of the athlete is too fast, and the risk of injury will be further increased if the quality of the sudden stop, direction change and landing movements are decreased at the same time. In order to quantify the contribution of high-risk actions to damage prediction, an action impact score is constructed as follows.

$$D_t = \sum_{k=1}^{K} \rho_k \, n_{k,t} \cdot s_{k,t} \tag{19}$$

where, $n_{k,t}$ represents the number of occurrences of the KTH high-risk action in window t, $s_{k,t}$ represents the impact intensity or posture abnormality degree corresponding to the action, and ρ_k represents the risk weight of different action types. For example, sharp deceleration and landing buffer movements were more likely to cause lower extremity joint impact, while body confrontation and top fight movements were associated with ankle knee twist and low back impact. Through this score, the model can transform the action recognition results into risk variables that can be used for injury prediction.

The damage prediction process based on the fusion of training load and action features is shown in Figure 4. Figure 4 shows the calculation path from action recognition results to injury risk output, where action risk, load change, physiological fatigue and injury background jointly enter the risk prediction module, rather than a single indicator directly determining the prediction result.

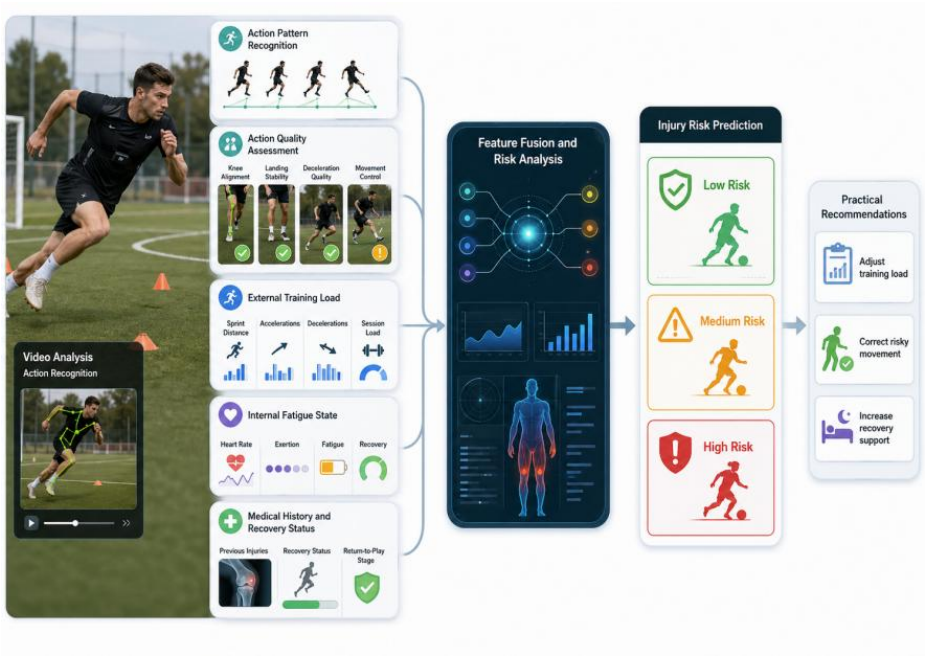

Figure 4: Sports injury prediction process of training load and action feature fusion

In the model prediction stage, the action impact score, load accumulation index, physiological fatigue characteristics and injury background characteristics are concatenated into a unified risk vector:

$$V_t = [D_t, ACWR_t, I_t, H_t, Q_t] \tag{20}$$

The risk prediction network outputs the probability of a training-related injury in the next week:

$$P_t = \frac{1}{1 + \exp[-(W_r V_t + b_r)]} \tag{21}$$

Here, P_t represents the probability of damage risk, and W_r and b_r are the model parameters. P_t<0.35 was considered low risk, 0.35≤P_t<0.65 was considered medium risk, and P_t≥0.65 was considered high risk. The threshold setting is not a direct replacement for team doctor judgment, but provides computer-aided basis for training arrangement, load adjustment and technical movement correction. Due to the low proportion of damage samples in the real training data, the weighted binary classification loss function is used for model training:

