OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

A Data-Driven Machine Learning Framework For Detecting Fatigue-Related Injury Risk Using Physiological And Motion Metrics

Vaana Gera
vaanagera16@gmail.com

ABSTRACT

Injury prevention in sport requires early identification of physiological fatigue and maladaptation before performance decline or injury occurs. However, traditional monitoring approaches often rely on isolated indicators, limiting their ability to capture the multifactorial nature of injury risk. This study aimed to investigate the factors associated with elevated injury risk in athletes.

For this study, data from the publicly available Real Time Physiological Monitoring in Sports dataset (Kaggle) was used. (colabsss, 2025). The dataset contained 23,400 observations from 234 athletes across five sport disciplines. To ensure independence of observations and reduce within-athlete information leakage, a controlled athlete-level sampling approach selected one representative observation per athlete, resulting in a dataset of 234 samples. A LightGBM classification framework was trained using physiological and motion based metrics.

The model obtained ROC-AUC of 0.705, recall of 0.52, accuracy of 0.64, precision of 0.67, and F1 score of 0.59. The biggest predictors of injury risk, according to SHAP analysis, were age, heart rate, motion-based variables and oxygen saturation. Severe departures from normal physiological patterns were linked to increased injury risk.

These findings suggest that combining physiological and motion-based data can help detect early signs of fatigue and increased injury risk. Although challenges such as personalization and defining optimal thresholds remain, wearable-driven analytics shows promise for supporting better training and injury prevention decisions.

Keywords: Health monitoring, Wearable technology, Injury risk, Fatigue detection, Sports analytics, Cardiovascular response, Feature importance, Data driven analysis

July 2026
Vol 9. No 3.

INTRODUCTION

Physical activity and participation in sport are fundamental components of a healthy lifestyle across the lifespan and are strongly advocated by healthcare professionals. Despite its well-documented benefits, sport participation is associated with a substantial injury burden, highlighting the need for rigorous research into injury prevention strategies across all age groups and sporting disciplines.

Youth and young adults demonstrate the highest participation rates in organized sport, yet they also experience the highest injury rates. Sport is the leading cause of injury in youth populations. Over 3.5 million children under the age of 14 get hurt while playing sports or engaging in recreational activities every year. Furthermore, overuse injuries alone account for approximately 50% of all juvenile sports injuries, highlighting the urgent need for efficient monitoring and early identification of physical stress and exhaustion in players (StriveOn, 2025).

Lower extremity injuries (see Appendix A) dominate across nearly every sport, with ankle and knee joint injuries comprising nearly 50% of all acute injuries. Beyond immediate physical consequences, sports injuries may also contribute to long-term public health concerns. Approximately 8% of youth discontinue sport participation annually due to injury or fear of injury, leading to reduced physical activity levels and increased risk of obesity and post-traumatic osteoarthritis (StriveOn, 2025).

Beyond the physical impact, sports injuries also impose a substantial economic burden on healthcare systems and societies worldwide. Estimates suggest that the global economic impact of sports injuries exceeds $300 billion annually, including healthcare expenses, rehabilitation costs, and productivity losses (Chen, 2026). These significant financial implications further highlight the importance of implementing evidence-based injury prevention strategies to reduce injury risk and associated costs (see Appendix A).
Given the extensive physical, psychological, and economic implications, reducing the burden of sports injuries has the potential to significantly improve quality of life and promote sustained engagement in physical activity. Early detection of fatigue and maladaptation plays a crucial role in injury prevention, optimized recovery, and long-term athlete health.

A widely used framework in sports injury prevention is the four-step model by Willem van Mechelen (1992), which provides a structured method to understand, prevent, and evaluate sports injuries (Mechelen, 1992). Researchers assess prevention strategies using designs such as randomized controlled trials (RCTs), cohort, case–control, and quasi-experimental studies, although practical and ethical constraints often limit the use of RCTs and may introduce bias.

Despite progress, implementing evidence-based interventions in real sports settings remains challenging. Caroline Finch addressed this through the Translating Research into Injury Prevention Practice (TRIPP) Framework, which extends the original model by incorporating real-world context and implementation (Finch, 2006). Program feasibility depends on factors such as participant age, level of play, sport type, and organizational structure, as highlighted by Carolyn Emery and Kati Pasanen (2019).

Despite these advances, existing research has important limitations. Many studies examine physiological and biomechanical indicators of injury risk in isolation, without considering how these factors interact. Although wearable technologies enable the collection of both data types, their combined role in understanding injury risk remains underexplored. This study addresses this gap by applying machine learning to jointly analyse physiological and motion-based metrics, enabling a more comprehensive examination of the factors associated with elevated injury risk in athletes.

PHYSIOLOGICAL AND MOTION-BASED METRICS FOR ATHLETE MONITORING

Physiological and motion-based metrics (see Appendix A) form the foundation of modern athlete monitoring systems, providing coaches and sports scientists with objective, actionable data to optimize training loads, prevent injury, and enhance performance.

