CDE-Driven Time-Series Analysis of Training Load and Injury Risk Prediction for Football Players

Meng Wang
[1]Sports Department, Zhengzhou Vocational College of Finance and Taxation, Zhengzhou 450000, Henan, China
Corresponding author: Meng Wang (e-mail: wangmeng09090222@126.com)

ABSTRACT Effective monitoring of training load and timely prediction of injury risk are essential for optimizing athletic performance and reducing sports-related injuries. This study proposes a Context–Data–Event (CDE)-driven framework for time-series analysis of football training loads and injury risk prediction. Multi-source physiological and motion data are collected through wearable sensing devices, including acceleration, heart rate, speed, displacement, and recovery indicators. Context-aware feature extraction is combined with long short-term memory networks to model temporal load variations, while a Bayesian risk assessment module dynamically estimates injury probability by analyzing interactions between physiological responses and training intensity fluctuations. Experimental evaluation using longitudinal training data from professional football players demonstrates high prediction accuracy and stable performance across varying load conditions. The proposed framework provides an effective solution for intelligent athlete monitoring and offers methodological references for wearable sensing technologies, physiological signal analysis, and wireless health-monitoring systems.

INDEX TERMS injury risk prediction, training load analysis, wearable sensing, physiological signal analysis, wireless health monitoring

I. INTRODUCTION

SCIENTIFIC monitoring and evaluation of training load, facilitated by high-performance smart sensors integrated into athletic gear, has become an essential component of competitive performance and injury prevention in football training. These advanced functional materials allow for the seamless collection of biomechanical and physiological data, ensuring a more precise and objective assessment of an athlete's physical status. Conventional ways of load assessment usually use individual indicators of physiological or athletic state, and it is hard to believe that it is really an accurate picture of the training condition of an athlete and his or her recovery. As the competitive matches become more intense, the issues that the training load monitoring systems have to deal with include lack of sufficient dimensions, poor timeliness, and limited predictive accuracy, which causes delays in detecting the risk of injury and prevents dynamical changes and personalized control of training plans.

The study is highly important in enhancing the scientific aspect of training football players and minimizing cases of injuries. With the creation of a model of systematic load monitoring and predicting the risk, a quantitative analysis and prediction of the trends of training data are obtained, which will become the basis of sports medicine intervention. Meanwhile, this study can be practically guiding in terms of optimizing the basis of the decision-making process of the coaching team, training intensity/recovery rhythm, and sustainable athlete development.

Data ingår)-driven time-series time modeling approach oneytowards the idea of multi-source information integration. It is based on this approach in combination with an LSTM (Long Short yer Term Memory) network to reflect the dynamic nature of training load and a Bayesian risk assessment mechanism to provide a probabilistic outcome of the risk of injury. This study's innovativeness is rooted in the combination of contextual information and behavioral event features captured by high-performance smart sensors, effectively bypassing the limitations of traditional models through an individualized approach to training load monitoring. By integrating these advanced sensing technologies with a rigorous theoretical framework, the research provides sensor-based intelligent technical support for more precise and scientific football training management.

II. RELATED WORK

The classical methods of training load monitoring are mainly based on the statistical analysis and the linear models including the average training volume, internal/external load ratio, and rapid/slow load ratio used to consider the load.

Although these are straightforward to compute, computationally intuitive, they miss out on the differences in physiological and training responses in an individual, and therefore, it is hard to adequately describe complex nonlinear interactions. Linear models can not capture the true dynamic training state when there is a delay or coupling effect between training stimuli and physiological feedback, which can cause a large difference between the load prediction and fatigue measurement.

As the data acquisition technology and smart wearable devices have become popular, machine learning algorithms have been slowly implemented in the sports load analysis. The support vector machine, random forests and neural network models have been utilized by researchers to derive features of multidimensional training data to gain a greater level of accuracy in the classification of load and injury risk. Nevertheless, these approaches mostly concentrate on outcome forecasting but do not consider time dependency structures and are inadequate in discovering risk-accumulation characteristics throughout the continuous training [1,2]. Moreover, sample size and insufficient interpretability of features are also a weakness of the existing models because they act as bottlenecks in large-scale model application.