$$\mathcal{L}_{risk} = -\omega_1 y_t \log(P_t) - \omega_0 (1 - y_t) \log(1 - P_t) \tag{22}$$

where y_t represents the true damage label, ω_1 and ω_0 represent the class weights of damaged and undamaged samples, respectively. This design can reduce the bias caused by sample imbalance, so that the model can maintain high sensitivity in identifying low-frequency but high-risk injury events. The prediction results can be used in combination with the training plan: when the model outputs medium-high risk continuously, the system can prompt the coach to reduce the proportion of high-intensity change direction and sprint training, and increase the recovery training and movement technique modification, so as to convert the action pattern recognition results into the basis for injury prevention decision-making.

5 Experimental Results

5.1 Ablation experiments

In order to test the actual contribution of each component module of the multimodal deep learning model in this paper to the action pattern recognition and sports injury prediction of soccer players, this paper carries out ablation experiments under the same data set, the same training rounds and the same evaluation index. The experimental data is composed of video skeleton key points, IMU-GPS motion sequences, training load records and damage labels. The training set, validation set and test set are divided according to 7:1.5:1.5. The complete model includes a video skeleton branch, an IMU-GPS motion branch, a training load branch, an attention fusion module, and a multi-task joint learning structure. The control model removed one of the core modules to observe the change of action recognition results and damage prediction results. The experimental results are shown in Table 3.

Table 3: Results of ablation experiments

Model Configuration	Action Recognition Accuracy (%)	Macro-F1 (%)	Injury Prediction AUC	High-Risk Injury Recall (%)
Complete multimodal model	94.6 ± 0.8	93.8 ± 0.9	0.921 ± 0.012	88.7 ± 1.5
Without video skeleton branch	89.2 ± 1.1	87.5 ± 1.3	0.874 ± 0.018	81.6 ± 2.0
Without IMU-GPS motion branch	90.1 ± 1.0	88.4 ± 1.2	0.862 ± 0.020	80.9 ± 2.1
Without training load branch	93.1 ± 0.9	91.9 ± 1.0	0.846 ± 0.022	78.3 ± 2.4
Without attention fusion module	91.4 ± 1.2	89.6 ± 1.4	0.879 ± 0.017	82.5 ± 1.9
Without multi-task joint learning	92.0 ± 1.0	90.7 ± 1.1	0.861 ± 0.019	79.8 ± 2.2

As can be seen from Table 3, the full multimodal model achieves the best results on the four indicators, the accuracy of action recognition reaches 94.6%, Macro-F1 reaches 93.8%, the AUC of damage prediction reaches 0.921, and the recall rate of high-risk damage reaches 88.7%. After removing the skeleton branch of the video, the accuracy of action recognition decreases to 89.2%, and Macro-F1 decreases to 87.5%, which indicates that the skeleton key point sequence has a strong distinguishing effect on the actions such as emergency stop, change direction, fight for the top and landing buffer. After removing the motion branch of IMU-GPS, the recall of high risk impairments decreases to 80.9%, indicating that inertial signals and trajectory loads are able to supplement action intensity information under video occlusion, body overlap, and high speed motion. After removing the training load branch, the accuracy of action recognition still remains at 93.1%, but the AUC of injury prediction decreases to 0.846, and the recall rate of high-risk injury decreases to 78.3%, indicating that the contribution of training load to injury risk judgment is greater than its contribution to action category recognition. After removing the attention fusion module, the model cannot dynamically allocate the modal weights according to the action scene, and all indicators are significantly decreased. After removing multi-task joint learning, the shared representation between action recognition and injury prediction is weakened, and the ability of the model to characterize the transmission of high-risk actions to injury risk is reduced.

In order to further observe the recognition stability of the full model on different soccer specific actions, this paper counted the recall rate of the full model for various types of actions, and the results are shown in Figure 5. Figure 5 shows that the model has high recall rates of 96.2%, 95.3% and 95.8% for sprint, shot and pass actions, respectively. The recall rates for actions such as emergency stop, change direction, fight for top and landing buffer are relatively low, but all remain above 90%. This result indicates that high-risk soccer actions are usually accompanied by body occlusion, short action duration, and abrupt joint Angle changes, and the recognition difficulty is higher than that of regular technical actions, but the multimodal input can still maintain good classification stability.

