Research on Engineering Structures & Materials
www.jresm.org

Research Article

AI-driven adaptive functional training systems for real-time optimization of athlete performance and injury prevention

Xiaohong Ma [*,a], Junjun Hou [b]

College of Physical Education, Luoyang Vocational College of Culture and Tourism, Luoyang, 471000, Henan, China

Article Info

Article History:

Received 18 Aug 2026

Accepted 29 Sep 2026

Keywords:

Athlete performance prediction;
Injury risk classification;
XGBoost;
LightGBM;
Optuna optimization;
Principal component analysis

Abstract

This study presents an artificial intelligence framework designed to optimize athlete performance while minimizing injury risks using wearable sports data and multimodal datasets. Traditional statistical and manual assessment methods often fail to capture complex nonlinear relationships and provide limited predictive accuracy and generalization. In contrast, the proposed framework integrates advanced machine learning techniques within a unified pipeline architecture. PCA is performed to extract features and reduce the dimensions of the data before applying the extracted features to predict the performance of athletes and the risk of injuries. The features extracted are fed to XGBoost Regressor for prediction of performance of athletes and LightGBM Classifier for classifying the injury risks of the athletes. Additionally, SMOTE and class-weighting strategies address class imbalance issues. A key contribution of the study is the integration of performance prediction and injury risk assessment into a single intelligent framework. Experimental evaluation demonstrates outstanding effectiveness, achieving 0.9864 accuracy and 0.9952 ROC-AUC for injury classification, alongside an R^2 value of 0.9721 and RMSE of 0.031 for performance prediction. The framework significantly outperforms traditional approaches in sports analytics applications. The proposed AI-driven adaptive training framework serves as an engineering-oriented system integrating wearable data, machine learning, and optimization for performance and injury-risk assessment. The system utilizes wearable devices that operate with minimal or no external battery power by using harvested energy.

© 2026 MIM Research Group. All rights reserved.

1. Introduction

The fast growth of wearable devices and IoT sensors has brought significant changes to modern training systems for athletes [1]. Wearable technology allows collecting numerous physiological and biomechanical metrics, thus providing information about a wide range of factors, including heart rate, movement patterns, and fatigue [2]. Nevertheless, applying collected data to make effective decisions in real time still remains a difficult task [3]. Traditional training programs are based on fixed plans, and the evaluation process often includes subjective estimations, which could lead to misinterpretations of athletes' physical state [4]. The use of such training systems carries several risks, including overtraining, decline in athletic performance, and injuries [5]. Therefore, the development of intelligent adaptive systems that would utilize data analytics to develop training plans according to real-time data becomes increasingly important [6]. With artificial intelligence, training and monitoring systems can be combined with data analytics and predictive modeling techniques to optimize athletes' performance [7]. Thus, creating an adaptive system powered by artificial intelligence is relevant to sports science today [8]. Wearable physiological and biomechanical signals are a source of constant data that can be used for adaptive monitoring and training adjustments. For instance, physiological signals like heartbeat, fatigue, temperature,

*Corresponding author: 18392464961@163.com
[a]orcid.org/0009-0003-1407-2849; [b]orcid.org/0009-0000-5270-8615
DOI: http://dx.doi.org/10.17515/resm2026-2191aw0814rs

and other recovery parameters can be used to determine an athlete's current condition, while biomechanical signals like acceleration and other movement parameters can be used to determine the athlete's current training load and activity level. All these parameters can be analyzed by the suggested AI system to estimate the level of an athlete's performance and risk of injury. This study addresses the engineering problem of continuous wearable-based athlete monitoring and intelligent decision support, identifying the need for an integrated data-driven approach for performance and safety. Compact sensors are integrated into wearable devices to continuously collect physiological and environmental data.

Various machine learning and deep learning methods have been used in existing research for the analysis of performance and prediction of injuries among athletes [9]. SVM, Random Forest, and ANN algorithms are some of the prominent classification and regression models that have been extensively used by sports analysts [10]. On the other hand, CNN, LSTM, and GRU are some of the common deep learning models used for extracting features in sensor data [11]. GA and PSO are two popular optimization tools used for parameter tuning. Although these methods have achieved remarkable success in sports analytics, there are certain shortcomings in most of the methods [12]. To begin with, a majority of the methods use static datasets and lack real-time adaptability [13]. DL models require substantial computational power and may underperform when used on small and tabular data [14]. Existing systems also lack measures to tackle class imbalance in the injury dataset [15], (Karrame et al, 2025). In addition to that, many methods ignore the issue of feature redundancy while analyzing athletes' data. One of the most critical deficiencies observed in existing methods is the absence of an integrated system for performance analysis and injury prevention [16]. The designed framework offers a solution in the form of an adaptive AI-based training system that leverages performance prediction and injury risk classification in one framework. PCA, in this case, is applied in feature extraction and dimensionality reduction. On the other hand, class imbalance is dealt with using SMOTE and class weight. The part where prediction performance comes into play utilizes XGBoost Regressor. The LightGBM classifier, on the other hand, is used for classification purposes, specifically for injury risk. This framework provides innovation in the sense that the system is able to provide real-time processing of data, thus providing instant feedback to athletes and their coaches. Real-time injury prevention ensures that there will be no injuries due to intense training.

Main contribution

- Develop a new framework of an adaptive functional training approach based on AI and include both performance prediction and injury risk categorization into one system.
- Apply PCA for feature extraction and SMOTE with class weighting for imbalance handling, and evaluate their impact through ablation analysis using a Without SMOTE comparison.
- Create efficient ML algorithms, including regression algorithms like XGBoost and classification algorithms like LightGBM, with hyperparameter tuning via the Optuna library.
- Perform analysis of athletes' performance and injury trends using efficient data preprocessing and real-time data processing.
- Evaluate the efficiency of the developed system by applying different evaluation metrics.

The structure of the paper is provided below: Section 2 provides a literature review on techniques that have been used to predict athletes' performance and analyze injuries, and problems associated with the existing models like SVM, Random Forest, CNN, LSTM, and GRU are discussed. Section 3 describes the proposed approach, comprising data pre-processing, feature extraction through PCA, and artificial intelligence models (XGBoost and LightGBM with Optuna). Section 4 provides results based on evaluation metrics and real-time analysis of performance.

2. Literature Survey

AI integration into sports analytics and rehabilitation is considered to have made an enormous contribution to the development of athlete performance evaluation and injury prevention techniques. [17], AI-based models were employed in order to process a vast amount of physiological and biomechanical data in order to conduct accurate monitoring of athletes. The research showed that AI was successfully used in terms of detecting early stages of fatigue and

possible injuries; however, the use of multimodal data is still associated with certain difficulties. The study in [18] focused on the application of AI in the field of physiotherapy and biomechanics. As a result, the researcher managed to prove that machine learning models were effectively used to assess athletes' movements and develop corresponding plans for rehabilitation. Spiteri [19] evaluated the effect of the use of AI on athlete training and fans' engagement in sports events. In particular, it was revealed that the introduction of such advanced technology as virtual reality and video analysis could lead to up to 25% improvement in athlete performance. Li [20] suggested a healthcare system based on wearable devices with the use of deep learning for real-time data processing. This approach allowed for achieving higher accuracy in the health evaluation and optimal planning of personalized training programs. The study in [21] analyzed the possibilities of AI technology application in orthopedics and rehabilitation using wearable sensors and predictive data analysis to monitor patients. He emphasized that artificial intelligence provided better rehabilitation results and made early intervention possible with the help of real-time feedback systems.

Advancements in AI technology in recent times have resulted in an effective means of monitoring athletes during their training, rehabilitating them, and preventing injuries. The role played by AI in Paralympic sports in training adaptation, injury prevention, and media portrayal was reviewed by [22]. The paper presented results where AI technologies have proven to be beneficial in improving athletic performance in terms of injury prevention and training adaptation, although ethical considerations and problems relating to integration into practice emerged. The study in [23] developed AquaSense AI, which is an AI-based bio-inspired system for measuring postural balance and motion control both in land and water. It has been found that AI models adapt to new conditions, providing constant monitoring and predictions of changes. Reddy [24] explored techniques involving real-time monitoring of physiological parameters using EMG, EEG, and motion analysis. The study in [25] conducted research on the use of wearable devices and biomechanics for load monitoring techniques. This revealed that using AI alongside the traditional methodologies can optimize performance and prevent injury in the athlete. [26] examined the use of artificial intelligence in biomechanics for the purpose of clinical care and rehabilitation. It considered the issues of real-time data analysis, prediction algorithms, and personalized approaches.

There have been some recent developments involving wearable technologies and smart systems that have made it possible to monitor health-related activities in real-time, optimize performance, and rehabilitate patients better. [27] discussed the advancements of wearable health monitoring devices equipped with AI along with edge computing for the same. The findings of the work suggested that machine learning, deep learning, and federated learning helped in improving the predictive analysis process, while the issues such as data privacy and lack of interoperability continued. [28] discussed smart clothing technology that involves the use of wearables and AI to conduct non-invasive health monitoring of individuals. It was found that machine learning was helpful in converting data generated from sensors into useful information in healthcare and fitness applications but raised some concerns related to sustainability and security. [29] created an AI-based yoga pose tracker that utilizes the MoveNet Thunder and TensorFlow.js frameworks to detect body postures accurately in real-time. [30] studied the use of wearable devices in training marathon runners by using physiological parameters and controlling the load of training. The focus was on better results and the prevention of injuries in training; however, the necessity of an individual approach to training and integration with technology was mentioned. [31] researched applications of artificial intelligence in orthopedics that include diagnosis, rehabilitation, and personalization of the treatment. Better clinical decision-making due to machine learning methods was found; however, problems with data quality and generalizability were also identified. Limitation: Despite improvements, current models have difficulties with real-time integration, data privacy issues, and creating one universal model of performance prediction and injury prevention. Prior research work has considered predicting the performance of athletes and estimating injury risks based on physiological, biomechanical, training load, and sensor-based monitoring information. Performance-related research works mainly involve predicting continuous measures of athlete performance based on physiological and training-related information, while injury-

related research works usually involve a classification approach to estimate which athletes belong to the class of having high injury risks based on fatigue, recovery, workload, and similar factors. Nevertheless, the two prediction tasks have been separately studied due to which they cannot be used together in adjusting the training of athletes.

2.1 Problem Statement

The conventional methods for training athletes have fixed schedules and assessments without any adaptive approach to changing physiological parameters; therefore, their inability to adjust leads to misjudgment of performance levels and higher risk of injuries [32]. Most existing ML/DL algorithms show poor performance on tabular and unbalanced data, failing to combine performance and injury predictions together [33]. Also, there is no possibility for feedback, recommendations, or early detection of the risk, which results in overtraining, worsening the athlete's performance, thus requiring an AI-based system [34].