In response to the need for dynamic prediction of injury risk, the CDE (Context-Data-Event) framework and time series analysis methods have gradually attracted attention. CDE can realize the correlation of multisource data by integrating environment, individual status and training events. Combined with time series networks such as LSTM and GRU, it can effectively identify the trend and potential risks of training load changes. Compared with traditional static analysis, this type of model can reflect the fluctuation of athletes' status in real time, improve the timeliness and accuracy of injury warning, and is an important direction for promoting the intelligent development of sports science [3,4].

III. METHODS

A. DATA ACQUISITION AND PREPROCESSING

This paper developed a multi-source data collection system on football training situations. The GPS motion units and triaxial accelerometers were utilized to record the changes in player speed, displacement tracks, and body load variations during the training and high-frequency sampling were applied to record the motion acceleration properties. Data in the form of a proprioceptive state were collected using a synchronous heart rate monitor that recorded the distribution of the heart rate interval and levels of variability, a sleep recovery tracker that recorded the rate of nighttime recovery and fatigue levels. In addition, to comprehensively assess the baseline risk of athletes, we systematically collected each athlete's age, detailed history of past injuries (including location, severity and recovery status) before the start of the experiment, and recorded their mental fatigue status using a standardized mental fatigue scale before each day's training. Timestamps were recorded at milliseconds and all the devices were synchronized to facilitate the wireless transfer of data to edge nodes where they could be fused [5,6]. To minimize processing latency while ensuring data quality, a layered preprocessing strategy is employed. First, the raw signal undergoes low-latency sliding window mean filtering with a window size of 0.1 seconds (i.e., 10 sampling points) to eliminate high-frequency transient noise while maintaining sensitivity to fast-moving events. Subsequently, outlier detection based on dynamic thresholds is used to remove signal drift and outlier peaks. To address the issue of inconsistent sampling frequencies among multiple sensors (10 Hz GPS/accelerometer and 1 Hz heart rate monitor), we employ a Gaussian process regression method based on cubic spline interpolation for time series alignment and reconstruction. This method first establishes an independent time series model for each data source, and then performs interpolation on a unified high-frequency time grid (10 Hz). Gaussian process regression quantifies the uncertainty in the interpolation process through its covariance function (Matern 5/2 kernel), thus providing a smoother and physiologically more physiologically plausible continuous estimate when reconstructing low-frequency heart rate signals. The data of the calibrated motion trajectory was corrected to coordinate differences to account spatial offset and the data of heart rate was integrated with exercise intensity to obtain reliable segments to guarantee continuity and comparability of the training load characteristics. Using the above method, we effectively fused data from different frequencies into a multivariate time series that is consistent in time dimension and has clear physical meaning, providing a reliable data foundation for the subsequent construction of the CDE framework and LSTM time feature mapping, and minimizing the alignment error that may be introduced by simple interpolation.

B. CDE FRAMEWORK CONSTRUCTION

The CDE is designed with the context description of a training process, modeling of data flow, and an event triggering logic in order to obtain the dynamic association of multi-source information. At the Context level, the training environment features, such as weather, location type, and type of training task are measured with the help of the site boundary determination and the choice of the stable positioning signal fragments. The system provides an analysis of the deployment information of the coach in real-time, plots the training form, load target, and location distribution to be mapped to a contextual scene variable and only constrain the extent of data to be interpreted. The Data layer comprises multimodal data streams of exercise, heart rate and recovery measurements and employs timestamp indexing and semantic labeling fusion to convert individual loading intensity, movement and pattern and physiological responses to continuous temporal objects. At the same time, this layer also integrates structured data from athlete profiles (age, history of injuries) and daily questionnaires (mental fatigue status), using them as constant or time-varying background features. The multidimensionalally normalized data is then given to the feature buffer layer and the internal consistency and temporal integrity of the data is upheld by a sliding queue mechanism [7,8]. The Event layer is responsible for identifying key events in the training process, such as sprints, sudden stops, adversarial situations, rapid acceleration, and sudden changes in fatigue response. The system uses sequence density clustering and probability threshold judgment to detect sudden events and generates event labels based on contextual conditions to achieve semantic expression of the training process. Through the ternary mapping relationship of CDE, load changes and physiological responses can be precisely aligned on the time axis, thereby establishing a traceable temporal correlation logic and providing a structured data foundation for understanding the dynamic relationship between training load and physiological responses.