The primary metrics used in this field include:

Heart Rate Metrics

Heart rate (HR) remains one of the most widely used indicators of cardiovascular stress and recovery (see Appendix A). Measures such as resting heart rate, maximum heart rate, and heart rate variability (HRV) provide insights into autonomic nervous system balance, training adaptation, and early signs of fatigue or overtraining. In the general adolescent population, normal resting HR typically ranges from 60–100 beats per minute, whereas well-trained endurance athletes often exhibit lower resting values between 40–60 beats per minute, and in some elite cases as low as 30–40 beats per minute due to increased cardiac efficiency (Cleveland Clinic, 2025).
During exercise, heart rate increases significantly to meet the body's higher oxygen demand, often reaching 50–85% of an individual's maximum heart rate depending on the intensity of activity (American Heart Association, 2024). Rather than absolute values, a sustained elevation or reduction relative to an individual's baseline is considered abnormal and may indicate physiological stress or inadequate recovery.

Oxygen Saturation (SpO_2)

Oxygen saturation (SpO_2) reflects the efficiency of oxygen transport within the blood (see Appendix A). Although less commonly used in routine training environments, SpO_2 monitoring is particularly relevant for athletes training at altitude or those with underlying respiratory or cardiovascular considerations. Normal SpO_2 values at sea level typically range from 95–100%, while persistent readings below 90–92% are considered abnormal and may indicate compromised oxygen delivery or increased physiological strain (Holland, 2023).

Muscle Activity

Muscle activity, commonly measured using electromyography (EMG), captures the electrical signals generated during muscular contraction. These measurements provide insights into muscular load, neuromuscular coordination, and the onset of fatigue during sport-specific movements. In research contexts, EMG activity is commonly normalized relative to maximal voluntary contraction (MVIC) to allow comparison across individuals and tasks. Based on EMG analyses, muscle activation is typically classified into different intensity levels: low activation (<20% MVIC), moderate activation (21–40% MVIC), high activation (41–60% MVIC), and very high activation (>60% MVIC). These classifications help researchers evaluate the mechanical demands of different movements and understand how muscular effort varies across sport-specific activities (Edwards, Ebert, Littlewood, Ackland, & Wang, 2017).

Motion-Based Data

Motion data (see Appendix A) collected via accelerometers, gyroscopes, and magnetometers across three axes (X, Y, and Z) enables detailed biomechanical analysis. In this dataset, the X-axis represents forward–backward or lateral movement, the Y-axis corresponds to vertical movement such as jumping or ground impact, and the Z-axis reflects rotational or side-to-side motion depending on sensor orientation. Motion values close to 0 on the X and Y axes, approximately −0.66 to 0.66, indicate low-intensity or stable movement, while values between 0.66 and 1.5 represent moderate movement intensity. Values exceeding ±1.5 suggest high-intensity or unstable motion patterns. For the Z-axis, which reflects gravitational orientation and rotational movement, low values occur below 9.1, moderate movement ranges between 9.1 and 10.5, and values above 10.5 indicate higher mechanical loading or abrupt movement changes. Analysing movement across these axes allows quantification of movement volume, impact forces, technique variability, and cumulative mechanical stress.

Training Intensity

Training intensity refers to the level of physical effort required during exercise and can be evaluated using physiological indicators such as heart rate and subjective measures like the Rating of Perceived Exertion (RPE). Exercise intensity is generally categorized into low, moderate, and vigorous levels. Low-intensity activities involve minimal increases in heart rate and breathing, such as warm-ups, cool-downs, or light movement. Moderate intensity leads to noticeable increases in heart rate and breathing while still allowing conversation, typically during steady aerobic activities. Vigorous or high intensity significantly elevates heart rate and breathing and occurs during demanding efforts such as sprint intervals, high-speed running, or competitive activity (Mayo Clinic, n.d.).

METHODOLOGY

Study Design and Data Source

This study aims to develop a machine-learning–based model for identifying injury risk in athletes using physiological, biomechanical, and demographic indicators. The dataset employed in this research was obtained from Kaggle, titled “Real-Time Physiological Monitoring in Sports” (colabsss, 2025). It contains sensor-derived physiological and performance metrics collected from athletes across multiple sports disciplines.

The dataset used in this study is simulated, consisting of synthetically generated physiological and motion data rather than real-world observations from athletes. Based on the dataset's documentation, the records have been created to mimic the athletes' training data that includes sensor readings, the measures of the level of training and also the risks of injuries. But there is no information about the process of creating this data from this documentation. Consequently, it is unclear whether the data were produced through perturbation of real athlete records, simulation-based approaches, generative modelling techniques, or a combination of methods.

The creation of a synthetic dataset has the benefit of being an effective tool for testing methodologies and developing models because it creates a controlled scenario whereby machine learning techniques can be tested while avoiding any ethical issues that come with handling human data. It is, however, worth noting that since the data used are not based on any actual athletes, the results cannot completely generalize to the real-world. The results should thus be taken as indications only and further research will need to test the model using actual athlete data.

The dataset includes repeated observations for each athlete, enabling analysis of both inter-individual and intra-individual variability in physiological patterns associated with injury risk. All data used in this study were archival in nature and accessed directly from the publicly available Kaggle repository.

