Comparative Study of Approaches for Injury Risk Prediction in Athletes

Akanksh Rao S R
Arizona State University
Tempe
asanthoo@asu.edu

Anuj Abhay Joshi
Arizona State University
Tempe
ajosh104@asu.edu

Reshma Panibhate
Arizona State University
Tempe
rpanibha@asu.edu

Mansi Nandkar
Arizona State University
Tempe
mnandkar@asu.edu

Abstract

Athlete injuries pose significant physical and psychological challenges, impacting both team performance and individual careers. With the increasing availability of data from wearable sensors, IoT devices, and physiological monitoring systems, a variety of data-driven approaches have emerged to predict injuries and improve athlete safety. This study presents a comparative analysis of established machine learning (ML) and deep learning (DL) techniques for predicting injury risk in athletes. It focuses on implementing and evaluating these models to assess their performance, interpretability, and applicability in real-world sports scenarios, with particular emphasis on prediction accuracy and response time. By benchmarking these techniques, we aim to highlight their strengths, limitations, and practical relevance, offering insights for future research and deployments in sports analytics.

***K*eywords** : Injury prediction · Machine learning · Deep learning · Sport analytics · Wearable sensors

1 Introduction

In the rapidly evolving field of sports analytics, injury prevention has become a critical focus, driven by the increasing availability of real-time performance data from wearable sensors and IoT devices [6, 11]. These technological advancements provide the foundation for applying AI-driven methods to predict and mitigate injury risks in athletes. Traditional injury prediction models typically rely on static datasets [9], which often overlook the temporal dependencies that are crucial for identifying injury precursors [2]. In contrast, recent advancements in ML and DL based techniques have enabled the modeling of these sequential dependencies, offering new opportunities for injury risk assessment [12]. This comparative study explores both traditional statistical models and modern sequential architectures to predict injury risks more effectively. By establishing a comprehensive framework that balances prediction accuracy, computational efficiency, and real-world applicability, this study aims to provide actionable insights for coaches and sports medicine professionals, contributing to the broader goal of enhancing athlete safety and performance optimization [1].

2 Motivation

Injury prevention plays a crucial role in maintaining athlete performance and longevity. Athletes are constantly exposed to the risk of both acute and chronic injuries, which can significantly impact their careers and the overall success of their teams [5]. Traditional injury prediction methods, such as those based on biomechanical evaluations or historical injury data, are limited in their ability to accurately predict injuries in real time or consider the dynamic nature of athletic performance [2]. Recent advancements in wearable technology and sensor systems have generated a wealth of real-time data, offering an opportunity to incorporate ML and DL based techniques into injury risk prediction models [11, 13]. These data-driven approaches can account for complex patterns in the athlete's physiological responses, training loads, and even environmental factors, which were previously difficult to capture with traditional methods [3]. These methods can provide not only more accurate predictions but also deeper insights into the underlying factors contributing to injury risks, enhancing the safety and effectiveness of sports training programs [7]. This motivation drives the exploration of these advanced techniques, aiming to bridge the gap between theoretical models and their practical application in the real-world sports context.

3 Literature Review

Recent studies have underscored the potential of Artificial Intelligence (AI) in transforming injury prediction paradigms in sports by integrating data from wearable sensors, medical imaging, and biomechanical analyses [4]. For instance, a comprehensive review on diagnostic applications of AI highlighted the role of deep learning in processing complex imaging and sensor data to detect injury patterns early [6, 13]. Similarly, a scoping review of machine learning (ML) approaches in sports emphasized the variability in prediction performance and the need for standardized evaluation metrics [2]. Research on youth soccer players has shown promising results with ML models that incorporate physiological and biomechanical variables, though challenges such as data quality and model generalizability persist [7]. Further, IoT-based systems using Recurrent Neural Networks (RNNs) have been explored for real-time injury prediction in martial arts, emphasizing the importance of temporal feature analysis [11, 12]. Additionally, investigations have compared edge wearable device data with conventional methods, demonstrating that deep learning architectures such as Convolutional Neural Networks (CNNs) and RNNs can significantly enhance predictive performance and safety [3, 8]. Other notable works have employed hybrid approaches to fuse data-driven analytics with expert domain knowledge, paving the way for more robust and interpretable models [5, 9]. Collectively, these studies establish a foundation for the comparative analysis proposed in this work, while also highlighting existing gaps in interpretability and scalability.