C. TEMPORAL FEATURE MAPPING MODEL

The temporal feature mapping model takes the dynamic changes of training load as its construction target and transforms the multi-source temporal signals output by the CDE framework into a learnable sequence feature structure. The model fragments the motion and physiological data through a sliding time window, and each fragment retains the continuously changing motion acceleration, heart rate curve and displacement velocity sequence, and maintains a consistent mapping between behavior and physiological response in the time dimension. At this stage, the model input feature vector not only includes these dynamic motion and physiological indicators, but also concatenates static age and past injury history codes, as well as psychological fatigue scores that change over time, enabling the LSTM network to learn how these individual baseline features regulate the relationship between dynamic load and physiological response. The input features are normalized and differentially enhanced before entering the Long Short-Term Memory (LSTM) network, which uses its gating mechanism to jointly represent the short-term fluctuations and long-term trends of the training process. This study employs a three-layer LSTM network structure, a depth chosen to balance model complexity and performance. A single-layer LSTM struggles to fully capture the multi-level nonlinear mappings between kinematics, physiology, and injury risk; while deeper networks (e.g., four or more layers) significantly increase the number of parameters, potentially leading to overfitting with limited training data. The three-layer structure has proven sufficient to extract adequate abstract features from multi-source time-series data while maintaining good generalization ability. The network superimposes multi-scale feature channels inside the hidden layer, and the model can capture the nonlinear correlation patterns of load increase, recovery fluctuations and abnormal fatigue through adaptive weight adjustment. The output layer uses a time-aligned probability distribution to describe the continuity of training state changes and feeds the prediction results back to the feature extraction module to achieve iterative optimization [9,10]. To avoid information decay and address the vanishing gradient problem that may arise from 12 weeks of continuous training data, the model introduces residual channels in its temporal connection structure and employs a gradient clipping strategy during training. Residual channels allow gradients to propagate directly across layers, effectively mitigating the vanishing gradient phenomenon in deep networks. Gradient clipping, by limiting the gradient norm, prevents gradient explosion that may occur due to excessively long time steps, thus ensuring stable training of the model on long-term sequences and enhancing its sensitivity to key segments. The entire mapping process maintains its responsiveness to individual training rhythm and environmental conditions, achieving a unified expression of high-frequency signals and low-frequency physiological data.

D. BAYESIAN RISK ASSESSMENT MODULE

The Bayesian risk assessment module dynamically quantifies the relationship between training load fluctuations and physiological responses through a probabilistic inference mechanism. The system receives a continuous data stream from the time-series feature mapping output, integrates exercise intensity, speed changes, heart rate fluctuations, and recovery rate into a set of state features, and constructs a prior distribution structure based on historical data. Unlike traditional Bayesian models, the prior distribution in this module is not a universal distribution, but rather a personalized distribution calibrated by incorporating the athlete's age, history of injury, and current state of mental fatigue. For example, athletes with a history of injury will be assigned a higher prior probability of risk under specific loads. Upon data input, the module updates the individual training state probabilistically, generating a posterior distribution in real time to reflect the potential injury risk under load changes. To facilitate the transition from high-frequency input to stable risk assessment and to construct a statistically meaningful window, this module employs a dynamic sliding calculation method with a 10-minute risk analysis cycle. It's important to note that this 10-minute interval is not an inherent delay in the system's response. Internally, the system continuously updates the feature vector at a rate down to the second, with each new data point triggering a real-time update of the state features. The explicit output of the Bayesian posterior probability and the reassessment of the risk level are performed at the end of each 10-minute cycle to ensure sufficient statistical stability of the accumulated physiological and exercise load data, avoiding misjudgments due to single abnormal fluctuations. When the load growth rate exceeds a set threshold or the stability of the heart rate interval weakens, the system re-estimates the current risk distribution and records the trend. Risk updates are achieved through weighted smoothing, ensuring that short-term abnormal fluctuations do not interfere with overall judgment and maintaining the ability to capture fatigue accumulation trends. The risk index output by the module evolves continuously over time and is input into the subsequent early warning and control system in the form of a numerical matrix, achieving a quantitative mapping from training load to injury probability. This process ensures the traceability of load monitoring and the dynamic consistency of assessment results. The output data from the Bayesian risk assessment module is shown in Table 1:

