International Journal of Basic and Applied Sciences, 14 (8) (2025) 447-452

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/hf5g9c06
Research paper

Enhancing Runner Safety: A Machine Learning Framework for Injury Prediction

Ashwini Sawant [1] *, Sangeeta Oswal [2], Manisha Chattopadhyay[1], Utsav Mutadak [1]

[1] Department of Electronics and Telecommunication Engineering, Vivekanand Education Society's Institute of Technology (VESIT), Mumbai University (MU), Mumbai, Maharashtra, India
[2] Department of Artificial Intelligence and Data Science Engineering, Vivekanand Education Society's Institute of Technology (VESIT), Mumbai University (MU), Mumbai, Maharashtra, India
**Corresponding author E-mail: ashwini.sawant@ves.ac.in*

Received: October 30, 2025, Accepted: December 9, 2025, Published: December 19, 2025

Abstract

Avoiding injuries is crucial for athletic performance. Despite the difficulty in predicting injuries, new technologies and data science applications may offer valuable information. The goal was to apply machine learning to predict in juries in runners based on complete training data. Over a seven-year span, 74 elite middle- and long-distance runners were used to test injury prediction. Two methods of analysis were used. Initially, a time series was created that represents the training load for the preceding seven days, with ten features describing the training for each day. These characteristics combined subjective information on training effectiveness and training effort level with data from a watch (such as duration and distance). The average area under the ROC-AUC curve for the day approach were produced to be 0.9978 by a prediction system based on Random Forest Classifier machine learning model. The machine learning-based method predicts a significant percentage of injuries when the model is built using training data recorded on a daily basis in athlete training. All things considered, the results showcase the benefits of applying machine learning to forecast injuries and also customize athlete training regimens.

Keywords: *Random Forest Classifier; Machine Learning; Injury Prediction; Data Sampling.*

1. Introduction

Injuries in endurance sports like running disrupt training, impair performance, and jeopardize athletes' careers, often caused by intense training loads, poor biomechanics, and inadequate recovery [1, 2]. Traditional injury prevention methods are subjective and fail to address dynamic risk factors, such as variations in train ing intensity and fatigue levels [3]. Machine learning offers a data-driven approach to injury prediction by analyzing vast amounts of training and movement data, enabling proactive measures [4], [5].

Individualized load monitoring and stress-recovery balance are critical in reducing injuries across sports like soccer, rugby, and cricket [6 – 8]. Recent advancements in machine learning, such as XGBoost and GPS data integration, have improved injury prediction accuracy and interpretability [9 – 11]. However, challenges remain, including sport-specific limitations and the need for larger datasets [12], [13]. This paper proposes a machine learning framework to predict injuries in runners, leveraging training load data to enhance athlete safety and optimize performance [14 – 16]. The framework integrates both subjective and objective data, such as GPS metrics, to provide accurate and actionable insights for injury prevention [17 – 19]. Sudden increases in weekly running distance beyond 30 percent raise injury risks in novice runners, highlighting the need for gradual progression [20]. Aerobic fitness measures failed to predict success in young cyclists, underscoring the need for broader talent identification criteria [21].

This research [22] demonstrated that accelerometers can validly and reliably collect kinematic data from surfers and their boards, producing waveform signals comparable to motion-capture systems. The study found that different surfing manoeuvres generate distinct acceleration patterns and that significant differences exist between complete and incomplete bottom-turn–to–top-turn combinations. Overall, it provides the first detailed evaluation of lower-body and surfboard acceleration during common surfing manoeuvres, contributing important new evidence to surfing performance research.

This study [23] examined how physiotherapists reason clinically when identifying and treating upper-limb movement difficulties after stroke, using a mixed-methods design involving surveys, interviews, and recorded treatment-session reflections. Findings showed that clinical decisions were mainly influenced by therapists' experience, theoretical knowledge, service structure, and holistic factors such as sensory and emotional aspects of stroke, while service users valued sensory experiences of movement and desired more collaborative discussion. Overall, the study highlights the flexible yet structured nature of physiotherapy decision-making post-stroke and emphasizes the importance of integrating service-user perspectives in rehabilitation.