Data Description and Structure

Each observation in the dataset corresponds to a real-time sample collected from an athlete, accompanied by demographic fields including Age, Gender, and Sport category. The data utilized in this study consists of a comprehensive collection of physiological and biomechanical measurements collected from a cohort of athletes. The dataset comprises 23400 total observations. These observations were obtained from 234 unique participants, with each athlete contributing exactly 100 temporally indexed measurements.

Participant Demographics

The study cohort consisted of N = 234 athletes with a mean age of 23.53 years. The gender distribution was near parity, with females comprising 52.14% (n=122) of the cohort and males comprising 47.86% (n=112). Participants were drawn from five distinct sport disciplines. The distribution across these specializations is detailed in Table 1. The largest representation was from Track athletes, while Swimming had the lowest number of participants.

Table 1: Distribution of Unique Participants by Sport Specialization

Sport Discipline	**Number of Participants (n)**	**Proportion (%)**
Track	54	23.1
Soccer	51	21.8
Tennis	50	21.4
Basketball	41	17.5
Swimming	38	16.2
Total	**234**	**100.0%**

The athletes are anonymized using a unique Athlete_ID, which additionally allows grouping of records originating from the same individual. This structure enables modelling that avoids within-athlete leakage while capturing repeated-measures variability.

Measures and Variables

The dataset includes a mixture of physiological, biomechanical, and training-related variables. Physiological metrics consist of heart rate, oxygen saturation, and muscle activity, serving as indicators of cardiovascular strain, oxygen delivery efficiency, and neuromuscular engagement, respectively.
Biomechanical workload was represented through Motion_X, Motion_Y, and Motion_Z, which reflect tri-axial accelerometer data capturing displacement and movement intensity.
Training load was encoded through Training_Intensity (Low, Medium, High).
The outcome variable, Injury_Risk, is a binary classification representing whether an athlete was at elevated risk of injury or not at the time of the recorded measurement.

Data Pre-processing and Feature Handling

Data pre-processing was performed using Python, primarily with the pandas and NumPy libraries. The target variable, Injury_Risk, was already encoded as a binary indicator (0 = no injury risk, 1 = injury risk) and therefore required no label transformation. All records contained valid target values, ensuring a well-defined supervised learning setup.

The dataset contained multiple observations per athlete, which could introduce bias if all records were used directly for training the model. To address this, a controlled athlete-level random sampling strategy was applied. First, athletes were grouped by Athlete_ID, and the total number of injury-risk occurrences was computed for each athlete. Athletes were then ranked in descending order based on this cumulative injury-risk count.

From this ranking, the 117 athletes with the highest cumulative injury risk were identified. For each of these athletes, one observation with Injury_Risk = 1 was randomly selected. Similarly, the 117 athletes with the lowest cumulative injury risk were identified, and one observation with Injury_Risk = 0 was randomly sampled for each athlete. This resulted in a dataset comprising one observation per athlete, with a balanced distribution between injury and non-injury cases.

After sampling, the selected observations were randomly shuffled to remove ordering effects. The final dataset therefore contained 234 athlete-level samples, ensuring independence between observations while preserving meaningful injury-risk signals.

For model training, Athlete_ID and Injury_Risk were excluded from the feature matrix to prevent information leakage. All remaining predictor variables were retained. Categorical features, identified by an object data type (including Sport, Gender, and Training_Intensity), were explicitly converted to categorical format to enable LightGBM's native handling of categorical splits.

The processed dataset was then divided into training and validation subsets. Model predictions were generated as probabilities and subsequently thresholded at 0.5 to obtain final binary injury-risk classifications.

Justification of the Sampling Strategy

To achieve statistical independence between observations, this athlete level sampling strategy was used. Initially, the dataset contained 100 observations per athlete (234 athletes) and if the model were trained on these 23,400 observations, it would learn athlete-specific patterns and not general injury risk characteristics. Many times, the records from the same individual share physiological and training related similarities which could artificially inflate predictive performance. Therefore, to ensure that the analysis was conducted at the athlete level rather than an observation-level, using only one observation per athlete was suitable.

Another reason for this sampling strategy was the class imbalance of the original dataset, with the non-injury cases being more prevalent than the injury cases. So, apart from ensuring independence, this sampling strategy also helped create a balanced distribution between both the outcomes and also reduced potential bias during model training.

A limitation of this approach is that the final dataset consisted of 234 samples, which could reduce statistical power and limit the generalizability of the findings to broader athletic populations. However, cost, complexity and time required for data collection are often constraints to athlete-level datasets in sports science and other biological research domains which makes very large cohorts difficult to obtain. The sampling strategy was therefore adopted as a trade-off between maximizing sample size and maintaining independence between observations. Consequently, the findings should be interpreted as exploratory, and future studies using larger athlete cohorts and subject-aware validation techniques could further strengthen the robustness and generalizability of the results.

Model Training and Validation

After constructing the athlete-level sampled dataset, the data were randomly shuffled and split into training and validation sets using an 80/20 ratio. This split provided sufficient data for model training while reserving an independent subset for performance evaluation. To allow reproducibility of the data partitioning process, a random seed was fixed (random_state =42).