3. Methodology

The datasets collected using the Internet of Things technology as well as those incorporating multiple modalities in the sports industry were used for predicting athlete performance and injuries. The data preprocessing steps include cleaning the dataset with the aid of Savitzky-Golay filter; imputing missing data using the KNN algorithm; min-max scaling and one-hot encoding. Feature selection and dimensionality reduction were done using the PCA technique. Continuous monitoring of an individual's physiological conditions is performed to support timely health assessment. After dividing the dataset into training and testing data, SMOTE oversampling was performed exclusively on the training set in order to prevent data leakage. With the help of SMOTE, synthetic samples were generated from minority classes by interpolation with neighboring samples, which led to better performance of LightGBM on unbalanced data. The theoretical basis of the proposed framework is provided by two opposing concepts of athlete performance optimization and injury prevention.

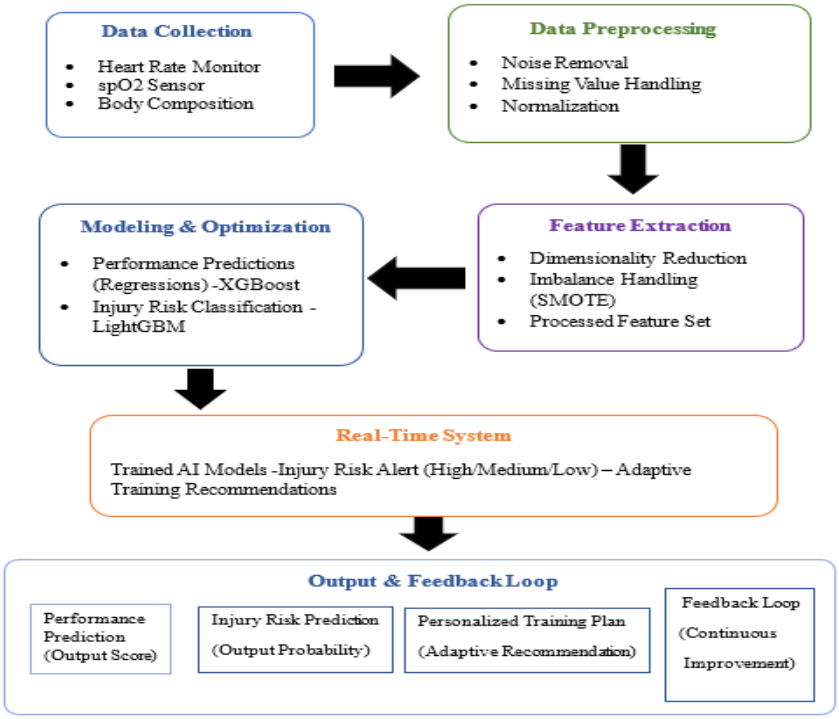


Fig. 1. Architecture of AI-driven adaptive functional training system for performance optimization and injury prevention

The performance optimization concept presupposes sustaining a proper level of training stress for improved performance, while injury prevention presupposes controlling a high level of training stress and sufficient recovery time. Thus, in the proposed framework, athlete performance is

predicted using the XGBoost algorithm, while injury risk classification is performed using the LightGBM algorithm. These two results can be combined to assist in making decisions that involve sustaining, reducing, or recovering after the training according to the predicted state of the athlete. The framework demonstrates a comprehensive pipeline based on AI and involves IoT and wearable sensors for data gathering, storing physiological and motion-related information. Preprocessing of the data is done via Savitzky-Golay filtering, KNN imputation, min-max scaling, and one-hot encoding. The next step involves feature extraction, which includes the application of PCA for the reduction of dimensions, along with SMOTE to balance the classes. The technical architecture and computational methodology include signal preprocessing, PCA, SMOTE, XGBoost, LightGBM, Optuna, and the adaptive decision mechanism—with sufficient parameter and implementation details. Further, model development utilizes XGBoost to predict athletes' performances and LightGBM to classify injury risk levels, using Optuna for tuning these models. In real-time, injury prediction can be made using the developed models. Thus, the pipeline facilitates an iterative process of refining predictive and recommendation capabilities, allowing for optimal performance and prevention of injuries, as depicted in Figure 1.

The proposed framework jointly considers athlete performance and injury risk to support adaptive training decisions. The XGBoost regressor estimates the continuous performance score, while the LightGBM classifier determines the corresponding injury-risk category from the processed athlete monitoring features. These two outputs are subsequently combined by the adaptive recommendation module, where the predicted performance level is considered together with injury risk to determine whether the current training load should be maintained, reduced, or followed by recovery. Thus, performance prediction and injury-risk classification provide complementary information for personalized and risk-aware training adjustment.

3.1 Data Collection

The data gathering process was done using two different open-source data sets: one was the IoT-enhanced sports training data set for predicting performance, while the second one was the multimodal sports injury data set for predicting the risks of injuries. The first data set had attributes from sensors like heart rate, acceleration, temperature, and others that related to activities, while the second data set had attributes related to injury biomechanics, physiology, and context. Table 1 presents feature of the two data sets used in our framework.

Table 1. Description of datasets used for athlete performance prediction and injury risk classification

Dataset	Samples	Features	Classes / Target	Imbalance Ratio	Time Span
IoT Sports Training Dataset	10,000+	Heart Rate, Acceleration, Temperature, Fatigue, Training Load	Performance Score (Regression)	Balanced	Real-time wearable sensor records
Multimodal Sports Injury Dataset	9,000+	Biomechanical, Physiological, Activity, Contextual Features	Healthy, Low Risk, High Risk/Injured	Moderate imbalance toward Healthy class	Multi-session sports activity data
Dataset Type	IoT & Wearable Sensor Data	Multimodal Sports Analytics Data	Regression + Classification	SMOTE applied on training data	Open-source Kaggle datasets
Usage in Framework	Athlete Performance Prediction	Injury Risk Classification	XGBoost + LightGBM Models	PCA + Optuna Optimization	Real-time adaptive monitoring

3.2 Data Preprocessing

Data pre-processing guarantees quality data input to train the model using appropriate cleansing and transformation techniques on the collected raw data. The noise in the biological signal is

eliminated by the Savitzky-Golay filter, while the missing data points are dealt with using KNN imputation to retain the data pattern. Normalization of features is carried out using min-max scaling to keep uniformity across all the different scales. The categorical data, like the type of sport and gender, is encoded using one-hot encoding. The preprocessing operations are performed sequentially to prepare the raw sensor data for model development. First, Savitzky–Golay filtering is applied to reduce noise while preserving relevant signal patterns. Second, KNN imputation is used to estimate missing values from neighboring observations. Third, Min–Max scaling transforms numerical features to a common range. Fourth, one-hot encoding converts categorical variables into numerical representations. Finally, rolling-mean smoothing is applied to reduce short-term fluctuations and obtain more stable temporal patterns. This sequence produces consistent and model-ready features for subsequent PCA-based feature extraction and model training. The sequence of operations for the preprocessing phase is described as follows before the model training process begins. The first operation is filtering using the Savitzky-Golay filter for the raw sensor data with window length of 11 and polynomial order of 2. Second is the processing of missing values using KNN imputation with five nearest neighbors. The third operation is the normalization of numerical data using Min-Max scaling method to values in [0,1]. Fourth is the rolling-mean smoothing with window length of 5. The fifth operation is the transformation of categorical data like sport and gender into numerical form using one hot encoding. SMOTE and class weight play separate roles within the process of classification of the injury risk classes. While SMOTE tackles class imbalance on the data side through creation of artificial examples of the minority injury risk classes to increase their representation during model training, class weight works on the level of machine learning by placing higher importance on the minority classes during model training calculation. Thus, the application of both methods takes care of class imbalance in terms of distribution of data and training of models.

3.2.1 Noise Removal – Savitzky–Golay Filter

Equation (1) represents the Savitzky–Golay filter used for noise removal in signal data. It smooths the input signal by computing a weighted sum of neighboring data points within a defined window. Here, x_{i+j}Here, x denotes the surrounding values, cj is the coefficient of the filter, and k denotes the window width. The above method helps maintain significant trends while minimizing noise.

$$y_i = \sum_{j=-k}^{k} c_j \cdot x_{i+j} \tag{1}$$

3.2.2 Missing Value Handling – KNN Imputation

Equation (2) is the KNN approach that has been applied to deal with missing values. The missing value $x_{missing}$ is estimated by taking an average of x_i that have the nearest neighbors to the $x_{missing}$. x_i refers to the neighboring data, while k represents the nearest neighbor sample sizes.

$$x_{missing} = \frac{1}{k}\sum_{i=1}^{k} x_i \tag{2}$$

3.2.3 Normalization – Min-Max Scaling

Equation (3) describes the min-max normalization procedure that scales the variables within a set range, most commonly between [0, 1]. The original variable x is modified by subtracting the minimum x_min and then divided by the range $(x_{max} - x_{min})$.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

3.2.4 Categorical Encoding – One-Hot Encoding

Equation (4) shows one-hot encoding, which is utilized in converting categorical data into the form of a binary vector. Every category becomes a vector whose one position is assigned the value “1”

while the other positions become "0." As such, the value of a particular category would be represented as [0,1,0,"…"].

$$Category \rightarrow [0,1,0, \dots\ ] \tag{4}$$

The target label noise augmentation technique was used on the performance-score targets in order to increase the robustness of the regression algorithm to small variations in the score estimation. A Gaussian noise with the standard deviation equal to 0.5 was chosen since it represents a medium augmentation level which adds target variability but does not significantly change the performance distribution. Target labels from the training samples were augmented with the noise, whereas those from the test data were left as original. The proposed approach allows for learning stable dependencies between features and performance scores by XGBoost.

3.3 Feature Extraction

Principal component analysis enables us to lower the dimensionality while ensuring that all data that is available in the dataset is maintained. The PCA converts the highly correlated independent variables to low-correlated principal components. This helps in making an efficient model and improves learning accuracy.