Injuries in endurance sports like running can disrupt training, impair performance, and jeopardize athletes' careers. Traditional injury prevention methods are often subjective and fail to address dynamic risk factors. There is need for a data-driven approach using machine learning to predict injuries, enabling proactive measures to enhance athlete safety and optimize training.

The following paper begins with an introduction that highlights the importance of injury prevention and the limitations of traditional methods in machine learning. The methodology details a five-stage workflow which includes data pre-processing to model deployment. The results and discussion section evaluates model performance using visualizations like confusion matrix and ROC curve. Finally, the conclusion summarizes findings, suggests future improvements for broader applicability, and concludes with references

2. Methods

The five sequential steps of the athlete injury-risk prediction workflow are shown in Figure 1: data preprocessing, correcting class imbalance, feature selection, model training, and deployment through a user interface. High predictive accuracy and useful interpretability for sport-science professionals were the two main goals of the workflow's design. Each step converts unprocessed athlete monitoring data into useful and physiologically based risk insights.

Stage 1: Data Preprocessing

Structured preprocessing was applied to raw athlete-monitoring data from a professional Dutch endurance squad from 2012 to 2019. Training load variables, session kinds, and strength sessions were all documented on a weekly basis.

Preprocessing is involved in:

- Using domain-appropriate imputation (such as forward-fill for weekly summaries) to handle missing values.
- Eliminating physiologically implausible outliers, such as negative session counts.
- Standardisation where necessary to enable model-based load metrics comparison.

Because minor discrepancies might have a disproportionate impact on injury prediction models, our procedures guaranteed consistency throughout a longitudinal dataset, which is especially crucial.

In order to prevent leaking, temporal integrity was maintained by processing all data before splitting. A chronological train–test split was employed to preserve the temporal structure of the weekly data, with the first 80% of the data being used for training and the last 20% set aside for testing. This keeps future training weeks from affecting earlier forecasts.

Stage 2: Resolving Dataset Imbalance in Stage Two

The initial dataset (575 injury weeks vs. 42,223 non-injury weeks) was extremely unbalanced. Due to the rarity of injury episodes in elite endurance sports, this imbalance may cause a model to consistently predict "no injury".

They employed a combination approach:

1) ADASYN oversampling, concentrating on injury cases that are challenging to learn, to create artificial minority samples.
2) Random undersampling to shrink the size of the majority class.

As a result, there were 30,129 injury entries and 42,000 non-injury entries, making the distribution more evenly distributed.

Overfitting Precautions: The model may learn minority patterns as a result of aggressive oversampling.

- Oversampling was only used on the training set, never on the validation or testing sets, in order to decrease this risk.
- To prevent excessive replication of damage patterns, two distinct imbalance techniques (ADASYN + undersampling) were employed.
- To identify indications of oversampling-induced overfitting, precision-recall analysis and confusion-matrix interpretation are included in the later model evaluation (Section 3).

Stage 3: Feature Selection

Recursive Feature Elimination (RFE) on the training set with cross-validation was used for feature selection after balancing. The top ten characteristics kept for modelling were as follows:

Physiological explanation of specific traits

- Strength training (ranked highest): By influencing tissue exhaustion and motor-unit recruitment stability, variations in neuromuscular load might modify injury risk.
- High-intensity and interval training: Musculoskeletal strain is known to be increased by sudden increases in anaerobic load, particularly in endurance athletes.
- Frequency of rest days: Inadequate recuperation impairs tissue adaptability and raises the risk of soft-tissue injuries.
- Seasonality (Date): Pre-competition and early-season conditioning stages frequently exhibit unique injury patterns.

Incorporating these biologically interpretable characteristics improves the model's credibility and usefulness for athletes.

Stage 4: Model Training and Testing

The chronologically earlier part of the dataset was used to train three models (Logistic Regression, Random Forest (RF), and XGBoost), which were then assessed using the temporal hold-out set. Among the evaluation metrics were: Precision, recall, F1-score, and accuracy, andROC-AUC.