Due to LightGBM's native handling of categorical variables, the identified categorical variables were retained in their format. The model was then trained using a binary classification objective and evaluated using the Area Under Receiver Operating Characteristic curve (ROC-AUC). A learning rate of 0.05 and 31 leaves per tree were the key hyperparameters. Training could continue for a maximum of 1000 boosting rounds, with early stopping in case the validation performance failed to improve for 100 rounds consecutively. The validation dataset was supplied during training to monitor the model's performance and prevent overfitting.

Finally, the predicted probabilities generated by the model were converted to binary injury-risk classifications using a standard threshold of 0.5.

Model Choice and Rationale

A Gradient Boosting Decision Tree approach was selected via LightGBM because of its suitability for heterogeneous tabular data and its high efficiency when handling complex nonlinear relationships. LightGBM was further chosen due to:

- its ability to natively process categorical variables without one-hot expansion,
- its robustness to noisy physiological data.

Given the multifactorial nature of injury risk, such capabilities are particularly valuable.
Although the primary aim of this study is to examine factors associated with elevated injury risk rather than to develop a predictive system, *LightGBM is being used intentionally for exploration because it captures relationships that simpler models cannot*. Model interpretability techniques, including feature importance and SHAP analysis, were then used to identify and analyze the relative contribution of each factor.

Evaluation Metrics

Model performance was assessed through several complementary metrics:

- ROC-AUC (see Appendix A) served as the primary indicator of discriminatory ability.
- ROC curve visualization (see Appendix A) illustrated the sensitivity–specificity trade-off across thresholds.
- Classification report (precision, recall, F1-score) at a nominal 0.5 threshold provided insight into class-specific performance (see Appendix A).
- Confusion matrix (see Appendix A) quantified false positives and false negatives, which are critical in injury prediction contexts where missed injury risks are especially consequential.

Together, these metrics offered a holistic view of model performance, balancing predictive accuracy with practical interpretability in an athletic training environment.

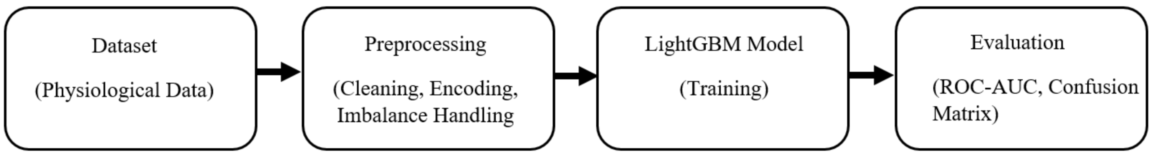
Software & Computational Tools

All analyses were conducted using:

- Python 3.10
- pandas and NumPy for data manipulation
- LightGBM for modelling
- Matplotlib for visualization

The computational workflow was fully reproducible through scripted execution, ensuring transparency and reliability of results.

Figure 1: Model Development Pipeline



RESULTS

Table 2: Summary of Model Performance Metrics

Metric	*Value*
Accuracy	63.8 %
Precision	66.67%
Recall	52.2%
F1-Score	58.5%
ROC- AUC	70.5%