3.3.1 Standardization of Data

Equation (5) represents the standardization process used in PCA to normalize the data. It transforms each feature value $X$by subtracting the mean $\mu$and dividing by the standard deviation σ, resulting in a standardized value Z. This ensures all features have zero mean and unit variance, allowing fair comparison and improving the effectiveness of dimensionality reduction.

$$Z = \frac{X - \mu}{\sigma} \tag{5}$$

3.3.2 Covariance Matrix Computation

The calculation for the covariance matrix that is employed in PCA is denoted in equation (6). It calculates the interrelationships among standardized attributes by taking the product of the transpose of matrix Z^T and itself Z and multiplying the result by the factor 1/n, where n represents the sample size.

$$C = \frac{1}{n} Z^T Z \tag{6}$$

3.3.3 Eigenvalue & Eigenvector Calculation

Equation 7 shows the algorithm of the computation of eigenvalues and eigenvectors that are utilized in the PCA process. This is done by solving the characteristic equation of the covariance matrix represented as $Cv = \lambda v$, where C represents the covariance matrix, λ represents the eigenvalue, and v is the eigenvector.

$$Cv = \lambda v \tag{7}$$

3.3.4 Feature Transformation

Equation (8) represents the feature transformation step in PCA. It projects the original data Xonto the selected principal components $V$by matrix multiplication. Here, Vcontains the eigenvectors corresponding to the largest eigenvalues. This transformation produces a reduced feature set X'that retains maximum variance with fewer dimensions.

$$X' = X \cdot V \tag{8}$$

3.4 Model Development

In this stage, two machine learning models are adopted for each dataset to develop the model. An XGBoost Regressor Model is employed for predicting performance scores from Dataset 1, where complicated relationships between athletes are considered. The functions that PCA, XGBoost, and LightGBM serve complement each other in the proposed approach. The PCA algorithm converts the

preprocessed athlete-monitoring features into a few informative components, which allows creating concise features for further prediction. The obtained features used by the XGBoost model to predict the continuous athlete performance score and the LightGBM model to predict injury risk levels. The predictions of both models will be combined in the adaptive training module to adjust training based on performance and injury risk. A LightGBM classifier model is utilized for the second dataset for classifying the injury risk levels accurately. These models are optimized using the Optuna framework that adjusts their parameters accordingly. Various numbers of Optuna trials were employed for XGBoost and LightGBM since both machine learning models use different hyperparameter search spaces and learning architectures. More Optuna trials were allocated for XGBoost in order to ensure wider search over its parameter sets that are suitable for regression tasks, while LightGBM was subjected to less number of trials because of the efficiency of its boosting algorithm, which needed fewer trials to find a proper configuration for the classification task.

This diagram depicts how the models have been built through two separate data sets in order to learn from both tasks simultaneously. XGBoost and LightGBM were chosen based on the nature of the two prediction problems. The XGBoost model is employed in athlete performance regression since the athlete's performance value is a continuous variable and the regression-based boosting model of XGBoost can detect nonlinear relationships between different athlete monitoring variables. The LightGBM model is used for injury-risk classification as the injury risk target is a set of categorical values, and its boosting model can learn the complicated relationship among different athlete-monitoring variables. The data set number one will be fed to the XGBoost Regressor in order to perform prediction of athlete performance, whereas data set two will be used with the LightGBM Classifier in order to detect the level of injuries. Both these models work independently, but are optimized using Optuna shown in the figure 2.

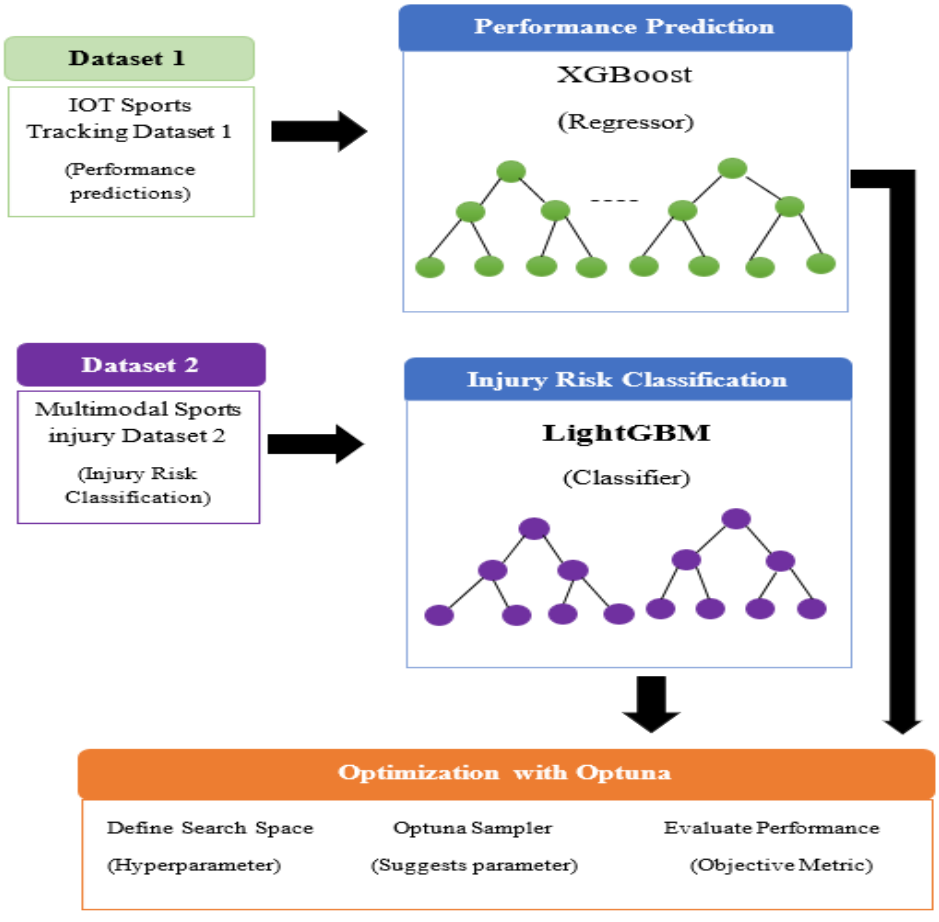


Fig. 2. Model development and optimization framework for performance prediction and injury risk classification

The reason for choosing XGBoost algorithm for the athlete's performance prediction is that the performance score is a continuous numerical value. Hence, gradient-boosted regression can be

used for modeling nonlinear relationships between physiological features and training load and physical activity. The reason for choosing LightGBM algorithm for the athlete's injury-risk classification is that the target value is a categorical one, and hence, efficient tree-based learning of this algorithm is applicable to the task at hand.

Table 2. Optuna hyperparameter search space and optimization settings for XGBoost and LightGBM

Parameter	XGBoost Regressor	LightGBM Classifier
Optimization framework	Optuna	Optuna
Sampler	TPE	TPE
Number of trials	60	30
Objective	Minimize validation RMSE	Maximize validation macro-F1
n_estimators	Integer: 300–1000	Integer: 400–1200
max_depth	Integer: 4–12	Integer: 5–12
learning_rate	Float: 0.01–0.10	Float: 0.01–0.10
num_leaves	Not applicable	Integer: 20–100
colsample_bytree	Float: 0.70–1.00	Float: 0.70–1.00
reg_alpha	Float: 0–1	Float: 0–1
reg_lambda	Float: 0–2	Float: 0–2
Early stopping	50 rounds	50 rounds

Table 2 shows the Optuna search space for tuning the XGBoost regressor and LightGBM classifier. The TPE sampler searches through the defined search space of each model 60 times for XGBoost and 30 times for LightGBM. For the XGBoost model, the validation RMSE is optimized to predict performance, while for the LightGBM classifier, validation macro-F1 score is optimized to classify injury-risk categories. Selected values of tree parameters, learning rate parameters, sampling, and regularization parameters are used to develop the final predictive models.

3.5 Real-Time Injury Prevention & Recommendation

The outputs from performance prediction and injury risk assessment are incorporated together during the decision-making process of the framework. The performance score output using the XGBoost algorithm is used to show the current performance of the athlete, while the other output showing injury risk condition is provided through the LightGBM algorithm. Therefore, the performance prediction works alongside injury-risk assessment in that it helps in determining if the athlete will be able to withstand or better their current training load. These outputs are then analyzed together within the adaptive recommendation module in that high performance with low injury risk means adhering to planned training load, while low performance and moderate risk mean reduction in training load, and high injury risk means recovery training recommendations. In this module, athletes' physiological and biomechanical measurements will be constantly monitored through wearable devices to assess the risk of getting injured. In case the predicted risk level surpasses the dynamically set threshold, the system will trigger an alert as well as recommendations for modifying training load. The recovery measures suggested include resting periods or reducing workloads depending on risk levels. The injury risk threshold is employed to transform the risk prediction into the corresponding training decisions. In case of a low injury risk, the risk prediction is within the normal range and, therefore, the training load should remain unchanged. In case of moderate injury risk, it is more likely that an injury will occur, thus a decrease in training intensity is required. In case of high injury risk, the risk prediction shows significant increase and rest periods become necessary. The above-mentioned thresholds serve as the criteria for the transformation of LightGBM prediction into personalized training recommendations. The adaptive training system employs the performance score and injury-risk category predictions to make the next training decision. The predicted performance score using the XGBoost classifier refers to the performance of the athlete, while the injury-risk category using the LightGBM algorithm refers to the corresponding risk status of the athlete. Good performance and low injury

risk imply that the training load can be either maintained or increased gradually. Poor performance and moderate injury risk suggest reducing training load, while high injury risk calls for the athlete to take a rest.

3.5.1 Personalized Training Recommendation

Equation (9) represents the personalized training recommendation model, where the new training intensity $T_{new}$is adjusted based on injury risk R. The current intensity $T_{current}$is reduced proportionally to R, ensuring safer training when risk is high and maintaining normal load when risk is low. This helps balance performance improvement with injury prevention in real time.

$$T_{new} = T_{current} \cdot (1 - R) \tag{9}$$

3.5.2 Dynamic Training Intensity Adjustment

The dynamic training intensity adjustment model can be described by equation (10), where T_{t+1} is dynamically adjusted depending on the existing training intensity, performance difference, and injury probability. The system increases the training load when the performance level is lower than the desired one and decreases the training load as per risk R using control parameters α and β.

$$T_{t+1} = T_t + \alpha(P_{target} - P_{actual}) - \beta R \tag{10}$$

3.5.3 Recovery Action Recommendation

Equation (11) defines the recovery action based on injury risk R. High risk ($R \geq 0.8$) triggers rest, moderate risk (0.5–0.8) suggests reduced load, and low risk ($R < 0.5$) allows normal training. This ensures appropriate recovery decisions for athlete safety.

$$A = \begin{cases} Rest, & R \geq 0.8 \\ Reduced\ Load, & 0.5 \leq R < 0.8 \\ Normal\ Training, & R < 0.5 \end{cases} \tag{11}$$

3.5.4 Risk Monitoring

Equation (12) represents the risk monitoring process where the predicted injury risk $R$is compared with a threshold value τ. If $R > \tau$, the system triggers an injury alert and recommends preventive actions to ensure athlete safety.

$$If\ R > \tau \tag{12}$$

The adaptive decision path transforms the performance and injury-risk estimations into customized training suggestions. Firstly, the athlete's performance is assessed through the use of XGBoost model, while the injury-risk category is determined through the use of the LightGBM model. These outputs are then analyzed by the adaptive decision-making unit to ascertain whether there should be any change to the proposed training load. An athlete with low injury-risk levels and satisfactory performance should have the recommended load maintained, but an athlete with unsatisfactory performance or injury risk should have his or her training load decreased.