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL DATA AND SETUP

The results of the experiment were based on the 12 weeks of training monitoring observations of a professional football club, in which 100 adult male players were involved. This study was approved by the Ethics Committee of Zhengzhou Vocational College of Finance and Taxation, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form. They composed the two data acquisition systems of a GPS positioning module, three axis accelerator, heart rate monitoring belt, and recovery assessment equipment with multi-source synchronous timed to milliseconds. Before the experiment began, we collected data on the age and injury history of all players. Throughout the 12-week monitoring period, the players' mental fatigue was assessed using the RESTQ-Sport scale before each daily training session. Exercise load, displacement, speed, heart rate variability, and nighttime recovery status were monitored in every training session. The sampling rate of the GPS was 10 Hz and the frequency of the heart rate signal was 1 Hz. The training material discussed endurance running, sprinting, ball control training. The data was after cleaning and normalization was keyed into the CDE-driven framework of feature extraction and time modeling. The LSTM network was set in the form of three layers of a hidden unit structure, grid search was used to find the most optimum learning rate, and early stop mechanism was added to avoid overtraining. The Bayesian risk assessment module generates a unique prior distribution for each player based on their training history, age, injury history, and mental fatigue status. For players lacking an injury history, the model employs a group prior strategy. This involves first constructing an initial prior distribution based on load risk statistics of players in the same position and age group, and then making personalized adjustments by incorporating basic physical fitness test and low-intensity acclimatization training data from the player's first three weeks with the team. This effectively avoids risk assessment bias during the "cold start" phase. The module updates the posterior distribution every 10 minutes. This was implemented on a full platform based on an NVIDIA RTX 3090 and the results of the model were shown as a dynamic risk curve. Table 2 shows the experimental data settings and acquisition parameters.

The table shows the main data sources and corresponding sampling settings. Exercise load indicators (speed, acceleration) are sampled at high frequency to ensure accuracy, physiological and recovery data are updated at low frequency to support long-term trend modeling, and environmental data is used to correct the training context to ensure the model's adaptability to scene changes.

B. EXPERIMENTAL RESULTS

This paper was a systematic model of the dynamic interaction between the individual training load and the risk of injury in terms of a context-driven (CDE) time-series analysis model. The model was shown to have good nonlinear mapping capacity in learning the interactive properties of multidimensional physiological indicators using 12 weeks of continuous training monitoring data, and was capable of revealing possible injury risks under different load trends. In particular, by incorporating age, history of injury, and state of mental fatigue, the model is able to distinguish different risk probabilities for athletes with different baseline risks under the same load. For example, an athlete with a history of injury may be identified by the model as having a risk level comparable to that of a healthy athlete under a high load under a moderate load. The model evaluation part took five athletes as regular individuals, and compared the mean load of training, the values of risk as predicted, and the actual rates of injuries. The results are shown in Figure 1.

The deviation between the model's predicted injury risk and the actual injury incidence was controlled within $\pm 3\%$, indicating that the model can maintain stable predictive performance under different load levels. Overall, the injury risk and actual incidence of players with higher training loads (such as P04 and P02) were significantly higher than those of the low-load group (such as P03), reflecting the model's good sensitivity to the coupling relationship between load and risk.