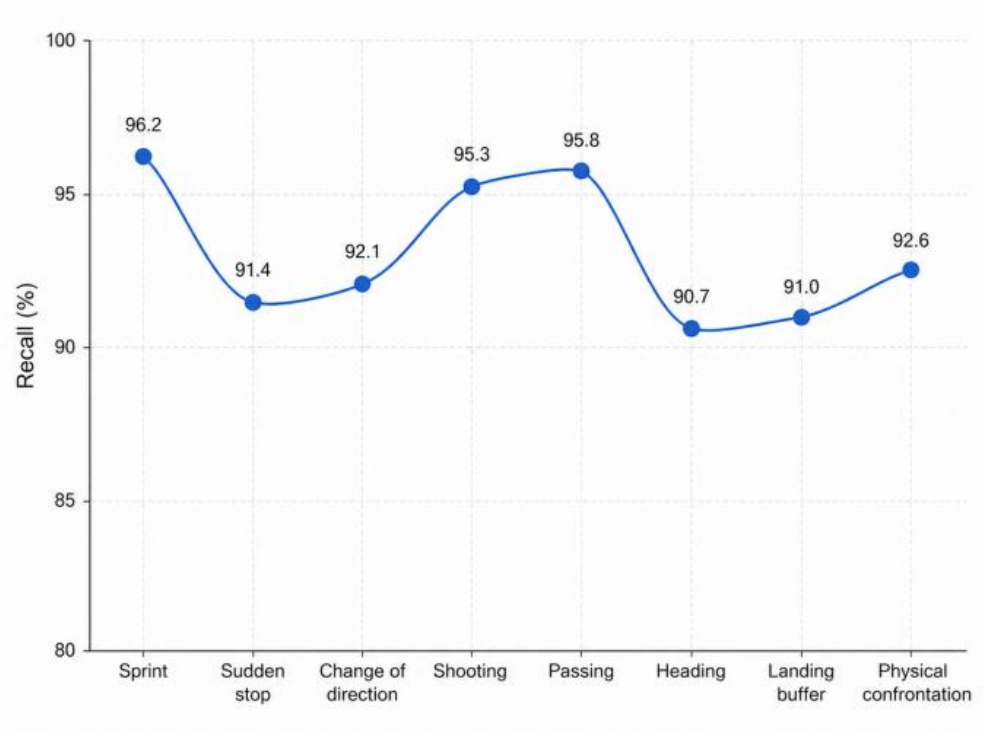

Figure 5: Recall of different action categories for the full model

It can be seen from Table 3 and Figure 5 that the advantages of the proposed model are not only reflected in the overall index improvement, but also in the stable recognition of football-specific action details. The video skeleton branch mainly improves the action structure discrimination ability, the IMU-GPS motion branch enhances the capture ability of high-speed actions and impact actions, the training load branch improves the injury risk prediction effect, and the attention fusion module enables different modalities to complement each other according to the action scene. Ablation experiments show that the multimodal deep learning model proposed in this paper is suitable for soccer player action pattern recognition and sports injury prediction, and can provide reliable data basis for subsequent training load adjustment and injury warning.

5.2 Performance analysis of action pattern recognition of soccer players

In order to evaluate the action pattern recognition ability of the proposed model in different football training and competition scenarios, this paper divides the test samples into six categories according to the scene complexity: fixed technology training, regular confrontation training, small field competition, full field competition, high-intensity physical confrontation and strong occlusion complex scenes. Each scene contains action categories such as sprinting, stopping, changing direction, shooting, passing, fighting for the top, landing buffer and physical confrontation, and the action recognition accuracy and Macro-F1 are used as evaluation indicators. The recognition results are shown in Figure 6.