The model with the greatest AUC (0.9997) was Random Forest. However, considering the exceptionally high performance, possible oversampling-related overfitting was investigated using:

- Balanced accuracy
- Analysis of false positives and false negatives
- Comparison with simpler models (Logistic Regression)

Section 3 provides a detailed discussion of model performance.

Stage 5: Front-End User Interface

The trained RF model was serialised as a pickle object and implemented into a front-end interface for practical use by coaches. Weekly load metrics and outputs can be entered into the interface.

- Injury-risk probability
- Explanation of contributing features (via feature importance ranking)

This closes the gap between the results of machine learning and practical coaching processes.

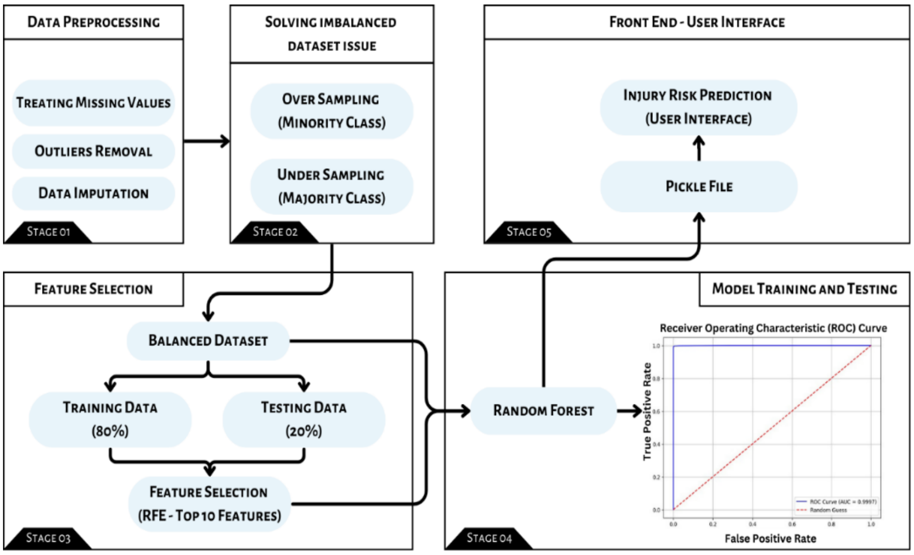

Fig. 1: Proposed Workflow for Athlete Injury-Risk Prediction.

3. Results and Discussion

This chapter treats the entire development process that dates back to the first stages of design and ends with the later stages of testing and implementation. The purpose of any particular section would be to give a clear understanding of how the solution was constructed and how it was implemented in a practical way.

3.1. Data pre-processing and balanced dataset

Only 1.34% of the 42,798 weekly athlete records in the raw dataset were marked as injury weeks. The initial and balanced distributions are displayed in Figures 2 and 3.
Using density-based sampling in ADASYN, more injury-labelled samples are balanced with realistic physiological variance.

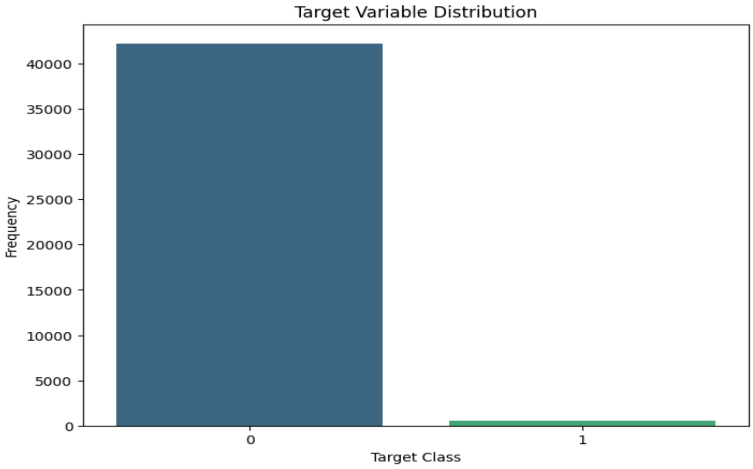

Fig. 2: Injury Data Count Before Sampling.