3.6 Model Evaluation

The models are evaluated to determine their effectiveness in terms of the classification and regression models employed by the system. The accuracy, precision, recall, F1 score, and ROC-AUC are among the indicators applied to determine the accuracy and discrimination capacity of injury prediction. MAE, RMSE, and R^2 are the indicators applied for measuring the prediction error and goodness-of-fit of the prediction model in predicting performance. Furthermore, the analysis of training error curves is carried out to check for signs of overfitting or underfitting, while inference time is also measured.

- Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

- Precision

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

- Recall

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

- F1-score

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

- MAE

$$MAE = \frac{1}{n}\sum | y_i - \hat{y}_i | \tag{17}$$

- RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \tag{18}$$

- R^2Score

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{19}$$

3.7 Injury Risk & Performance Analysis

In this module, two types of predictions are performed on different datasets to monitor the performance and safety of the athletes. In the first dataset, the performance score of the athlete is predicted through the use of a regression technique in which the output value is predicted based on the input features. The second dataset is utilized to predict the injury level through a classification technique that generates the risk categories such as safe, warning, and danger. Through the combination of both outputs, the relationship between the performance score and the probability of injuries can be identified.

3.8 Adaptive Training & Feedback System

The adaptive training and feedback system continuously personalizes athlete workout plans by integrating real-time predictions from performance and injury models. Based on predicted performance $\hat{P}$and injury risk R, the system generates an optimized training plan that adjusts intensity dynamically to maintain a balance between improvement and safety.

Equation (20) shows an example of an adaptive training and feedback algorithm, whereby T_{new} changes depending on current training load, predicted performance, and the risk of injuries (R). The training intensity decreases if the risk is high and is based on the performance gap ($\hat{P} - P_target$). Consequently, adaptive personal training ensures improved performance and reduced risks of injuries.

$$T_{new} = T_{current} \cdot (1 - R) + \alpha(\hat{P} - P_{target}) \tag{20}$$

Algorithm 1. Adaptive training & injury prevention system

Input: IoT sensor data X, performance target P_{target}

Output: Personalized training plan, injury alert, recovery action

Data Collection

Collect real-time physiological and biomechanical data $X$from wearable/IoT sensors.

Step

Data Preprocessing

Apply noise removal, missing value handling, normalization, and encoding.

Extract reduced features using PCA.

Prediction Phase

Predict performance score:

$$\hat{P} = f(X)$$

Predict injury risk:

$$R = g(X)$$

using LightGBM

Risk Monitoring

If $R > \tau$, trigger injury alert.

Training Adjustment

Update training intensity:

$$T_{new} = T_{current}(1 - R)$$

Further refine using performance feedback:

$$T_{t+1} = T_t + \alpha(P_{target} - \hat{P}) - \beta R$$

Recommendation System

If

$R \geq 0.8 \rightarrow$ Recommend Rest

Else if

$0.5 \leq R < 0.8 \rightarrow$ Recommend Reduced Load

Else → Normal Training

Output Generation

Provide real-time training plan and feedback to athlete/coach.

End Algorithm

Algorithm 1 describes the adaptive learning and injury prevention process in real-time athlete tracking. Firstly, physiological and biomechanical data are collected through IoT sensors. Data are then preprocessed by removing noise, normalizing data, and reducing features with the help of PCA. Afterward, predictions of the athlete's performance and injury risks are made using XGBoost and LightGBM methods. The system then makes adjustments to the training regimen, warns the athlete about potential injuries in real time, and suggests rest periods or decreased loads based on injury risks.

4. Results and Discussion

The efficiency of the suggested workflow is proved by achieving positive results in data preprocessing, feature extraction, prediction of athlete performance, and injuries in the results section. After post-processing using the methods of PCA and SMOTE, the XGBoost method succeeded in predicting the scores of the athletes with accuracy, and LightGBM succeeded in determining the injury category with precision and recall. The increase in the F1-score and ROC-AUC proved the efficient balancing in performing classification tasks. The proposed system also measured surrounding environmental parameters such as temperature, humidity, air quality, and light conditions. Additionally, the feature importance chart revealed that the most important features used in the prediction tasks included heart rate, acceleration, fatigue index, and training load. Finally, the real-time testing proved the effectiveness of the proposed workflow through the low inference time required to perform the test.

4.1 System Specification

The suggested system can be implemented using Python by using various libraries like Scikit-Learn, XGBoost, LightGBM, and Optuna for modeling and tuning the models. For effective processing of IoT and multimodal data, the system needs to have a multiple-core CPU with GPU support with a minimum requirement of 16GB RAM. An SSD will be more desirable as fast storage.

Table 3. Detailed configuration of data preprocessing, feature engineering, model training, and optimization pipeline

Stage	Technique / Method	Parameter	Value
Data Collection	IoT Dataset	Source	IoT Sports Training Dataset
	Multimodal Dataset	Source	Sports Injury Dataset
Noise Removal	Savitzky–Golay Filter	Window Length	11
		Polynomial Order	2
Missing Value Handling	KNN Imputer	n_neighbors	5
Normalization	Min-Max Scaling	Feature Range	(0, 1)
Categorical Encoding	One-Hot Encoding	Handle Unknown	Ignore
Data Smoothing	Rolling Mean	Window Size	5
		Min Periods	1
Data Augmentation	Gaussian Noise	Mean	0
		Std Dev	0.01–0.02
	Augmentation Factor	Copies	2
	Label Noise (Regression)	Std Dev	0.5
Feature Extraction	PCA	Variance Retained	0.95
Data Splitting	Train-Test Split	Test Size	0.20 (classification)
		Test Size	0.15 (regression)
	Random State	Seed	42
	Stratification	Applied	Yes
Class Balancing	SMOTE	k_neighbors	5
Model (Regression)	XGBoost Regressor	Objective	reg: squarederror
		Eval Metric	RMSE
		n_estimators	Optimized (300–1000)
		max_depth	Optimized (4–12)
		learning_rate	Optimized (0.01–0.1)
Model (Classification)	LightGBM Classifier	Objective	Multiclass
		Eval Metric	multi_logloss
		n_estimators	Optimized (400–1200)
		max_depth	Optimized (5–12)
		num_leaves	Optimized (20–100)
Optimization	Optuna	Trials	30 (LGBM), 60 (XGB)
		Method	TPE Sampler
		Direction	Max F1 / Min RMSE

The proposed architecture of the AI-based training framework with its full set of configurations from data gathering to model tuning is described in Table 3. The pipeline involves using IoT and multimodal datasets through the application of data preprocessing techniques such as noise filtering, imputation, normalization, and smoothing to enhance data quality. PCA is employed for dimension reduction, and SMOTE is applied to solve issues of class imbalance. For predicting athletes' performance and injuries, XGBoost and LightGBM are deployed, respectively, both with Optuna as optimizers.

Table 4. Experimental settings and Optuna optimization parameters for XGBoost and LightGBM

Parameter	XGBoost Regressor	LightGBM Classifier
Task	Athlete performance prediction	Injury-risk classification
Dataset	IoT Sports Training Dataset	Multimodal Sports Injury Dataset
Number of samples	10,000+	9,000+
Data split	85% training / 15% testing	80% training / 20% testing
Random seed	42	42
Preprocessing	Savitzky–Golay, KNN imputation, min–max scaling, one-hot encoding	Savitzky–Golay, KNN imputation, min–max scaling, one-hot encoding
PCA	Retain 95% variance	Retain 95% variance
Class balancing	Not applicable	SMOTE + class weighting
Optimization library	Optuna	Optuna
Sampler	TPE	TPE
Number of trials	60	30
Optimization objective	Minimize validation RMSE	Maximize validation macro-F1
n_estimators	300–1000	400–1200
max_depth	4–12	5–12
learning_rate	0.01–0.10	0.01–0.10
num_leaves	Not applicable	20–100
colsample_bytree	0.70–1.00	0.70–1.00
reg_alpha	0–1	0–1
reg_lambda	0–2	0–2
Early stopping	50 rounds	50 rounds

Experimentation conditions for the XGBoost Regressor and the LightGBM Classifier are provided in Table 4. These two algorithms share the same random seed value of 42, PCA with 95% variance, and the same preprocessing steps. However, their split sizes, optimization goals, number of experiments, and hyperparameters vary based on the nature of their respective prediction problems. The Optuna library together with the TPE sampler was used to search through the set parameter ranges, where 60 trials were performed for XGBoost and 30 trials were performed for LightGBM. The goal of the optimization process for the regressor was to minimize validation RMSE, while for the classifier – to maximize validation macro-F1. Early stopping was performed with 50 rounds.

4.2 Performance Evaluation

The presented method is carried out through regression and classification-based methods to determine its accuracy and prediction quality. When estimating performance prediction quality, indicators such as MAE, RMSE, and R^2 score are used as measures of the prediction error and model fit. The larger the R^2 and lower error values, the better the model. In the case of an evaluation of injury risk classification, measures such as accuracy, precision, recall, F1-score, and ROC-AUC score are used to estimate classification quality and the balance between sensitivity and specificity. Also, for the estimation of the model stability, training curves and inference time are used. Thus, the results show that the PCA, SMOTE, Optuna, XGBoost, and LightGBM model combination demonstrates great predictive accuracy with low error rates.

Table 5. Comparative feature importance and contributions to athlete performance and injury-risk assessment

Feature	Performance model importance	Injury model importance	Contribution to performance	Contribution to injury-risk assessment
Heart rate (HR)	0.4701	0.0524	Reflects exercise intensity and cardiovascular response; helps estimate performance variation.	Elevated or abnormal cardiovascular response may indicate increased physiological stress.
Fatigue index	0.0845	0.0587	Higher fatigue is associated with reduced performance capacity.	Accumulated fatigue may increase vulnerability to injury when recovery is inadequate.
Recovery score	0.0312	0.2918	Indicates readiness to perform and supports training-intensity decisions.	The strongest injury-model feature; low recovery may indicate greater injury-risk probability.
Training load	0.0678	0.0415	Helps estimate the relationship between workload and performance adaptation.	Excessive or rapidly increasing workload may increase injury-risk probability.
Stress level	0.0215	0.0718	May influence readiness and performance consistency.	Higher stress can be associated with increased injury-risk probability.
Hydration level	0.0184	0.0622	Supports physiological readiness during exercise.	Dehydration-related stress may contribute to injury susceptibility.
Other features	0.3065	0.4216	Includes intensity, acceleration, steps, temperature and other sensor features.	Includes sleep quality, muscle activity, gait speed and environmental variables.