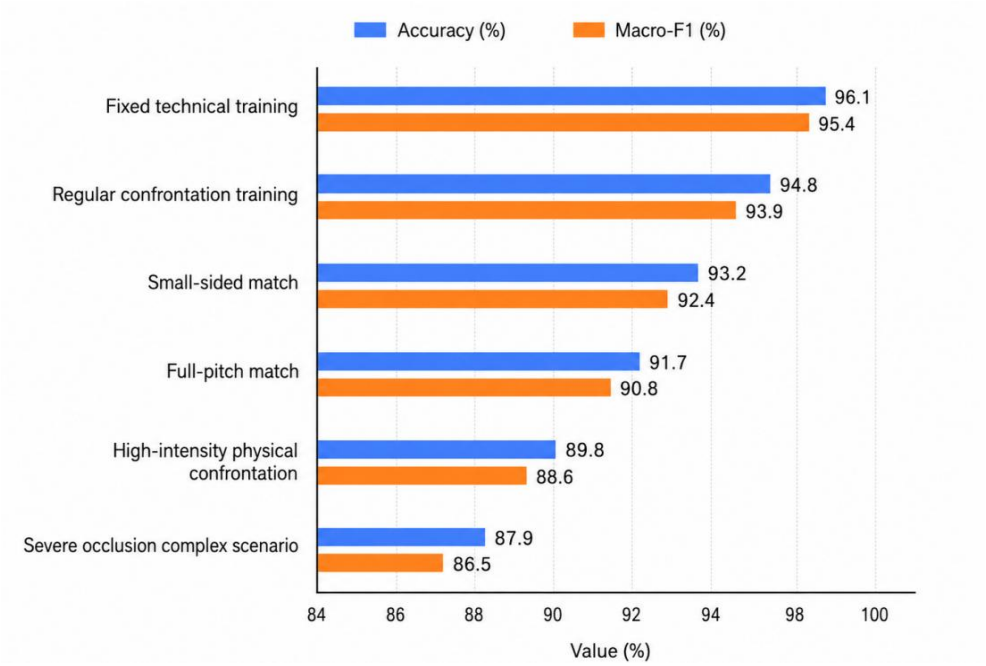

Figure 6: Performance of action pattern recognition in different soccer scenarios

It can be seen from Figure 6 that the model has the best recognition effect in the fixed technology training scenario, with the accuracy of action recognition reaching 96.1% and Macro-F1 reaching 95.4%. In this scene, the movement trajectory of the athlete is relatively clear, the body is less occluded, and the key points of the video skeleton and the IMU signal can form a stable correspondence, so the model has high discrimination ability for passing, shooting and sprinting. In conventional adversarial training and small games, the accuracy rates are 94.8% and 93.2%, respectively, and Macro-F1 is 93.9% and 92.4%, respectively, indicating that the model can still maintain good recognition stability when the movement speed is improved, the frequency of turning is increased, and the defensive interference is enhanced.

When the scene enters the stage of full-court competition and high-intensity physical confrontation, the performance of action recognition decreases. The accuracy of the whole game is 91.7%, and Macro-F1 is 90.8%. The accuracy of high-intensity physical confrontation scene is further reduced to 89.8%. This change is obviously related to the strong action continuity, frequent body contact and occlusion between players in football matches. Especially in actions such as top fight, landing buffer and adversarially turning, single frame attitude information is easily affected by occlusion, and the model needs to rely more on IMU angular velocity, impact peak value and GPS velocity change to complete the judgment.

In the strong occlusion complex scene, the model accuracy is 87.9%, and Macro-F1 is 86.5%, which is the lowest set of results in all scenes. This result shows that complex illumination, multi-person overlap, and high-speed direction change will weaken the quality of video features, but the model still maintains Macro-F1 above 85%, indicating that multi-modal fusion can provide compensation when visual information is incomplete. In general, the model in this paper is more stable for normative technical action recognition, and difficult for high confrontation, high occlusion and short burst action recognition, but the overall performance can meet the needs of football training monitoring and injury prediction pre-analysis.