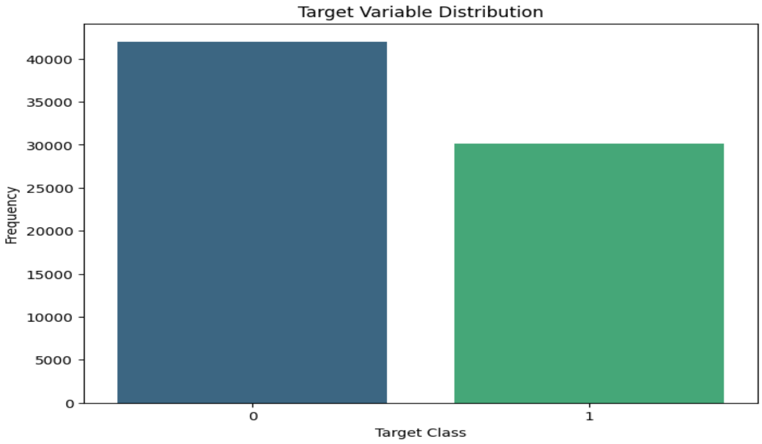

Fig. 3: Injury Data Count After Sampling.

3.2. Feature selection

Figure 4 depicts the relative significance of characteristics chosen using Recursive Feature Elimination (RFE) based rankings, which showed that:

- High-intensity/tough sessions were likewise highly predictive;
- Strength-training count and variability were top predictors; and
- Rest-day metrics and interval-session frequency moderately predictive
- Seasonality (Date) that records annual load cycles

These findings are consistent with the injury-risk theory in endurance sports, where inadequate recovery and abrupt load spikes are significant factors.

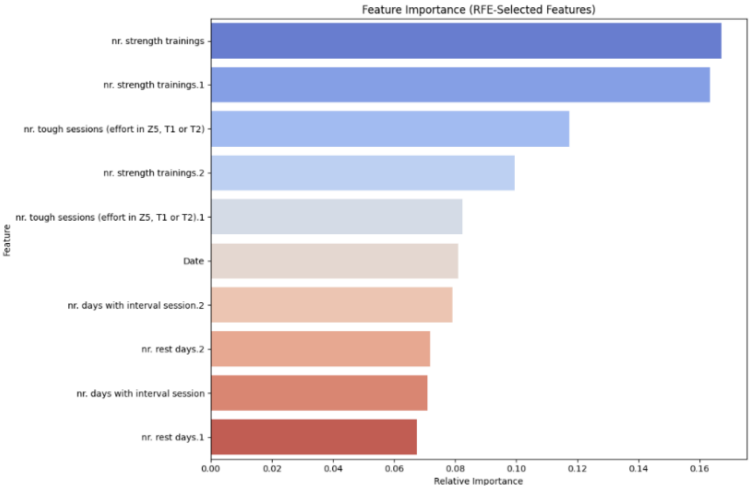

Fig. 4: Bar Plot for RFE.

3.3. Model training and testing

Table 1 provides an overview of the findings. Because injury-load connections are non-linear, logistic regression performed poorly. Metrics for Random Forest and XGBoost were exceptionally high: RF AUC is 0.9997, and XGB AUC is 0.9976.

Interpretation of confusion matrix:

RF yielded the following results despite correctly predicting the majority of injury and non-injury samples: 119 false negative results and 12 false positive results. Strong performance is indicated by such few errors, but they also raise the possibility of oversampling-driven overfitting, which needs to be noted in restrictions.