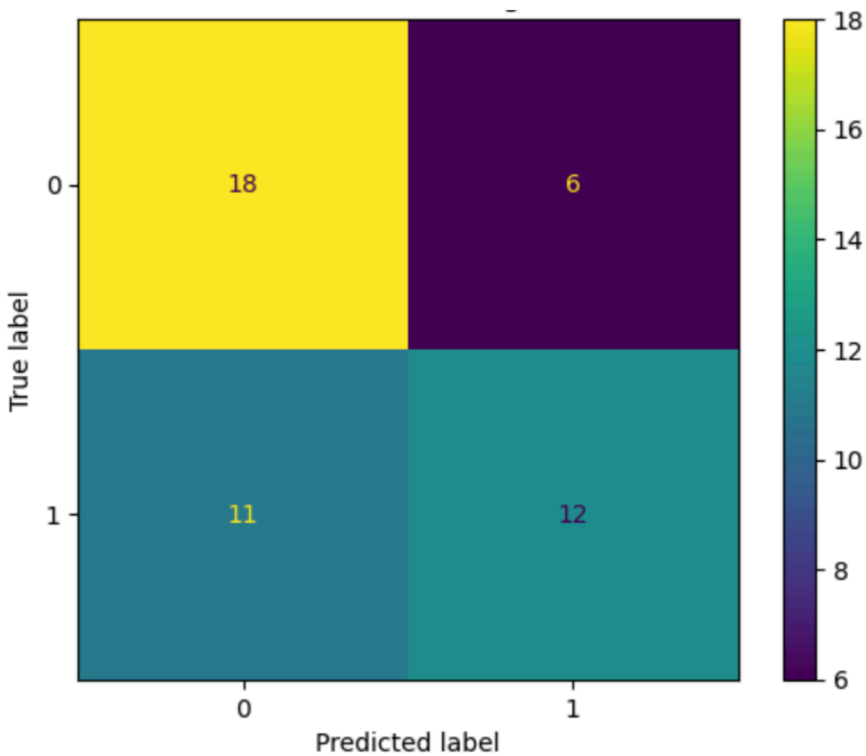


Figure 2: LightGBM Model Confusion Matrix

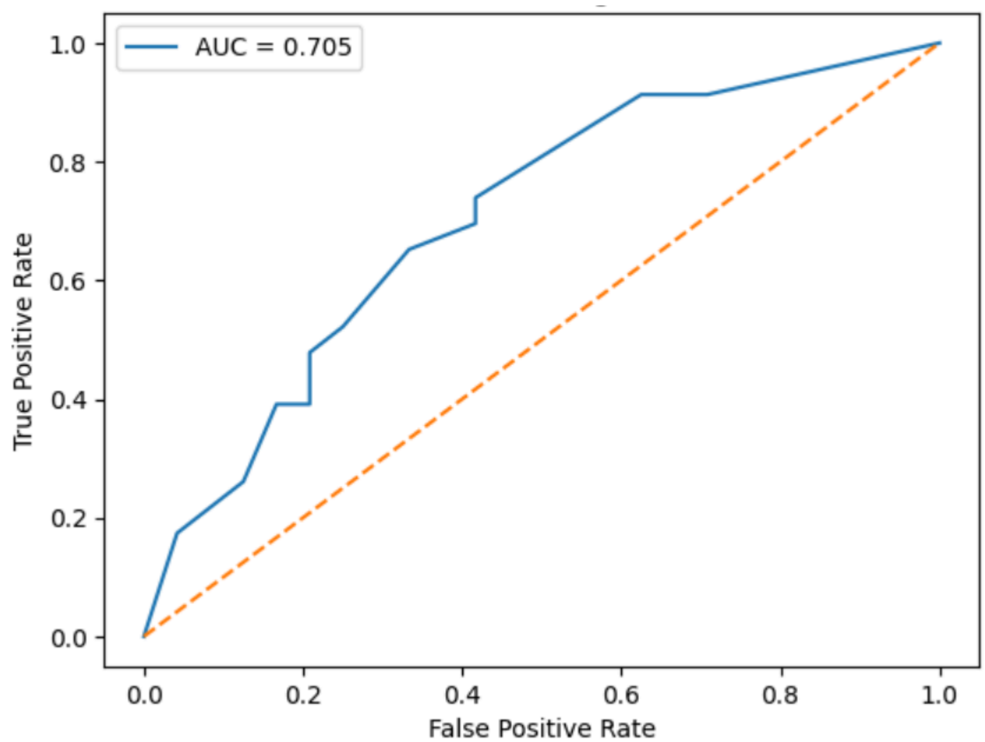


Figure 3: ROC Curve LightGBM

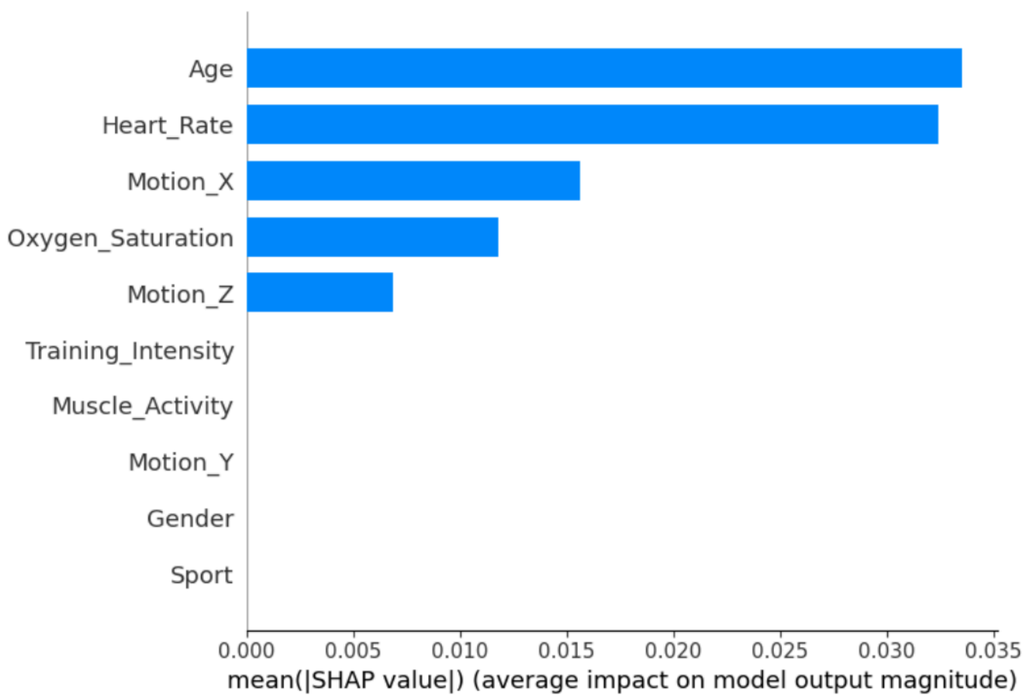


Figure 4: Feature Importance Based on SHAP Values

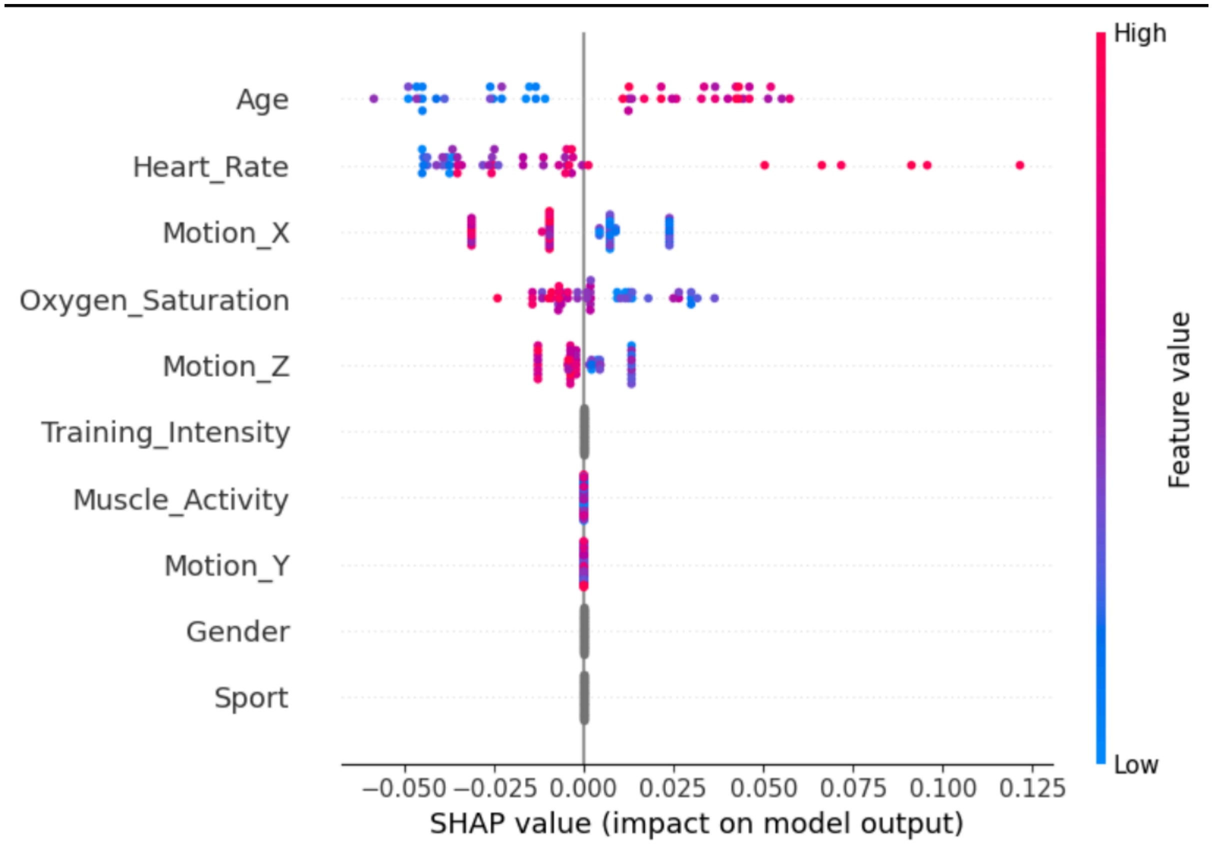


Figure 5: SHAP Summary Plot

ANALYSIS AND DISCUSSION

Model Performance Summary (LightGBM)

As shown in Table 2, overall accuracy is 63.8%, indicating that approximately 64% of predictions are correctly classified. Precision (66.7%) suggests that when the model predicts an injury, it is correct two-thirds of the time. Recall (52.2%) indicates that just over half of actual injury cases are successfully identified. The F1-score (58.5%) reflects a balanced trade-off between precision and recall. The ROC–AUC score of 70.5% indicates acceptable discriminative performance, showing the model can distinguish between injured and non-injured athletes by more than 40% of a random chance.

While the model demonstrates only moderate predictive performance, this limits its suitability for direct predictive application. It also illustrates the multifaceted and intricate nature of injury risk, implying that not all pertinent aspects may be captured by the variables presented. Even so, the findings are still useful for comprehending feature-level correlations, and SHAP-based interpretations shed light on the variables most closely associated with increased injury risk.

Role of Age and Gender in Physiological Response

As shown in Figure 4 and 5, age emerges as the top predictor, as reflected in the SHAP (see Appendix A) analyses, indicating that physiological resilience and recovery capacity vary across age groups. Older athletes may exhibit altered cardiovascular response, reduced neuromuscular efficiency, or slower recovery kinetics, which can amplify fatigue-related risk under comparable training loads (Markus et al., 2022).

Gender contributes minimally to the model’s predictions, suggesting individual variability within gender groups exceeds differences between them.

The model demonstrates partial effectiveness as an early warning system. A recall of 52.2% indicates that more than half of high-risk cases are detected before confirmed injury, making the model useful for proactive monitoring. The improved precision (66.7%) means that alerts are more reliable and less likely to result in unnecessary interventions.