5.3 Sports injury prediction results of football players

In order to evaluate the stability and risk discrimination ability of the proposed model in the sports injury prediction task of football players, this paper uses 10-fold cross validation to evaluate the weekly risk samples, and redivides the training subset and the validation subset at each compromise. The input structure of training load, action characteristics, physiological fatigue characteristics, and historical injury variables were kept consistent for each fold, and the evaluation metrics included AUC, sensitivity, specificity, and high-risk recall. AUC is used to measure the ability of the model to distinguish injury samples from non-injury samples, sensitivity reflects the detection level of the model for actual injury athletes, specificity is used to measure the ability of the model to exclude low-risk samples, and high-risk recall is used to evaluate the identification effect of the model for key warning objects. The damage prediction results under different cross-validation folds are shown in Figure 7.

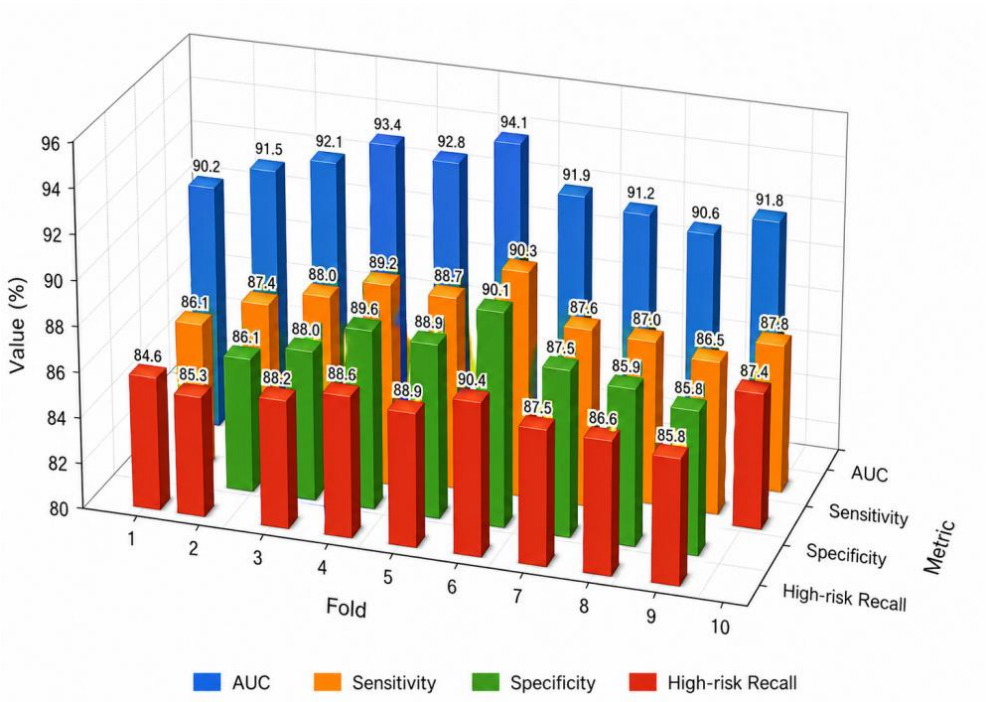

Figure 7: Sports injury prediction results under 10-fold cross validation

As can be seen from Figure 7, the AUC remained between 90.2% and 94.1% after displaying according to the percentage system, corresponding to the original AUC of 0.902-0.941, indicating that the model has a good ability to distinguish damage risk under different data subsets. Among them, the AUC of the sixth fold reaches 0.941, the sensitivity is 89.1%, the specificity is 90.3%, and the high-risk recall is 90.4%, which is the most stable performance. The AUC of the first fold and the ninth fold are relatively low, 0.902 and 0.906, respectively, but still maintain above 0.90, indicating that the model is not significantly invalid due to the change of sample division. The sensitivity fluctuated between 84.6% and 89.1%, and the specificity varied between 86.1% and 90.3%. Both indicators were at a high level, indicating that the model can not only identify potential injury athletes, but also reduce the false warning of low risk athletes.