C. MODEL PERFORMANCE ANALYSIS

The model can capture the risk response caused by short-term load fluctuations and issue risk signals in advance when the load suddenly increases or recovery is insufficient, demonstrating a certain early warning capability. The trend of the predicted results is significantly positively correlated with the athlete's actual injury record ($r = 0.89$, $p < 0.01$), indicating that the model output can effectively reflect the individual's true physiological response state. The CDE-driven time series analysis model shows significant advantages in training load monitoring and injury risk prediction. Compared with traditional prediction models based on univariate factors (such as heart rate or exercise speed), the CDE model improves the matching accuracy in nonlinear load feature identification by approximately 12.6%, with an overall prediction accuracy of 91.3%. To more comprehensively evaluate the model's performance in the high-risk scenario of injury prediction, its sensitivity (true positive rate, measuring the model's ability to correctly identify players who have actually suffered injuries) and specificity (true negative rate, measuring the model's ability to cor-

AEM JOURNAL .ORG

TABLE 1. Schematic diagram of output data from the Bayesian risk assessment module

Time series fragments	Training load index (AU)	Heart rate fluctuation amplitude (ms)	Recovery score	Dynamic risk value (%)	Updated risk confidence level (%)
T01	376	43.2	0.86	17.4	92.1
T02	402	39.8	0.79	25.6	90.4
T03	428	37.5	0.75	31.2	91.3
T04	451	35.9	0.70	33.5	93.0
T05	389	41.5	0.83	22.4	94.6
T06	367	44.1	0.88	15.9	95.2

TABLE 2. Experimental Data Setup and Acquisition Parameters

Data types	Sampling frequency	Data acquisition equipment	Number of data indicators
GPS displacement and velocity	10 Hz	Catapult Vector S7	6
Acceleration data	100 Hz	Axivity AX6	3
Heart rate and variability	1 Hz	Polar H10	4
Recovery and fatigue state	Once a day	WHOOP Band 4.0	2
Environmental and tactical context recording	Real-time training	CDE context module	5
Baseline characteristics of athletes (age, history of injuries)	Before the experiment	Medical records/questionnaires	2
Mental fatigue	Before daily training	RESTQ-Sport Scale	1

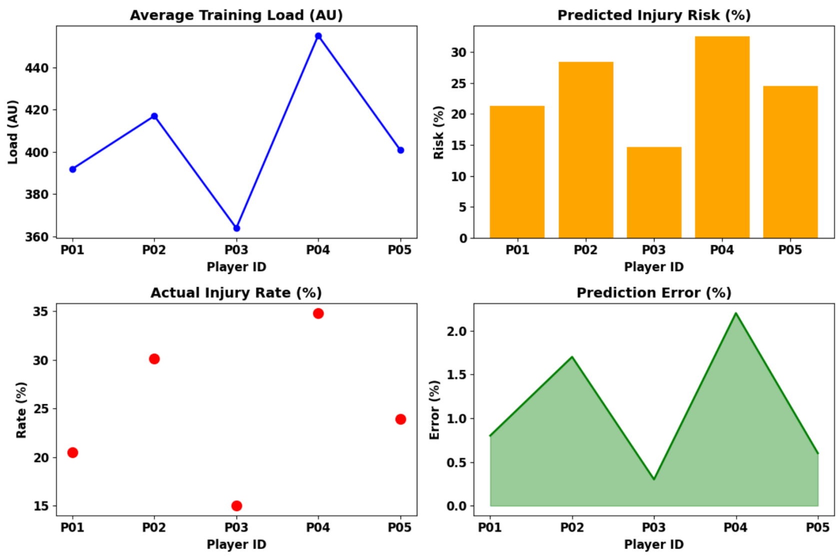

FIGURE 1. Experimental results data

rectly exclude uninjured players) were further calculated. Experimental results showed that the model's sensitivity was 88.2%, specificity was 91.5%, positive predictive value was 85.4%, and negative predictive value was 93.1%. The high negative predictive value indicates that when the model predicts a player is not at risk, the probability of them actually not suffering an injury is extremely high, which is of great value for developing safe training plans. The model demonstrates strong capabilities in short-term fatigue assessment and high-risk identification during periods of high workload, issuing risk warnings 1-2 days in advance, providing a basis for optimizing athlete training plans and preventing injuries.