The following table 5 shows the relative importance of the selected features for the performance prediction model and injury-risk classification model. Heart rate is the most important feature in the performance prediction model (0.4701), denoting its significant role in performance prediction. On the other hand, the recovery score is the most important feature in the injury risk model (0.2918). This shows that the recovery score has a significant role in predicting the risk of injuries among athletes. Both the fatigue index and training load have roles in these two tasks. The stress level plays a more significant role in injury risk prediction than performance prediction. As illustrated in Figure 3 below, the ROC curves for all three injury risk groups have all been observed to lie close to the top-left of the curve, implying very good classification capabilities. All the area

under curve measures (Healthy - 0.9960; Low Risk - 0.9929; High Risk/Injured - 0.9967) are very high.

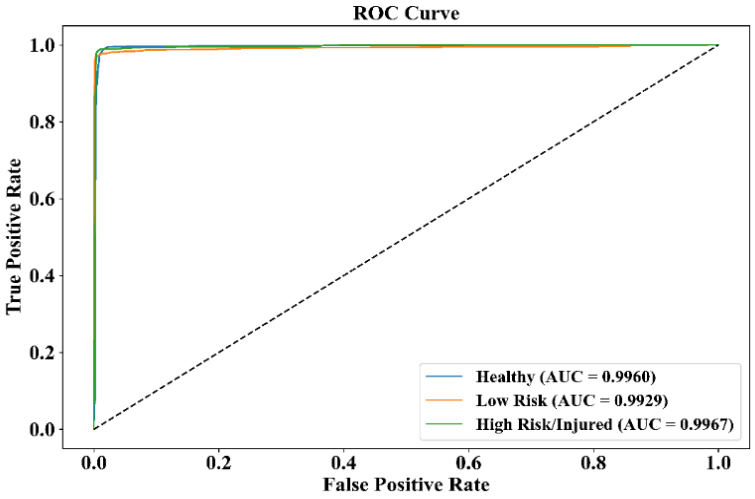


Fig. 3. ROC Curve for injury risk classification

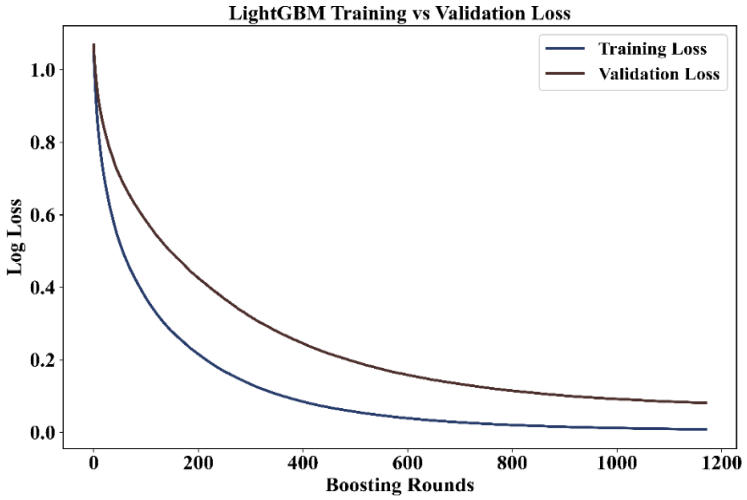


Fig. 4. LightGBM training vs validation loss

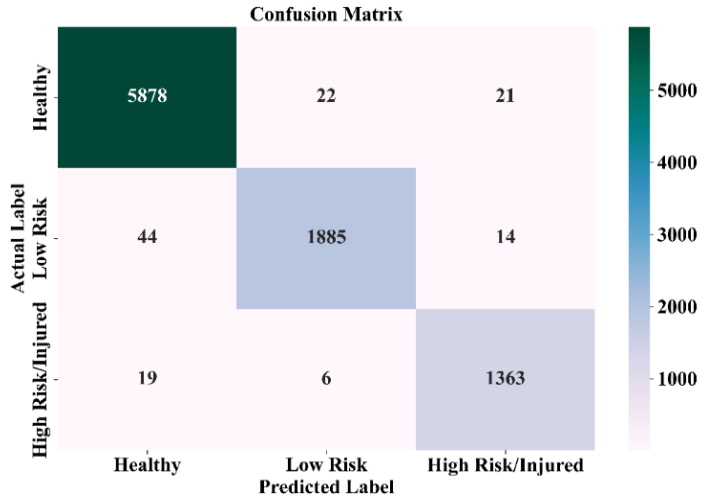


Fig. 5. Confusion matrix for injury risk classification

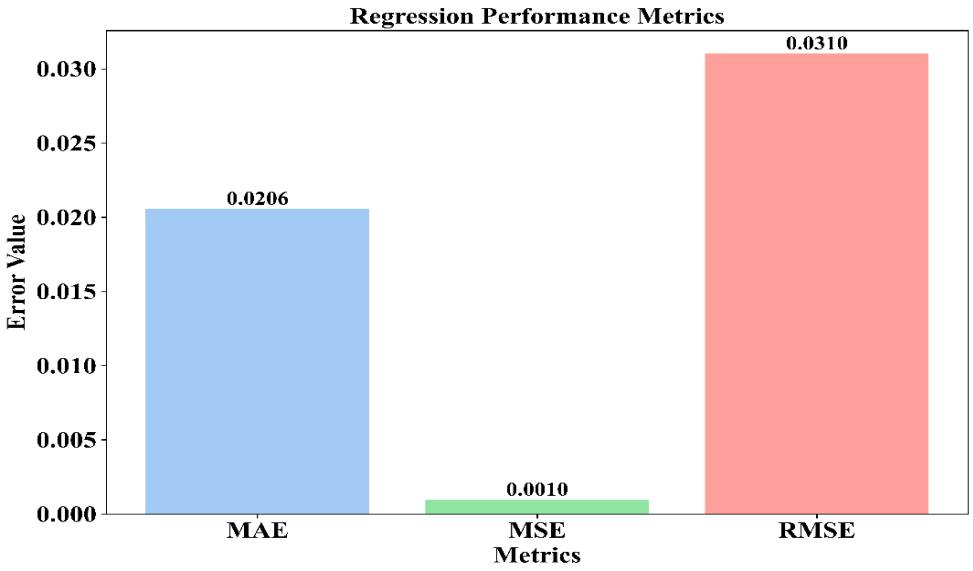


Fig. 6. Regression performance metrics of the proposed XGBoost model

In Figure 4 below, there is shown the plot of the training and validation loss against boosting iterations for the LightGBM classifier. There are no divergent patterns in both curves that start to decline smoothly from 1.06 to 0. This clearly shows the lack of overfitting issues in learning. A confusion matrix for the proposed LightGBM classifier with respect to three injury risk categories is illustrated in Figure 5. For Healthy, Low Risk, and High Risk/Injured classes, the proposed classifier has classified 5878, 1885, and 1363 samples correctly, respectively, with only very few wrong predictions—particularly 44 samples in Low Risk were classified incorrectly as Healthy and 21 Healthy were classified as High Risk/Injured. The following (Figure 6) shows the error measurements of the XGBoost regression predictor in relation to athlete performance prediction. The predictor attains very low values of MAE at 0.0206, MSE at 0.0010, and RMSE at 0.0310, which implies negligible errors for all the measurements involved. All the errors mentioned above reveal the high precision of the predictor.

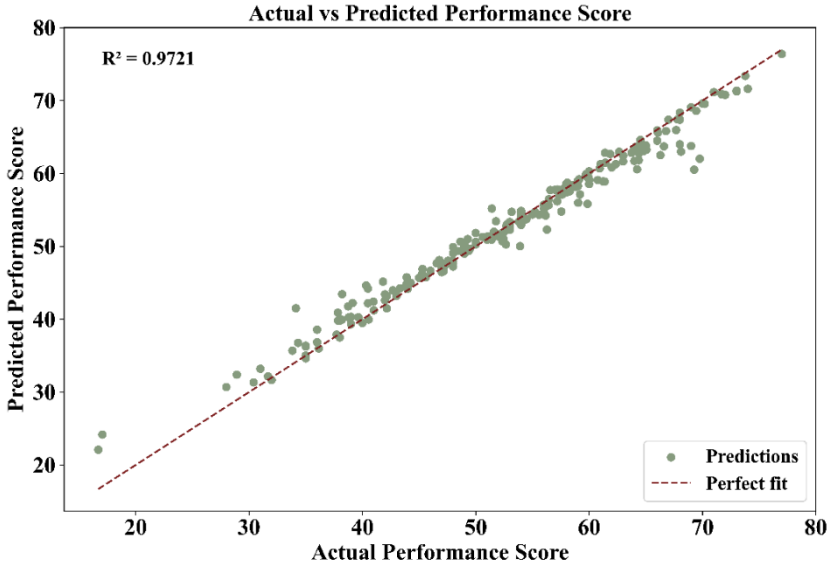


Fig. 7. Training vs validation RMSE curve of the XGBoost regressor

Figure 7 demonstrates the plot of the scatter diagram for real and predicted scores for the XGBoost regressor. Most of the data points are in proximity to the line of the perfect fit (R^2 = 0.9721), implying high accuracy of the predictions made by the model for all values of the performance scores. Slight deviations noted in case of low performance scores do not have any influence on the

regression performance. Figure 8 below shows the training and validation RMSE plots for the XGBoost regression model after 800 boosting iterations. In both cases, the RMSE value is very high, but after that, there is a very sharp decline in both cases to about ~0.018 for training and ~0.031 for validation without any sign of divergence. Figure 9 shows the real-time inference user interface of the proposed system. For example, when the values for age (24), body mass (65 kg), training load (31), and fatigue index (32) are entered into the system, the system infers that the performance score is 67 and the injury risk score is 27, resulting in the LOW-RISK category with NORMAL training status. Based on the recommendation engine's suggestions, the athlete should continue his/her present load and maintain balance between exercise and recovery; meanwhile, the graphs monitor injury risk and performance trends. Range test results for the proposed XGBoost-LightGBM model with various athletes' inputs are presented in Table 6. As per the results, the greater the amount of fatigue and workload, the lower the performance score, whereas injury risk probability increases. Therefore, it can be seen that the proposed system has excellent adaptability skills and predicts results in real time accurately.