From the perspective of high risk recall rate, the results of the model in each fold are more than 85%, and the fourth, fifth and sixth fold reach 89.6%, 88.9% and 90.4% respectively. This result shows that after the fusion of training load and action features, the model can

better capture the association between high-risk actions such as rushing to stop, changing direction, landing buffer, and physical confrontation and fatigue accumulation. For football training scenarios, the high risk recall rate has direct application value, because the missed identification of high risk athletes may lead to the continued increase of training load, thereby increasing the probability of muscle strain, knee and ankle injuries, and excessive fatigue injuries.

Overall, the 10-fold cross-validation results show that the proposed model can maintain high prediction accuracy and stability under different data partitions. The action impact score, ACWR load index, physiological fatigue state and previous injury information jointly improve the reliability of risk prediction. The results show that the multimodal deep learning method can provide more robust computational support for football players' injury early warning, and can be used as an important basis for coaches to adjust training intensity, arrange recovery training and correct technical movements.

5.4 Analysis of comprehensive application effect of multimodal fusion model

In this paper, the comprehensive application effect is defined as the auxiliary value of the model in real football training management, which mainly includes the decrease of injury incidence, the increase of risk warning advance, the increase of motion correction completion rate and the improvement of training load adjustment accuracy. The injury event was confirmed by the team doctor based on clinical examination and training interruption records. The statistical standard was that the athlete's training interruption caused by muscle strain, joint sprain, excessive fatigue or acute confrontation injury was not less than 48 hours. In order to evaluate the practical effect of the multimodal fusion model, 60 athletes with complete training records and team doctor follow-up records were selected from 126 samples to carry out application verification, and the traditional empirical monitoring scheme was compared with the model-assisted scheme in this paper. Both schemes cover the same training period, the same training venue and the same coaching staff arrangement. The comprehensive application results are shown in Table 4.

Table 4: Comparison of comprehensive application effects of multimodal fusion models

Evaluation Indicator	Traditional Experience-Based Monitoring Scheme	Proposed Model-Assisted Scheme	Difference	95% Confidence Interval	t	p	Cohen’s d
Injury Incidence Rate (cases/1000 h)	4.82 ± 0.76	2.91 ± 0.58	-1.91	[-2.24, -1.53]	9.37	<0.001	2.41
Lead Time of High-Risk Warning (d)	1.6 ± 0.5	3.8 ± 0.7	2.2	[1.84, 2.53]	11.26	<0.001	2.89
Action Correction Completion Rate (%)	71.4 ± 5.8	86.9 ± 4.6	15.5	[12.7, 18.2]	8.74	<0.001	2.26
Training Load Adjustment Accuracy (%)	74.8 ± 6.1	89.3 ± 4.2	14.5	[11.9, 17.4]	8.19	<0.001	2.11

It can be seen from Table 4 that the model-assisted scheme in this paper is significantly better than the traditional empirical monitoring scheme in terms of comprehensive application effect. The incidence of injury decreased from 4.82 times /1000 h to 2.91 times /1000 h, and the difference was -1.91 times /1000 h, and the difference reached a significant level ($p<0.001$). This result shows that the multimodal fusion model can detect the abnormal performance of athletes in emergency stop, change direction, landing buffer and physical confrontation in advance through the joint analysis of action recognition, load monitoring and risk prediction, and reduce the risk of injury caused by continued training under high load conditions.

From the perspective of risk warning effect, the proposed model increases the amount of high-risk warning advance from 1.6 days to 3.8 days, indicating that the system can capture the signal of load accumulation and action quality degradation earlier before damage occurs. For football training, identifying risks more than 2 days in advance has strong practical significance. Coaches can adjust the proportion of sprint training, reduce continuous high-intensity confrontation, and arrange recovery training or special technical modification. The completion rate of motion correction increased from 71.4% to 86.9%, indicating that the motion risk information output by the model could be better translated into training intervention content, especially suitable for the correction of knee joint buckle, insufficient landing buffer and unstable variable support.