Table 1: Model Comparison Table

Model	Class	Precision	Recall	F1-Score	ROC-AUC Score	Accuracy
Logistic Regression	0	0.69	0.75	0.72	0.7142	0.6567
	1	0.59	0.52	0.55		
XGBoost	0	0.99	0.96	0.99	0.9976	0.9918
	1	0.96	0.99	0.99		
Random Forest Classifier	0	0.99	0.99	0.99	0.9997	0.9929
	1	0.98	0.98	0.99		

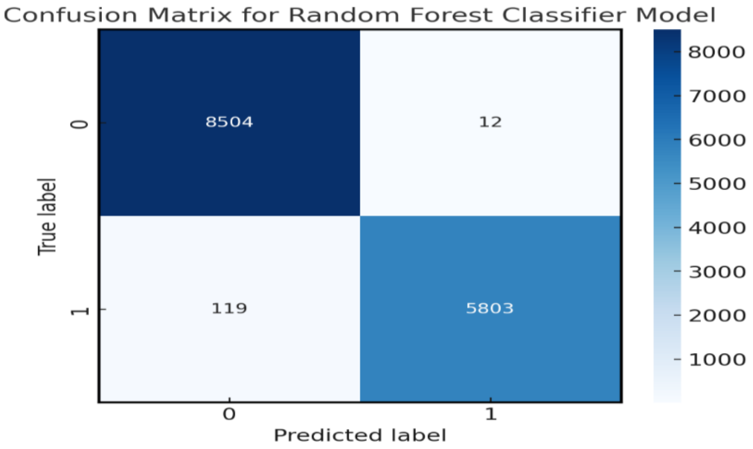

Fig. 5: Confusion Matrix for Random Forest Model

In order to forecast an athlete's risk of injury, this ROC curve in Figure 6 compares the outcomes of models like XGBoost, Logistic Regression, and Random Forest Classifier. The top-left region is hugged by the RF and XGB curves, indicating strong discriminative power. The curve of logistic regression is far closer to performance at the chance level.

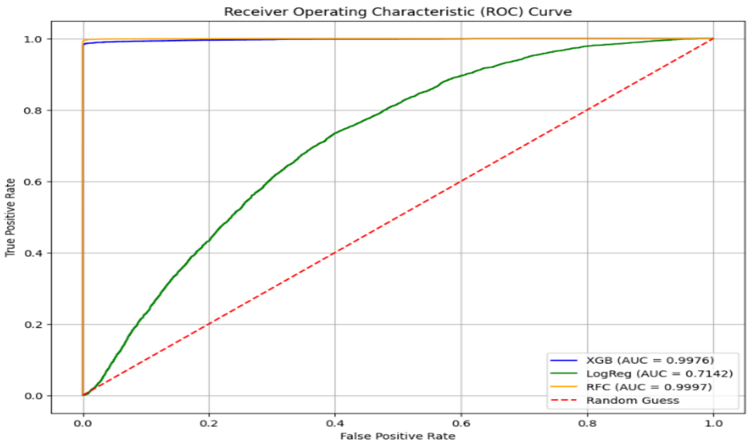

Fig. 6: ROC Curve for Random Forest Classifier.

3.4. Practical implications for coaches

This model helps coaches:

- Track weekly shifts in risk instead of depending solely on instinct;
- Recognise high-load weeks that need for proactive recovery techniques
- Using objective risk assessments to support return-to-play decisions;
- modifying the micro cycle structure (interval sessions, rest days); and
- Planning customised strength training based on injury susceptibility

The interface offers comprehensible insights that complement regular coaching procedures.

4. Limitations and Future Work

- Generalizability: Patterns may vary for sprinters, team sports, or young athletes; data comes from a single elite Dutch endurance squad.
- Oversampling risks: ADASYN might provide artificial instances that are too similar to noise, which would boost model performance.
- Temporal correlation: Daily load swings within a week may not be captured by weekly statistics.
- Feature omissions: Sleep and HRV, among other physiological indicators, were excluded.
- Model complexity: Despite feature importance analysis, highly effective ensemble models might still be somewhat opaque.

Future research could investigate different imbalance-handling techniques including focus loss or cost-sensitive learning, validate the model on external cohorts, and incorporate richer physiological information.

5. Conclusions

The research aims to predict the injury risk in athletes using machine learning approach. Of all the models evaluated, the random forest classifier shows the highest performance in the data set used. Achieving an accuracy of 99.29% and an outstanding ROC-AUC score of 0.9997, the model effectively identifies injury risks with remarkable reliability. These precise predictions can aid in early intervention and the implementation of preventive strategies to safeguard athletes.