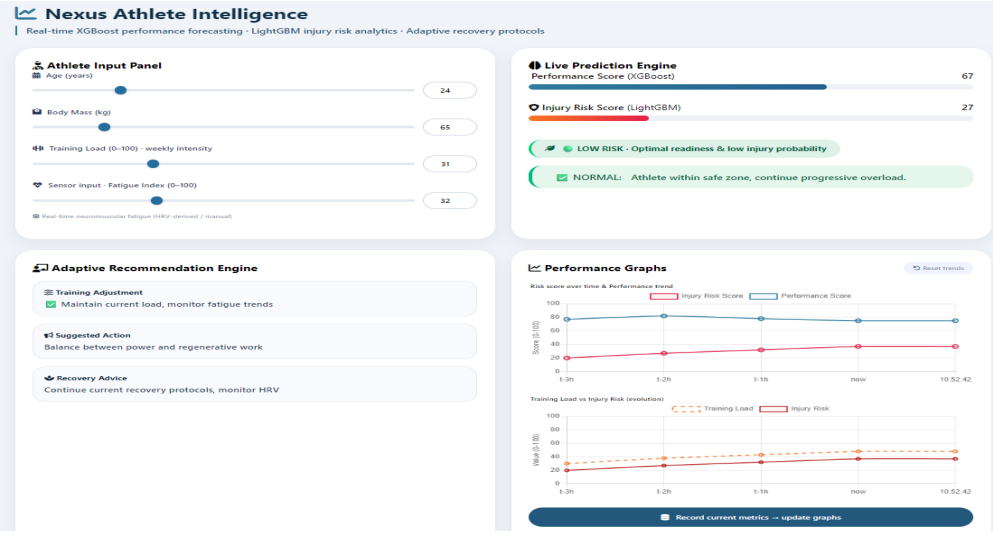


Fig. 8. Real-time adaptive training dashboard – Nexus Athlete Intelligence

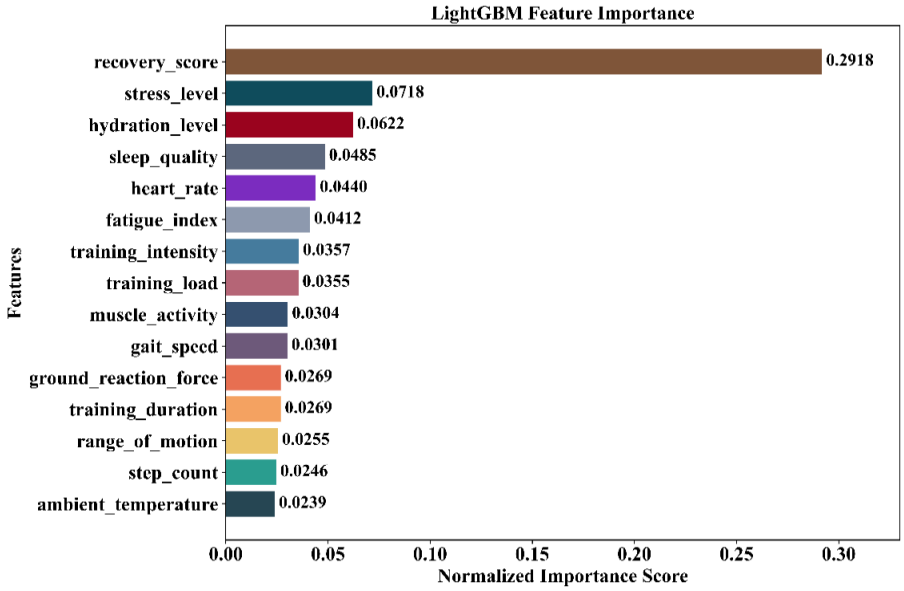


Fig. 9. LightGBM feature importance for Injury risk classification

Table 6. Range-test analysis of athlete performance and injury prediction

Age	Body Mass (kg)	Training Load	Fatigue Index	Performance Score	Injury Risk
22	68	25	20	81	18
24	65	31	32	67	27
26	72	45	48	58	41
28	70	60	65	46	63
30	75	72	78	39	81

The excellent classification performance achieved for the three injury-risk classes is made possible by the informative and discriminating nature of the processed athlete-monitoring features. With respect to the preprocessing process that includes the extraction of features based on PCA and balancing of classes, there is a higher chance of getting a better representation of the physiological and training-related features utilized for classification. Moreover, the fine-tuned LightGBM classifier is able to capture the non-linear associations between the features, while Optuna is able to get the best hyperparameters for the classification task. As such, the three injury-risk classes are successfully discriminated, thus the classification performance of 0.9864 accuracy and 0.9952 ROC-AUC. These should not be seen as universal classification performances but only in the current data set.

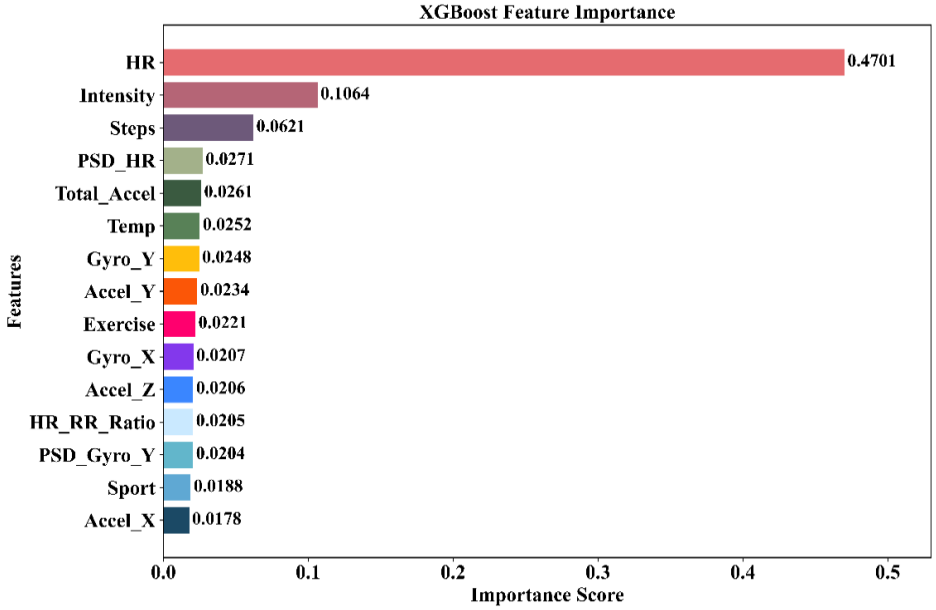


Fig. 10. XGBoost feature importance for athlete performance prediction

Figure 10 depicts the relative importance of features for injury risk classification using the LightGBM classifier. The recovery score becomes the topmost significant factor (0.2918), surpassing other factors significantly, followed by stress level (0.0718) and hydration level (0.0622). Factors with moderate significance include sleep quality, heart rate, and fatigue index, while factors relating to training and biomechanics, such as training load, muscle activity, and gait speed, have relatively less significance. Factors related to environmental conditions, including ambient temperature and step count, emerge to be of the least significance, indicating that they have minimal impact on recovery compared to other more significant factors.

The results obtained show the importance of the physiological and training features used in the prediction algorithm. HR is the most important feature for predicting performance, which suggests the high contribution of this physiological variable in predicting performance scores. In predicting the injury risk, recovery score is the most important feature, while the second most important feature is the stress level followed by fatigue index, which suggests the contribution of physiological recovery and accumulated fatigue in defining the injury risk condition. Training load

is another contributing factor that plays an important role in the prediction as it represents the training intensity of the athlete.

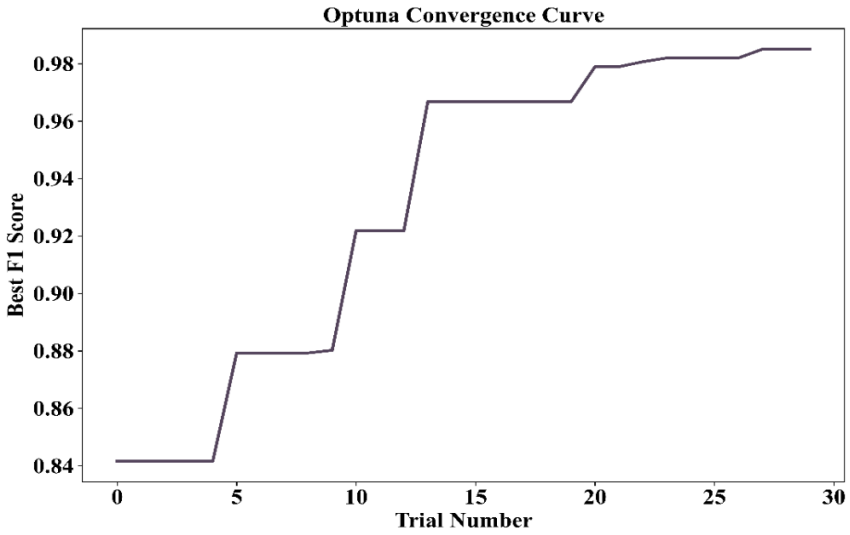


Fig. 11. Optuna convergence curve for LightGBM hyperparameter optimization

Figure 11 depicts the Optuna hyperparameter tuning convergence for 30 trials. It can be seen that the best F1 score progressively improves starting from 0.84 to begin with, experiencing significant rises at trials 5, 10, and 13, respectively, and stabilizing after 20 trials with a score greater than 0.98. It is evident that the Optuna TPE sampler works perfectly fine.

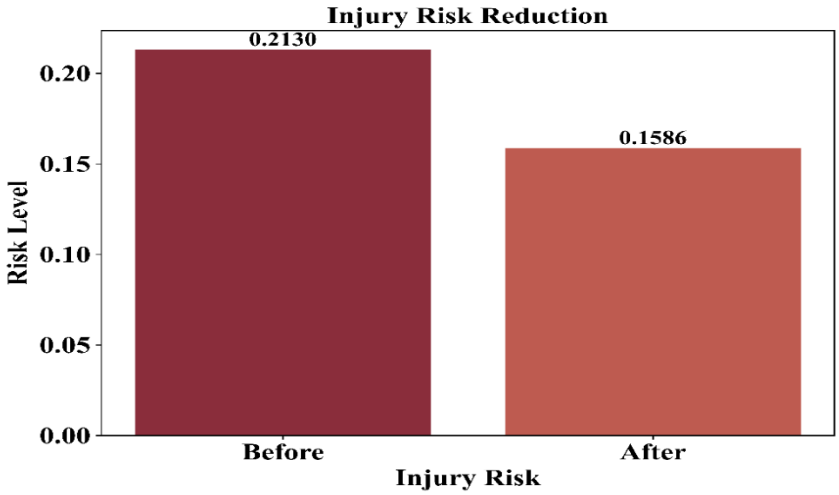


Fig. 12. Injury risk reduction before and after adaptive training intervention

Figure 12 depicts the effect of the suggested adaptive training system on the injury risk levels. The level of risk falls from 0.2130 before to 0.1586 after the system intervention, representing a 25.5% reduction in the risk level. This is an indication that the system works effectively in minimizing the injury risks of athletes.