The accuracy of training load adjustment is improved from 74.8% to 89.3%, indicating that the model in this paper can not only identify the action pattern, but also combine GPS, IMU, heart rate, RPE and previous injury information to determine whether the load is suitable. Cohen's d is greater than 2.0, indicating that the differences have a strong practical effect. In summary, the multimodal fusion model can connect the action recognition results, training load changes and injury risk warning, and provide stable data support for football player training monitoring, load distribution and injury prevention.

6 Discussion

The proposed model shows good stability in the action pattern recognition and sports injury prediction of football players. The experimental results show that the accuracy of action recognition of the full multimodal model reaches 94.6%, the AUC of injury prediction reaches 0.921, and the recall rate of high-risk injury reaches 88.7%, which indicates that video skeleton key points, IMU-GPS motion sequences, training load and injury records can effectively complement each other. Compared with single video recognition or single load prediction, multi-modal fusion not only improves the accuracy of action category judgment, but also enhances the model's ability to explain high-risk actions such as emergency stop, change direction, landing buffer and physical confrontation.

From the results of action recognition, the model performs well in fixed technology training and conventional adversarial training, but the accuracy drops to 87.9% in strong occlusion complex scenes. This phenomenon is related to the overlapping of multiple people, the short duration of actions, and the frequent physical contact in soccer scenes. When the video pose information is affected by occlusion, IMU angular velocity, impact peak and GPS velocity change can provide compensation, but it is still difficult to completely eliminate the error caused by complex scenes. The damage prediction results show that the AUC in 10-fold cross validation remains between 0.902 and 0.941, indicating that the model has good generalization stability. After the training load branch was removed, the AUC of injury prediction decreased significantly, indicating that load accumulation, fatigue state, and previous injuries were still key factors in risk judgment.

The strength of this study is to connect the action recognition results with the damage prediction process, so that the model can provide auxiliary judgment from the chain of "abnormal action -load accumulation -elevated risk". However, there are still some limitations in this study. The samples mainly come from male soccer players in the same training cycle, and the applicability of the model in different age groups, women's soccer and teams with different competitive levels still needs to be further verified. Follow-up studies can expand sample sources, introduce more actual game data, and combine team doctor intervention records and rehabilitation process data to improve the interpretability and application stability of the model in real training management.

7 Conclusions

Focusing on the problem of action pattern recognition and sports injury prediction of soccer players, this paper constructs an analysis model based on multimodal deep learning. In this study, video skeleton key points, IMU-GPS motion sequences, training load, physiological fatigue state and previous injury records are incorporated into a unified modeling framework. Through spatio-temporal feature extraction, attention fusion and multi-task learning, the collaborative association between action recognition results and injury risk prediction is realized. Experimental results show that the accuracy of action recognition of the complete model reaches 94.6%, the AUC of injury prediction reaches 0.921, and the recall rate of high-risk injury reaches 88.7%, indicating that multimodal fusion can effectively improve the reliability of football specific action recognition and injury warning.

The research value of this paper is mainly reflected in three levels: data organization, model collaboration and training application. Firstly, a multi-source data processing flow for soccer training scenarios is constructed, which enhances the joint expression ability of action structure, exercise intensity and load changes. Secondly, a deep learning model shared between action pattern recognition and damage prediction is designed, so that high-risk action features can participate in subsequent risk judgment. Thirdly, the ablation experiment and comprehensive application analysis verify the application value of the model in training monitoring, motion correction and injury prevention. There are still problems of limited sample sources and scenario coverage in this research. In the future, football samples of different ages, genders and competitive levels can be extended, and longer periods of game data and rehabilitation records can be introduced to improve the generalization ability and practical applicability of the model.

References

[1] Majumdar A, Bakirov R, Hodges D, et al. Machine learning for understanding and predicting injuries in football[J]. Sports Medicine-Open, 2022, 8(1): 73.

[2] Bullock G S, Mylott J, Hughes T, et al. Just how confident can we be in predicting sports injuries? A systematic review of the methodological conduct and performance of existing musculoskeletal injury prediction models in sport[J]. Sports medicine, 2022, 52(10): 2469-2482.