4 Problem Statement

Injury risk prediction in sports is a rapidly growing field, yet it faces significant challenges related to data quality, model generalizability, and interpretability. Reviews have revealed that many machine learning models for injury risk prediction suffer from inconsistent evaluation methods and limited adaptability across diverse sports environments [4]. However, the integration of domain-specific knowledge with data-driven models remains underexplored, leaving a critical gap in the development of robust, interpretable, and scalable solutions [1, 10]. This project aims to address these limitations by conducting a comparative analysis of ML and DL approaches for injury risk prediction, focusing on enhancing prediction performance, model transparency, and practical applicability in both resource-rich and resource-constrained settings.

5 Proposed Methodology

The study follows a systematic approach, comprising data collection, preprocessing, model development, evaluation, and validation. The methodology is outlined as follows:

5.1 Data Pipeline

Comprehensive sports performance and physiological data will be gathered from established public databases and partnering sports organizations. This data will then be meticulously cleaned and organized, with key performance indicators identified and enhanced, and measurements standardized across different sources. This careful preparation will ensure that coaches and trainers receive consistent, reliable information regardless of the data's origin.

5.2 Exploratory Data Analysis (EDA)

Through detailed statistical analysis and visual data mapping, we will identifies critical performance indicators (KPIs) that show strong connections to potential injury risks through statistical analysis and visualization. These findings are thoroughly reviewed and validated by experienced sports medicine professionals and trainers to ensure practical relevance in order refine feature selection.

5.3 Model Development

Implement ML models such as logistic regression, support vector machines (SVM), and random forests to evaluate baseline performance. In addition, we will develop deep learning (DL) models using recurrent neural networks (RNN) and long-short-term memory (LSTM) to capture temporal dependencies in the data. Transformers might also be used for advanced sequential injury risk assessment, eenhancingpredictive accuracy.

5.4 Comparative Study

Models will be evaluated based on accuracy, precision, recall, and F1-score to measure their predictive performance. Interpretability and computational efficiency will also be assessed to ensure practical deployment feasibility. This approach will help identify the most effective and applicable models for real-world use.

5.5 Validation and Reporting

The findings will be validated using historical athlete data and prospective simulated scenarios to ensure robustness. Continuous feedback from professionals will be incorporated to refine model accuracy and relevance which will be thoroughly documented in a detailed technical report.

6 Current Progress

6.1 Dataset Selection and Analysis

We selected the MHEALTH (Mobile Health) dataset due to its particular suitability for injury risk prediction in athletes. It is the Rich Multisensor Time-Series Data that explores real-world movement patterns through sensors placed on athletes' chest, wrist, and ankle, capturing vital information about how the body moves and responds during sports activities. Through detailed activity-based labeling, it has documented common sports movements like walking, running, cycling, jumping, and crouching, making it easier to understand athletic performance. The sequential nature of data helps trainers and athletes track continuous movement patterns over time, revealing how one motion flows into another. The realistic simulation of wearable data comes from standard sports monitoring devices, recording 50 measurements every second, just like the equipment used in professional training. As an open-Access and well-documented resource in the UCI Machine Learning Repository, this information is freely available to coaches, athletes, and researchers, helping advance our understanding of sports performance and injury prevention.

6.1.1 Dataset Characteristics

The MHEALTH dataset consists of recordings from 10 volunteers (8 male, 2 female) aged 20-35 who performed a series of physical activities while wearing sensors. It includes 12 defined activity labels ranging from stationary activities (standing, sitting, lying down) to dynamic movements (walking, running, jumping), making it highly relevant for athletic performance analysis.The data includes:

Sensor Modalities: Accelerometer, gyroscope, and magnetometer at the ankle and wrist; accelerometer and ECG at the chest.
Sampling Frequency: 50 Hz (50 samples per second)
Features: 24 sensor measurements plus activity labels and subject IDs.