4.3 Comparative Analysis

The comparative analysis involves comparison between the proposed models and traditional models in order to prove the superiority of the models in terms of performance. In performance prediction, XGBoost Regressor was compared with Linear Regression, Random Forest, and other models, where the model achieved a low mean absolute error and mean squared error and a high R^2, thereby performing better. The LightGBM Classifier was compared with Logistic Regression, SVM, and Decision Trees, among other classifiers, in injury risk classification, where it performed

better in terms of accuracy, precision, recall, and ROC-AUC. Optuna Optimization also boosted both models by tuning the hyperparameters to minimize overfitting.

Table 7. Comparative analysis of classification models for injury risk prediction

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest [35]	0.9728	0.9735	0.9721	0.972	0.9846
Support Vector Machine [36]	0.9689	0.9697	0.9682	0.968	0.9812
Proposed Model (LightGBM + Optuna)	0.9864	0.9864	0.9864	0.9864	0.9952

Table 7 displays the performance comparison between various classifiers used in predicting the risk of injury. From the table, it is evident that the proposed model with an Optuna-optimized LightGBM classifier gives the best performance with respect to all the measures such as accuracy (0.9864), precision, recall, F1 score, and ROC-AUC (0.9952), outperforming Random Forest and SVM classifiers.

Table 8. Comparative analysis of regression models for performance prediction

Model	R^2 Score	MAE	MSE	RMSE
Random Forest Regressor [37]	0.9485	0.0312	0.0023	0.048
Linear Regression [38]	0.9217	0.0398	0.0035	0.0592
Proposed Model (XGBoost + Optuna)	0.9721	0.0206	0.001	0.031

The comparison between the performances of different models is provided in Table 8 for predicting athlete performance. The XGBoost model with Optuna optimizer has been found to be the most effective among all other models due to the highest R^2 value of 0.9721 and minimum errors. It includes the lowest values of MAE, MSE, and RMSE.

Table 9. Comparative Analysis, after the paragraph introducing the baseline models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVM	0.9689	0.9697	0.9682	0.968	0.9812
Random Forest	0.9728	0.9735	0.9721	0.972	0.9846
ANN	0.9756	0.9762	0.975	0.9754	0.9873
CNN	0.9784	0.9789	0.9778	0.9782	0.9891
LSTM	0.9802	0.9807	0.9796	0.98	0.9904
GRU	0.9815	0.982	0.981	0.9813	0.9911
Proposed LightGBM + Optuna	0.9864	0.9864	0.9864	0.9864	0.9952

The following Table 9 presents a comparison between the proposed LightGBM–Optuna classifier and other conventional machine learning and deep learning classifiers including SVM, Random Forest, ANN, CNN, LSTM, and GRU. The proposed classifier outperformed other classifiers on all evaluation metrics where the values of accuracy, precision, recall, and F1-score were equal to 0.9864 while the value of ROC-AUC was equal to 0.9952. GRU gave the closest performance among the other baselines where the value of F1-score was 0.9813 while the value of ROC-AUC was 0.9911.

Table 10 below highlights the performance analysis of the various model setups used in performance prediction. From the table, it can be observed that the full model setup (PCA + Optuna + XGBoost) has a better prediction capability as indicated by the value of R^2 = 0.9721, MAE = 0.0206, MSE = 0.0010, and RMSE = 0.0310. However, when we compared our model to the one where no PCA was used, its RMSE is 0.0424. On the other hand, if we did not use Optuna optimization in our

approach, the RMSE would have been 0.0374. In short, we can conclude that PCA and Optuna optimization have enhanced the performance of the model. Table 11 represents the comparison among the different models in terms of their classification ability for injury risk prediction through LightGBM. In this regard, it is clear that the proposed model (with PCA and Optuna) yields the most optimal results with accuracy, precision, recall, and F1-score values of 0.9864 and an ROC-AUC of 0.9952. It proves that the inclusion of both PCA and Optuna improves classification ability substantially.

Table 10. Ablation study of proposed model with PCA and Optuna optimization

Model Configuration	R^2 Score	MAE	MSE	RMSE
Without PCA	0.9584	0.0275	0.0018	0.0424
Without Optuna	0.9642	0.0241	0.0014	0.0374
Full Proposed Model (PCA + Optuna + XGBoost)	0.9721	0.0206	0.0010	0.0310

Table 11. Ablation study of injury risk classification models with PCA and Optuna optimization

Model Configuration	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Without PCA	0.9724	0.9731	0.972	0.9718	0.9842
Without Optuna	0.9786	0.9792	0.9781	0.9775	0.9895
Full Proposed Model (pca+optuna+LightGBM)	0.9864	0.9864	0.9864	0.9864	0.9952

Table 12. Ablation analysis of PCA, SMOTE, class weighting, and Optuna for injury-risk

Configuration	PCA	SMOTE	Class Weighting	Optuna	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Baseline LightGBM	No	No	No	No	0.9521	0.9528	0.9514	0.9518	0.9715
PCA only	Yes	No	No	No	0.9586	0.9591	0.9578	0.9583	0.9784
PCA + SMOTE	Yes	Yes	No	No	0.9698	0.9705	0.9689	0.9696	0.9859
PCA + Class Weighting	Yes	No	Yes	No	0.9657	0.9664	0.9649	0.9655	0.9827
PCA + SMOTE + Class Weighting	Yes	Yes	Yes	No	0.9786	0.9792	0.9781	0.9775	0.9895
Full Proposed Model	Yes	Yes	Yes	Yes	0.9864	0.9864	0.9864	0.9864	0.9952

Table 12 provides the analysis for the ablation study of the proposed injury risk classification framework using LightGBM. The baseline LightGBM model had an F1-Score of 0.9518, which was raised to 0.9583 upon including PCA in the model. Inclusion of SMOTE raised the F1-Score to 0.9696, whereas use of class weighting gave an F1-Score of 0.9655 along with the inclusion of PCA. The inclusion of both SMOTE and class weighting further increased the F1-Score to 0.9775. Lastly, inclusion of Optuna resulted in the highest performing framework having an accuracy and F1-Score of 0.9864 along with an ROC-AUC of 0.9952.

4.4 Discussion

The interpretation of the results includes presenting the fact that the proposed models of XGBoost and LightGBM outperform baseline models in accuracy and error. The performance improvement is attributed to preprocessing, PCA-based feature selection, and hyperparameter optimization using Optuna. Comparing with other existing research, it can be concluded that the proposed approach performs better in terms of accuracy and efficiency, particularly on tabular data.

However, it can be seen that one of the insights from the study is that the combination of regression and classification in one algorithm gives complete surveillance. Some of the drawbacks include reliance on the quality of the data sets used, lack of real-time testing of the model, and inability to generalize into other sports domains.

5. Conclusions

The analysis demonstrates the effectiveness of the proposed AI-driven framework for improving athlete performance prediction and injury risk classification through the integration of XGBoost and LightGBM machine learning algorithms optimized using the Optuna hyperparameter optimization package. The framework provides a systematic approach for processing multimodal athlete-related data, extracting informative features, and developing predictive models capable of supporting data-driven training decisions. In particular, the optimized XGBoost model enhances the prediction of athlete performance by identifying complex relationships among physiological, training, and recovery-related variables, while the LightGBM model supports injury-risk classification by learning patterns associated with potential physical stress and adverse training conditions. The integration of feature preprocessing, dimensionality reduction, optimized machine learning, and model-specific hyperparameter tuning contributes to improved prediction precision, reduced estimation errors, and enhanced model generalizability across different athlete conditions. Beyond predictive modeling, the framework supports intelligent and continuous athlete monitoring by combining learning-based analysis with efficient feature extraction and real-time computational processing. Such integration enables performance trends and injury-related risk patterns to be identified more systematically, providing a basis for adaptive training-load adjustment. The framework can therefore assist coaches, sports scientists, and performance-monitoring systems in making timely training decisions based on quantitative evidence rather than relying solely on conventional assessment practices. In addition, the potential integration of energy-autonomous wearable technologies, in which ambient energy sources such as body heat, motion, light, or electromagnetic energy are converted into usable electrical power, can further support the development of sustainable and continuously operating athlete-monitoring platforms. From an engineering perspective, the proposed framework contributes an integrated architecture that connects wearable sensing, data processing, feature extraction, machine learning, and optimization for real-time predictive assessment. Its practical applicability lies in enabling personalized monitoring, early identification of injury-related risks, and adaptive adjustment of training loads. Overall, the framework establishes a scalable foundation for intelligent sports engineering applications by combining predictive analytics, optimized learning algorithms, and continuous monitoring within a unified decision-support environment.

5.1 Future Work

The future research could be done in terms of incorporating real-time streaming data from wearables used in IoT-based systems in order to support decision-making on the go. Moreover, the system could be further developed by introducing more sophisticated deep learning frameworks like transformers or a combination thereof. Adding psychological and environmental factors into consideration would help create a more comprehensive picture of the athlete's state.