Suitability for Real-Time Wearable Integration

The model’s reliance on continuous physiological and motion data makes it well suited for integration into wearable monitoring systems. LightGBM’s computational efficiency enables low-latency inference, supporting near real-time fatigue or injury risk alerts. Nevertheless, careful threshold tuning and deployment-specific calibration are essential to balance sensitivity and false-alarm rates in real-world settings.

Dominant Predictors of Injury Risk

As shown in Figure 4, SHAP analysis identified Age, Heart Rate, Motion metrics, and Oxygen Saturation as the strongest predictors of injury risk. Oxygen saturation and heart rate together capture the balance between oxygen delivery and cardiovascular demand; SHAP patterns show that drops in oxygen saturation combined with elevated or variable heart rate significantly increase predicted injury risk, indicating inadequate physiological recovery. This reflects impaired aerobic efficiency and autonomic imbalance, which can accelerate fatigue, reduce neuromuscular control, and increase susceptibility to injury.

Motion-derived features further strengthen this prediction. Motion_X and Motion_Z consistently influenced model outputs, suggesting that biomechanical deviations in movement and stability emerge before high-risk classifications. These findings align with injury prevention research frameworks such as the model proposed by Willem van Mechelen, which emphasizes identifying risk factors prior to intervention (Mechelen, 1992). By detecting early physiological and biomechanical changes, the model supports data-driven monitoring approaches consistent with real-world injury prevention strategies highlighted in the Translating Research into Injury Prevention Practice (TRIPP) Framework developed by Caroline Finch (Finch, 2006).

Influence of Sport Type and Training Intensity

Although the dataset does not allow comparison across specific sports or training intensities, the model suggests context-dependent fatigue patterns. Despite a ROC–AUC of 0.705, using global decision thresholds may mask sport-specific physiological and biomechanical responses to training stress (Bittencourt et al., 2016).

Limitations of Model

Despite improved performance, several limitations remain:

- A recall of 0.522 indicates that a portion of high-risk cases are still missed.
- The absence of longitudinal athlete tracking limits individualized fatigue trend analysis.
- Sensor noise and potential simulation bias may affect feature reliability.
- A fixed classification threshold may not be optimal across different athletic contexts

Directions for Future Research

- Personalized baseline modelling for individual athletes, allowing injury risk to be evaluated relative to each athlete's normal physiological patterns rather than generalized population averages.
- Threshold optimization using cost-sensitive or adaptive learning approaches, improving the model's ability to detect high-risk states while minimizing false positives or missed injury warnings.
- Long-term injury and recovery tracking, which would enable the model to capture temporal patterns in fatigue, adaptation, and rehabilitation across training cycles.
- Integration of additional stress indicators such as sleep quality and psychological load, providing a more holistic representation of athlete well-being and improving the accuracy of injury risk prediction.

CONCLUSION

This study investigated an application of an interpretable LightGBM machine learning model for injury-risk assessment using physiological, biomechanical, and demographic indicators derived from wearable sensor data. The model showed moderate predictive performance. SHAP analysis was used and Age, Heart Rate, Motion Metrics and Oxygen Saturation were identified as the key factors highlighting the combined role of physiological resilience, cardiovascular response, and biomechanical movement patterns in injury risk. This study demonstrated the potential of machine learning to support data-driven athlete monitoring and provide insights into the factors associated with injury risk, using a combination of predictive modelling along with explainable artificial intelligence. Although further validation using larger real-world datasets is required before practical implementation, this study provides a strong foundation for future research into interpretable, wearable-based injury-risk prediction systems and contributes to the growing application of artificial intelligence in sports medicine.

REFERENCES

- colabsss. (n.d.). Real-time physiological monitoring in sports [Dataset]. Kaggle. https://www.kaggle.com/datasets/colabsss/real-time-physiological-monitoring-in-sports

- Yin, M. (2025). AI-driven medical image analysis for sports injury diagnosis and prevention. Scientific Reports, 15(1), 41484. https://doi.org/10.1038/s41598-025-20580-y

- Halson, S. L. (2014). Monitoring training load to understand fatigue in athletes. Sports Medicine, 44(2), 139–147. https://doi.org/10.1007/s40279-014-0253-z

- Van Mechelen, W., Hlobil, H., & Kemper, H. C. (1992). Incidence, severity, aetiology and prevention of sports injuries. Sports Medicine, 14(2), 82–99. https://doi.org/10.2165/00007256-199214020-00002

- Gabbett, T. J. (2016). The training—injury prevention paradox: Should athletes be training smarter and harder? British Journal of Sports Medicine, 50(5), 273–280. https://doi.org/10.1136/bjsports-2015-095788

- Schwellnus, M., Soligard, T., Alonso, J. M., Bahr, R., Clarsen, B., Dijkstra, H. P., … Engebretsen, L. (2016). How much is too much? (Part 2) International Olympic Committee consensus statement on load in sport and risk of illness. British Journal of Sports Medicine, 50(17), 1043–1052. https://doi.org/10.1136/bjsports-2016-096572