V. CONCLUSION

The CDE-based training load time-series analysis technique, enhanced by high-precision data from integrated smart sensors, is highly adaptable and reliable in uncovering the nonlinear connection between training load development and injury risk. This synergy between advanced sensing and algorithmic sensor-based modeling allows for the seamless capture of biomechanical variables, ensuring the stability and accuracy of risk predictions in complex athletic environments. The model can detect individual load variation to cause changes in the physiological responses through context-awareness and event association mechanisms, with the goal of producing dynamic mapping and the early detection of the risk states, gates the scientific support of the training plans optimization and of fatigue control. This model framework provides a connection between training data and physiological monitoring and risk assessment that increases the individuality and intelligence of prediction of injury risk. Nevertheless, the sample size and feature dimensionality still remains a limitation of the model. Further studies can explore the migration of cross-season data and adaptive feature selection to enhance the generalization performance of the model, specifically by integrating the durability and signal stability of advanced smart textile sensors into real-time feedback optimization. This technical synergy will ensure the long-term reliability of training load monitoring as the physical properties of the integrated textile fibers evolve through rigorous athletic use.

AUTHOR CONTRIBUTIONS

Meng Wang designed, collected and analyzed the data, and drafted the manuscript. Meng Wang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Meng Wang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the Ethics Committee of Zhengzhou Vocational College of Finance and Taxation, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] X. Wei, H. Chen, N. Wang, et al., "A review of research on football training load based on quantitative data," Journal of Chengdu Sport University, vol. 50, no. 6, pp. 106-116, 2024, doi: 10.15942/j.jcsu.2024.06.013.

[2] Z. Su, "Research on monitoring of football training load," Contemporary Sports Technology, vol. 14, no. 18, pp. 4-7, 2024, doi: 10.16655/j.cnki.2095-2813.2024.18.002.

[3] J. Wang and Z. Liu, "Effects of high-load resistance and low-load plyometric training on the explosive power of young female football players," Journal of Lanzhou University of Arts and Sciences (Natural Science Edition), vol. 39, no. 2, pp. 110-115, 2025.

[4] Z. Yu and R. Zheng, "A comparative study on the effects of SSG training on athletes' load-bearing capacity and physical fitness in football," Journal of Shangqiu Normal University, vol. 37, no. 3, pp. 93-95, 2021, doi: 10.3969/j.issn.1672-3600.2021.03.025.

[5] Z. Wang, C. Li, and L. Zuo, "Research on the Application of Intelligent Assistive Equipment in Youth Football Training," Contemporary Sports Technology, vol. 15, no. 11, pp. 7-10, 2025, doi: 10.16655/j.cnki.2095-2813.2025.11.003.

[6] M. Hostrup and J. Bangsbo, "Performance adaptations to intensified training in top-level football," Sports Medicine, vol. 53, no. 3, pp. 577-594, 2023, doi: 10.1007/s40279-022-01791-z.

[7] H. Silva, F. Y. Nakamura, M. Beato, et al., "Acceleration and deceleration demands during training sessions in football: a systematic review," Science and Medicine in Football, vol. 7, no. 3, pp. 198-213, 2023, doi: 10.1080/24733938.2022.2090600.

[8] C. Miranda, V. Rago, J. R. Silva, et al., "Effects of traditional vs. complex strength training added to regular football training on physical capacities in U19 football players: a team study," Sport Sciences for Health, vol. 18, no. 3, pp. 671-680, 2022, doi: 10.1007/s11332-021-00833-9.

[9] E. Shanley, C. Thigpen, M. Kissenberth, et al., "Heads up football training decreases concussion rates in high school football players," Clinical journal of sport medicine, vol. 31, no. 2, pp. 120-126, 2021, doi: 10.1097/JSM.0000000000000711.

[10] G. Lisenchuk, V. Leleka, K. Bogatyrev, et al., "Fitness training in functional preparedness of highly qualified football players," Journal of Physical Education and Sport, vol. 23, no. 2, pp. 502-509, 2023, doi: 10.7752/jpes.2023.02062.