Data Availability

https://www.kaggle.com/datasets/ziya07/iot-driven-sports-training-dataset
https://www.kaggle.com/datasets/anjalibhegam/multimodal-sports-injury-dataset

References

[1] Bisht K, Tejaswi DS, Manjulatha C, Das S, Gunupuru Y. Evolution of Wearable Technology in Sports and Fitness. In: Pathak J, Kumar A, Tripathi SL, Raj B, editors. Smart Textiles and Wearables for Health and Fitness. 1st ed. Wiley; 2025. p. 289-305. https://doi.org/10.1002/9781394302895.ch13

[2] Rebelo A, Martinho DV, Valente-dos-Santos J, Coelho-e-Silva MJ, Teixeira DS. From data to action: a scoping review of wearable technologies and biomechanical assessments informing injury prevention strategies in sport. BMC Sports Sci Med Rehabil. 2023;15(1):169. https://doi.org/10.1186/s13102-023-00783-4

[3] Yang X, Zhang S, Liu J, Gao Q, Dong S, Zhou C. Deep learning for smart fish farming: applications, opportunities and challenges. Rev Aquac. 2021;13(1):66-90. https://doi.org/10.1111/raq.12464
[4] Wackerhage H, Schoenfeld BJ. Personalized, Evidence-Informed Training Plans and Exercise Prescriptions for Performance, Fitness and Health. Sports Med. 2021;51(9):1805-1813. https://doi.org/10.1007/s40279-021-01495-w
[5] Reitzner SM, Brodin P. The Immune System as a Biosensor of Health and Athletic Performance. Scand J Med Sci Sports. 2026;36(3):e70258. https://doi.org/10.1111/sms.70258
[6] Verma A, Singhal N. Integrating Artificial Intelligence for Adaptive Decision-Making in Complex System. In: Das S, Saha S, Coello Coello CA, Bansal JC, editors. Advances in Data-Driven Computing and Intelligent Systems. Lecture Notes in Networks and Systems, vol. 892. Singapore: Springer Nature Singapore; 2024. p. 95-105. https://doi.org/10.1007/978-981-99-9521-9_8
[7] Shah M, Shah A, Patel K, Kshirsagar A, Sanghvi S, Sojitra V. Predictive analytics, strategic game analysis, and injury prevention in sports: the role of big data and artificial intelligence. Mach Learn Comput Sci Eng. 2025;1(1):17. https://doi.org/10.1007/s44379-025-00017-z
[8] Zhang Y, Tsai SB. Application of Adaptive Virtual Reality with AI-Enabled Techniques in Modern Sports Training. Mob Inf Syst. 2021;2021:1-10. https://doi.org/10.1155/2021/6067678
[9] Wu X, et al. Cloud-based deep learning-assisted system for diagnosis of sports injuries. J Cloud Comput. 2022;11(1):82. https://doi.org/10.1186/s13677-022-00355-w
[10] Tan X. Enhanced Sports Predictions: A Comprehensive Analysis of the Role and Performance of Predictive Analytics in the Sports Sector. Wirel Pers Commun. 2023;132(3):1613-1636. https://doi.org/10.1007/s11277-023-10585-z
[11] Nafea O, Abdul W, Muhammad G. Multi-sensor human activity recognition using CNN and GRU. Int J Multimed Inf Retr. 2022;11(2):135-147. https://doi.org/10.1007/s13735-022-00234-9
[12] Bouchekara HREH. Most Valuable Player Algorithm: a novel optimization algorithm inspired from sport. Oper Res. 2020;20(1):139-195. https://doi.org/10.1007/s12351-017-0320-y
[13] Highnam K, Puzio D, Luo S, Jennings NR. Real-Time Detection of Dictionary DGA Network Traffic Using Deep Learning. SN Comput Sci. 2021;2(2):110. https://doi.org/10.1007/s42979-021-00507-w
[14] Karlsson A, Wang T, Nowaczyk S, Pashami S, Asadi S. Mind the Data, Measuring the Performance Gap Between Tree Ensembles and Deep Learning on Tabular Data. In: Miliou I, Piatkowski N, Papapetrou P, editors. Advances in Intelligent Data Analysis XXII. Lecture Notes in Computer Science, vol. 14641. Cham: Springer Nature Switzerland; 2024. p. 65-76. https://doi.org/10.1007/978-3-031-58547-0_6
[15] Gao Y, Zhai P, Mosalam KM. Balanced semisupervised generative adversarial network for damage assessment from low-data imbalanced-class regime. Comput-Aided Civ Infrastruct Eng. 2021;36(9):1094-1113. https://doi.org/10.1111/mice.12741
[16] Meng L, Qiao E. Analysis and design of dual-feature fusion neural network for sports injury estimation model. Neural Comput Appl. 2023;35(20):14627-14639. https://doi.org/10.1007/s00521-021-06151-y
[17] Kamil NHM, Razali MNM, Norazman SH, Mansor MA, Jamaludin AS. AI-Driven Performance Metrics: Enhancing Athlete Monitoring and Injury Prevention. In: Hassan MHA, Sharir R, Syed Omar SF, Nunome H, Limroongreungrat W, editors. Proceedings of the 3rd International Conference on Innovation and Technology in Sports. Lecture Notes in Bioengineering. Singapore: Springer Nature Singapore; 2026. p. 503-511. https://doi.org/10.1007/978-981-95-0804-4_42
[18] Kadam N, Deshpande V. AI-Enhanced Biomechanics: Personalizing Physiotherapy for Optimal Recovery. In: Kumar A, Kumar TA, Ahuja S, Ananth JP, Manoj SO, editors. Physiotherapy Using Artificial Intelligence. 1st ed. Wiley; 2026. p. 379-398. https://doi.org/10.1002/9781394391561.ch18
[19] Spiteri BJ, Dingli A. AI in Sports: Shaping the Next Generation of Athletes. In: Dingli A, Musiolik TH, editors. Growing Up with AI. Studies in Computational Intelligence, vol. 1231. Cham: Springer Nature Switzerland; 2026. https://doi.org/10.1007/978-3-032-02260-8
[20] Li W, Feng J, Feng G, Feng S. Wearable technology for athletes: material innovations, performance monitoring, and emerging paradigms. Microchim Acta. 2026;193(2):76. https://doi.org/10.1007/s00604-025-07792-9
[21] Ipaktchi KR, Higinbotham SE. AI-Powered Rehabilitation and Monitoring Systems for Orthopedic Trauma Patients. In: De Simone B, Catena F, Gumbs AA, Kaafarani HMA, Kirkpatrick AW, editors. New Technologies, Precision Medicine, and AI in Emergency General Surgery. Hot Topics in Acute Care Surgery and Trauma. Cham: Springer Nature Switzerland; 2026. p. 303-311. https://doi.org/10.1007/978-3-032-09357-8_17
[22] Priyadarshini U, Jonjua M, Singh SK, Mishra P. Artificial Intelligence in Paralympic Sports: Advancing Training, Performance, and Media Representation. Macromol Symp. 2026;415(1):e70171. https://doi.org/10.1002/masy.70171
[23] Poongodi J, Kavitha K, Sathish S, Kumar RL. Hybrid AI-Driven Bio-Inspired Wearable Sensors With Aquasense AI Technology for Multimodal Health Monitoring and Rehabilitation in Dynamic

Environments. Trans Emerg Telecommun Technol. 2025;36(3):e70081. https://doi.org/10.1002/ett.70081
[24] Reddy KJ. Motor Rehabilitation and Biofeedback. In: Innovations in Neurocognitive Rehabilitation. Cham: Springer Nature Switzerland; 2025. p. 231-266. https://doi.org/10.1007/978-3-031-88117-6_11
[25] Shah J, Hudy A. Data Driven Injury Mitigation: Athlete Case Studies and Load Management Technologies. In: Raval MS, Kaya T, Artan NS, Taber C, editors. Sports Data Analytics. Studies in Big Data, vol. 184. Singapore: Springer Nature Singapore; 2026. p. 89-112. https://doi.org/10.1007/978-981-95-5132-3_5
[26] Anandh S, Parhate V. The Future of Physiotherapy: AI-Driven Biomechanical Rehabilitation Techniques. In: Kumar A, Kumar TA, Ahuja S, Ananth JP, Manoj SO, editors. Physiotherapy Using Artificial Intelligence. 1st ed. Wiley; 2026. p. 335-351. https://doi.org/10.1002/9781394391561.ch16
[27] Xi L, Li C, Anari MS, Rezaee K. Integrating wearable health devices with AI and edge computing for personalized rehabilitation. J Cloud Comput. 2025;14(1):64. https://doi.org/10.1186/s13677-025-00795-0
[28] Yadav A, Yadav K. Transforming healthcare and fitness with AI powered next-generation smart clothing. Discov Electrochem. 2025;2(1):2. https://doi.org/10.1007/s44373-025-00015-z
[29] Avram C, Taslauan A, Radu D, Astilean A. AI-Driven Posture Analysis and Training Management Application. In: Machado J, Trojanowska J, Soares F, Rea P, Butdee S, Gramescu B, editors. Innovations in Mechatronics Engineering IV. Lecture Notes in Mechanical Engineering. Cham: Springer Nature Switzerland; 2025. p. 92-104.
[30] Li J, Lu J. Research on Wearable Device Health Data Analysis and Personalized Training System Based on Artificial Intelligence. In: Jansen BJ, Ye J, Zhou Q, editors. Proceedings of the 4th International Conference on Cognitive Based Information Processing and Applications-Volume 1. Lecture Notes on Data Engineering and Communications Technologies, vol. 265. Singapore: Springer Nature Singapore; 2026. p. 27-34. https://doi.org/10.1007/978-981-95-2345-0_3
[31] Song J, et al. Artificial intelligence in orthopedics: fundamentals, current applications, and future perspectives. Mil Med Res. 2025;12(1):42. https://doi.org/10.1186/s40779-025-00633-z
[32] Di P, Hongye L, Donglin S. Autoregulated macro-microcycle training enhances physiological reserve and adaptability in competitive speed skaters. BMC Sports Sci Med Rehabil. 2025;18(1):47. https://doi.org/10.1186/s13102-025-01495-7
[33] Kuruwita A H, et al. A Hybrid Bidirectional Deep Learning Model Using HRV for Prediction of ICU Mortality Risk in TBI Patients. J Healthc Inform Res. 2025;9(4):629-655. https://doi.org/10.1007/s41666-025-00209-5
[34] Chaiya N, Tangkuaram T, Moonla C. Smart Skin and Beyond: Unlocking the Full Potential of Wearable Health Sensors. In: Mahato K, Chandra P, editors. Nano-bioelectronics for Precision Health Monitoring. Singapore: Springer Nature Singapore; 2025. p. 317-352. https://doi.org/10.1007/978-981-95-2117-3_11
[35] He K. Prediction Model of Juvenile Football Players' Sports Injury Based on Text Classification Technology of Machine Learning. Mob Inf Syst. 2021;2021:1-10. https://doi.org/10.1155/2021/2955215
[36] Cao Y, Li T. SVM action recognition model based on skeletal key point analysis with posture sensors to help sports training. BMC Sports Sci Med Rehabil. 2025;17(1):220. https://doi.org/10.1186/s13102-025-01260-w
[37] Zheng S, Man X. An Improved Logistic Regression Method for Assessing the Performance of Track and Field Sports. Comput Intell Neurosci. 2022;2022:1-10. https://doi.org/10.1155/2022/6341495
[38] Hacker S, Lenz C, Reichert L, Ringseis R, Zentgraf K, Kruger K. Vitamin D status and its determinants in German elite athletes. Eur J Appl Physiol. 2025;125(6):1549-1561. https://doi.org/10.1007/s00421-024-05699-6