6.2 Data Preprocessing

Our data preprocessing pipeline consists of several key stages designed to prepare the MHEALTH dataset for injury risk modeling:

6.2.1 Validation of Proxy Injury Risk Labels

Given the absence of true injury annotations in the dataset, we developed a proxy labeling strategy based on biomechanical and physiological indicators of injury risk. These proxy labels simulate real-world risk scenarios and allow us to frame our project as a supervised learning task.

6.2.2 Segmentation into Time-Series Windows

To support training of sequential models, we segmented the time-series data into overlapping windows. We divided continuous movement data into manageable time segments, focusing on overlapping 2-second windows that capture 100 distinct measurements. Each window provides a detailed snapshot of the athlete's movement patterns, with specific risk labels assigned based on either the most frequent or final movement in that segment. By organizing the data into these structured windows (categorized by samples, timesteps, and movement features), coaches and trainers can better understand how movement patterns evolve and potentially indicate injury risks.

Proxy Indicator	Rationale
High Impact Acceleration	Detected using sudden spikes in total chest acceleration (>3.5g); may reflect unsafe landings, falls, or jerky movement
Fatigue Signals	Based on unusually high heart rate (ECG) during low-intensity activities; suggests poor recovery or cardiovascular strain
Repetitive Stress	Extended duration of high-load activities with limited rest; simulates overuse injury risk
Postural Instability	Captures unstable body movement during transitions using gyroscope data; often a precursor to ligament injuries

Table 1: Proxy indicators used for injury risk labeling

6.3 Model Implementation and Evaluation

Our approach involves implementing and comparing both traditional machine learning models and sequential deep learning architectures.

6.3.1 Traditional Models (Baselines)

Random Forest: Implemented for baseline performance using aggregated features from each window.
Logistic Regression: Provides interpretable linear decision boundaries.
Support Vector Machines: Tests the effectiveness of margin-based classification for risk detection.

6.3.2 Sequential Deep Learning Models

LSTM Networks: Primary sequential model to capture temporal dependencies in movement patterns.
1D Convolutional Neural Networks: Explored for detecting local motion patterns that may indicate injury risk.

6.4 Preliminary Results

The early testing using movement pattern analysis shows promising results in identifying the difference between safe and risky athletic movements. Although basic assessment methods achieved 65-75% accuracy in identifying risk patterns, tracking movements over time proved more effective, reaching 78-82% accuracy. Our analysis highlighted that sudden changes in movement speed (acceleration) and heart activity (ECG data) were the most reliable warning signs of potential injuries.

6.5 Adjustments to Plan and Risk Mitigation

Based on our progress and preliminary findings, we have made several adjustments to our original plan:

Enhanced Validation Strategy: To address the limitation of proxy labels, we've implemented cross-validation with stratification to ensure robust model evaluation.
Class Imbalance Management: Given the predominance of "no risk" segments in the dataset, we've incorporated balanced sampling techniques and adjusted evaluation metrics to focus on recall for risk categories.
Interpretability Focus: We've placed additional emphasis on model interpretability through visualization tools and feature importance analysis to build trust in the predictions.

7 Updated Timeline

Phase	Activities	Timeline
Data Preparation	Dataset selection, preprocessing, proxy label refinement	Completed
Model Development	Implementation of traditional and sequential models	In Progress
Comparative Analysis	Systematic evaluation and comparison of model performance	Weeks 6-8
Visualization	Development of interpretability tools and visualizations	Weeks 7-9
Documentation	Final report preparation and presentation materials	Weeks 9-10

Table 2: Updated project timeline

8 Conclusion

The study highlights the critical importance of utilizing advanced machine learning (ML) and deep learning (DL) techniques for predicting injury risks in athletes. By comparing various models, we aim to identify those that not only excel in predictive accuracy but also offer interpretability and computational efficiency, which are essential for real-world application. The integration of data from wearable sensors and IoT devices enhances our understanding of injury dynamics, paving the way for improved athlete safety. Furthermore, the validation of these models through historical data and simulated scenarios ensures their robustness and relevance in practical settings . As we benchmark these techniques, we also acknowledge the existing gaps in interpretability and scalability, which future research must address. Ultimately, our findings will contribute valuable insights to the field of sports analytics, guiding future developments in injury prevention strategies.