[3] Page R M, Field A, Langley B, et al. The effects of fixture congestion on injury in professional male soccer: A systematic review[J]. Sports Medicine, 2023, 53(3): 667-685.

[4] Silva H, Nakamura F Y, Beato M, et al. Acceleration and deceleration demands during training sessions in football: a systematic review[J]. Science and Medicine in Football, 2023, 7(3): 198-213.

[5] Rico-González M, Pino-Ortega J, Méndez A, et al. Machine learning application in soccer: a systematic review[J]. Biology of sport, 2023, 40(1): 249-263.

[6] Nassis G, Verhagen E, Brito J, et al. A review of machine learning applications in soccer with an emphasis on injury risk[J]. Biology of sport, 2023, 40(1): 233-239.

[7] Robles-Palazón F J, Puerta-Callejón J M, Gámez J A, et al. Predicting injury risk using machine learning in male youth soccer players[J]. Chaos, Solitons & Fractals, 2023, 167: 113079.

[8] Haller N, Kranzinger S, Kranzinger C, et al. Predicting injury and illness with machine learning in elite youth soccer: a comprehensive monitoring approach over 3 months[J]. Journal of Sports Science & Medicine, 2023, 22(3): 476.

[9] Pillitteri G, Petrigna L, Ficarra S, et al. Relationship between external and internal load indicators and injury using machine learning in professional soccer: a systematic review and meta-analysis[J]. Research in sports medicine, 2024, 32(6): 902-938.

[10] Tsilimigkras T, Kakkos I, Matsopoulos G K, et al. Enhancing sports injury risk assessment in soccer through machine learning and training load analysis[J]. Journal of sports science & medicine, 2024, 23(3): 537.

[11] Giancola S, Cioppa A, Georgieva J, et al. Towards active learning for action spotting in association football videos[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 5098-5108.

[12] Mkhallati H, Cioppa A, Giancola S, et al. SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 5074-5085.

[13] Held J, Cioppa A, Giancola S, et al. VARS: Video assistant referee system for automated soccer decision making from multiple views[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 5086-5097.

[14] Xarles A, Escalera S, Moeslund T B, et al. Astra: An action spotting transformer for soccer videos[C]//Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports. 2023: 93-102.

[15] Denize J, Liashuha M, Rabarisoa J, et al. COMEDIAN: Self-supervised learning and knowledge distillation for action spotting using transformers[C]//Proceedings of the IEEE/CVF Winter Conference on applications of computer vision. 2024: 530-540.

[16] Psaltis A, Patrikakis C Z, Daras P. Deep multi-modal representation schemes for federated 3d human action recognition[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 334-352.

[17] Pajak I, Krutz P, Patalas-Maliszewska J, et al. Sports activity recognition with UWB and

inertial sensors using deep learning approach[C]//2022 IEEE international conference on fuzzy systems (FUZZ-IEEE). IEEE, 2022: 1-8.

[18] Mekruksavanich S, Phaphan W, Hnoohom N, et al. Recognition of sports and daily activities through deep learning and convolutional block attention[J]. PeerJ Computer Science, 2024, 10: e2100.

[19] Müller P N, Müller A J, Achenbach P, et al. Imu-based fitness activity recognition using cnns for time series classification[J]. Sensors, 2024, 24(3): 742.

[20] Mekruksavanich S, Jitpattanakul A. A residual deep learning method for accurate and efficient recognition of gym exercise activities using electromyography and IMU sensors[J]. Applied System Innovation, 2024, 7(4): 59.

[21] Ayala R E D, Granados D P, Gutiérrez C A G, et al. Novel study for the early identification of injury risks in athletes using machine learning techniques[J]. Applied Sciences, 2024, 14(2): 570.

[22] Teixeira J E, Encarnação S, Branquinho L, et al. Data mining paths for standard weekly training load in sub-elite young football players: a machine learning approach[J]. Journal of Functional Morphology and Kinesiology, 2024, 9(3): 114.

[23] Davis J, Bransen L, Devos L, et al. Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned[J]. Machine Learning, 2024, 113(9): 6977-7010.