- Injury Prevention in Sport. (2022, August 18). Physiopedia, . Retrieved 06:35, March 20, 2026 https://www.physio-pedia.com/index.php?title=Injury_Prevention_in_Sport&oldid=314100

- Fang, D., & Chen, C. (2025). Sports injury risk prediction based on temporal graph encoding and graph neural networks: A cross-sport transfer learning framework. Scientific Reports, 15(1), 37656. https://doi.org/10.1038/s41598-025-21613-2

- Leckey, C., Van Dyk, N., Doherty, C., Lawlor, A., & Delahunt, E. (2024). Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis. British Journal of Sports Medicine, 59(7), 491–500. https://doi.org/10.1136/bjsports-2024-108576

- Professional, C. C. M. (2025, September 18). Heart rate. Cleveland Clinic. https://my.clevelandclinic.org/health/diagnostics/heart-rate

- Holland, K. (2023, June 29). Is my blood oxygen level normal? Healthline. https://www.healthline.com/health/normal-blood-oxygen-level#results

- Edwards, P. K., Ebert, J. R., Littlewood, C., Ackland, T., & Wang, A. (2017). A systematic review of electromyography studies in normal shoulders to inform postoperative rehabilitation following rotator cuff repair. Journal of Orthopaedic and Sports Physical Therapy, 47(12), 931–944. https://www.jospt.org/doi/10.2519/jospt.2017.7271

- Mayo Clinic. (n.d.). Exercise intensity: How to measure it. Mayo Clinic. https://www.mayoclinic.org/healthy-lifestyle/fitness/in-depth/exercise-intensity/art-20046887

- StriveOn. (2025). Youth sports injuries statistics: Rates by sport. StriveOn. https://joinstriveon.com/blog/youth-sports-injuries-statistics

- Markus, I., Constantini, K., Goldstein, N., Amedi, R., Bornstein, Y., Stolkovsky, Y., Vidal, M., Lev-Ari, S., Balaban, R., Leibou, S., Blumenfeld-Katzir, T., Ben-Eliezer, N., Peled, D., Assaf, Y., Jensen, D., Constantini, N., Dubnov-Raz, G., Halperin, I., & Gepner, Y. (2022). Age differences in recovery rate following an Aerobic-Based exercise protocol inducing muscle damage among amateur, male athletes. Frontiers in Physiology, 13, 916924. https://pubmed.ncbi.nlm.nih.gov/35774290/

- Bittencourt, N. F. N., Meeuwisse, W. H., Mendonça, L. D. M., Nettel-Aguirre, A., Ocarino, J. M., & Fonseca, S. T. (2016). Complex systems approach for sports injuries: Moving from risk factor identification to injury pattern recognition—narrative review and new concept. British Journal of Sports Medicine, 50. https://doi.org/10.1136/bjsports-2015-095850

APPENDIX A: KEY DEFINITIONS

Athlete Injury Risk
The likelihood that an athlete may experience a sports-related injury based on physiological, biomechanical, and training-related factors.

Lower Extremity Injuries
Injuries affecting the lower part of the body, including the hips, thighs, knees, lower legs, ankles, and feet.

Physiological Metrics
Measurements that reflect the body's physiological functioning, such as heart rate, oxygen saturation, and muscle activity.

Heart Rate (HR)
The number of times the heart beats per minute, commonly used to measure exercise intensity and physiological stress.

Oxygen Saturation (SpO_2)
The percentage of oxygen carried by haemoglobin in the blood, indicating how effectively oxygen is being delivered to body tissues.

Motion-Based Metrics
Movement-related measurements captured using wearable sensors that record motion across different axes (e.g., Motion X, Motion Y, Motion Z).

SHAP Values (SHapley Additive Explanations)
A method used to interpret how different variables contribute to an outcome by estimating the influence of each feature on the result.

ROC Curve (Receiver Operating Characteristic Curve)
A graphical representation showing the relationship between the true positive rate (sensitivity) and the false positive rate across different classification thresholds.

ROC-AUC (Area Under the ROC Curve)
A measure of a model’s ability to distinguish between positive and negative cases. It represents the probability that a randomly selected positive example will be ranked higher than a randomly selected negative example.

Precision

Precision is the proportion of all the model's positive classifications that are actually positive. It is mathematically defined as:

Precision = $\frac{correctly\ classified\ actual\ positives}{Everything\ classified\ as\ positive}$

Recall

The true positive rate, or the proportion of all actual positives that were classified correctly as positives, is also known as recall. It is mathematically defined as:

Recall = $\frac{correctly\ classified\ actual\ positives}{all\ actual\ positives}$

F1 score

The F1 score is the harmonic mean (a kind of average) of precision and recall. Mathematically it is given by:

F1 score = $2\times\frac{precision\ \times recall}{precision+recall}$

Accuracy

It is the proportion of all classifications that were correct, whether positive or negative. It is mathematically defined as:

Accuracy = $\frac{True\ positive+True\ negative}{True\ positive+True\ negative+False\ positive+False\ negative}$

Confusion Matrix

A table used to summarize classification results by displaying the number of true positives, true negatives, false positives, and false negatives which are critical in injury prediction contexts where missed injury risks are especially consequential.

.