References

[1] Sheng Chen, Liya Guo, Rui Xiao, Jingfa Ran, Haidan Li, and Lino C. Reynoso. Establishing a cognitive evaluation model for injury risk assessment in athletes using rbf neural networks. *Soft Computing*, 27(17):12637–12652, 2023.

[2] Omar Farghaly and Priya Deshpande. Leveraging machine learning to predict national basketball association player injuries. In *2024 IEEE International Workshop on Sport, Technology and Research (STAR)*, pages 216–221, 2024.

[3] Mohd. Asif Gandhi, Surekha Khetree, N L Mishra, Arul Mary Rexy V, Z. Justin, and Harshal Patil. Sports risk prediction and evaluation model based on single layer feed forward neural network. In *2024 3rd International Conference for Innovation in Technology (INOCON)*, pages 1–6, 2024.

[4] Christopher Leckey, Nicol van Dyk, Cailbhe Doherty, Aonghus Lawlor, and Eamonn Delahunt. Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis. *British Journal of Sports Medicine*, 59(7):491–500, 2025.

[5] Nilamadhab Mishra, Beau Gray M. Habal, Precious S. Garcia, and Manuel B. Garcia. Harnessing an ai-driven analytics model to optimize training and treatment in physical education for sports injury prevention. In *Proceedings of the 2024 8th International Conference on Education and Multimedia Technology*, ICEMT '24, page 309–315, New York, NY, USA, 2024. Association for Computing Machinery.

[6] Carmina Liana Musat, Claudiu Mereuta, Aurel Nechita, Dana Tutunaru, Andreea Elena Voipan, Daniel Voipan, Elena Mereuta, Tudor Vladimir Gurau, Gabriela Gurău, and Luiza Camelia Nechita. Diagnostic applications of ai in sports: A comprehensive review of injury risk prediction methods. *Diagnostics*, 14(22), 2024.

[7] Francisco Javier Robles-Palazón, José M. Puerta-Callejón, José A. Gámez, Mark De Ste Croix, Antonio Cejudo, Fernando Santonja, Pilar Sainz de Baranda, and Francisco Ayala. Predicting injury risk using machine learning in male youth soccer players. *Chaos, Solitons Fractals*, 167:113079, 2023.

[8] Mohammad Mohsen Sadr, Mohsen Khani, and Saeb Morady Tootkaleh. Predicting athletic injuries with deep learning: Evaluating cnns and rnns for enhanced performance and safety. *Biomedical Signal Processing and Control*, 105:107692, 2025.

[9] Tushar Dhar Shukla, Divya Nimma, Kiran Sree Pokkuluri, Syed Najmusaqib, K.K. Sivakumar, and B Kiran Bala. Utilizing artificial intelligence for enhancing performance and preventing injuries in sports analytics. In *2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT)*, pages 1–6, 2024.

[10] Mengli Wei, Yaping Zhong, Huixian Gui, Yiwen Zhou, Yeming Guan, and Shaohua Yu. Sports injury prediction model based on machine learning. *Chinese Journal of Tissue Engineering Research*, 29(2):409–418, 2025.

[11] Shuning Xu, Xiao Zhang, and Ning Jin. Soccer sports injury risk analysis and prediction by edge wearable devices and machine learning. In *2023 International Conference on Artificial Intelligence of Things and Systems (AIoTSys)*, pages 38–43, 2023.

[12] H. Yao. An iot-based injury prediction and sports rehabilitation for martial art students in colleges using rnn model. *Mobile Networks and Applications*, 2024. Published: 07 September 2024, Accepted: 19 August 2024.

[13] Xiaoyu Ye, Yifan Huang, Zhen Bai, and Yucheng Wang. A novel approach for sports injury risk prediction: based on time-series image encoding and deep learning. *Frontiers in Physiology*, 14:1174525, 2023. Erratum in: Front Physiol. 2024 Jul 22;15:1441107. doi: 10.3389/fphys.2024.1441107.