Funding Information

The authors declare that there was no kind of funding involved in this research.

Conflict of Interest Statement

The authors state that the work described in this study could not have been affected by known competing financial interests or personal relationships. The authors do not disclose conflicts of interest.

Data Availability

The open data repository website Kaggle provided the dataset utilized in this investigation. The information underlying the research's conclusions is openly accessible at the given URL https://www.kaggle.com/datasets/shashwatwork/injury-prediction-for-competitive-runners

References

[1] Lovdal, S., den Hartigh, R., & Azzopardi, G., "Injury Prediction in Competitive Runners with Machine Learning", International journal of sports physiology and performance, 16(10), 1522–1531.2021. https://doi.org/10.1123/ijspp.2020-0518.

[2] Dennis van Poppel, Maarten van der Worp, Slabbekoorn A, et al., "Risk factors for overuse injuries in short- and long-distance running: a systematic review", Journal of Sport and Health Science, Volume 10, Issue 1, January 2021, Pages 14-28. https://doi.org/10.1016/j.jshs.2020.06.006.

[3] Xuelian Dong, Yuanhang Sun, “Injury Risk Prediction and Prevention Algorithm for Athletes Based on Data Mining,” Journal of Electrical Systems, pp.1717–1728, 2024. https://doi.org/10.52783/jes.3090.

[4] Carey, D. L., Ong, K., Whiteley, R., Crossley, K. M., Crow, J., Morris, M. E.,"Predictive Modelling of Training Loads and Injury in Australian Football”, International Journal of Computer Science in Sport vol. 17, Issue 1, 2018, doi: 10.2478/ijcss-2018-0002. https://doi.org/10.2478/ijcss-2018-0002.

[5] Raysmith BP, DrewMK., "Performance successor failure is influenced by weeks lost to injury and illness in elite Australian track and field athletes: a 5-year prospective study", Journal of Science and Medicine in Sport Volume 19, Issue 10, October 2016, Pages 778-783, PubMed ID: 26839047. https://doi.org/10.1016/j.jsams.2015.12.515.

[6] D’souza D., "Track and field athletics injuries–a one-year survey", British Journal of Sports Medicine, 1994;28:197–202. PubMed ID: 8000821 doi:10.1136/ bjsm.28.3.197. https://doi.org/10.1136/bjsm.28.3.197.

[7] Soligard T, Schwellnus M,AlonsoJ-M,etal., "How much is too much? (Part 1) International Olympic Committee consensus statement on load in sport and risk of injury", British Journal of Sports Medicine. 2016;50(17):10301041. PubMed ID: 27535989 https://doi.org/10.1136/bjsports-2016-096581.

[8] Brink MS, Visscher C, Arends S, Zwerver J, Post WJ, Lemmink KAPM, "Monitoring stress and recovery: new insights for the prevention of injuries and illnesses in elite youth soccer players", British Journal of Sports Medicine, 2010; 44(11):809–815, PubMed ID: 20511621. https://doi.org/10.1136/bjsm.2009.069476.

[9] Cross MJ, Williams S, Trewartha G, Kemp SPT, Stokes KA., "The influence of in-season training loads on injury risk in professional rugby union", Int J Sports Physiol Perform. 2016;11(3):350–355. https://doi.org/10.1123/ijspp.2015-0187.

[10] Dennis R., Farhart R, Goumas C, Orchard J., "Bowling workload and the risk of injury in elite cricket fast bowlers", Journal of Science and Medicine in Sport, Volume 6, Issue 3, September 2003, Pages 359-367 PubMed ID: 14609154 https://doi.org/10.1016/S1440-2440(03)80031-2.

[11] Huxley D J , O’Connor D, Healey PA., "An examination of the training profiles and injuries in elite youth track and field athletes", European Journal of Sport Science, 2014;14(2):185–192. PubMed ID: 23777449 https://doi.org/10.1080/17461391.2013.809153.

[12] Jaspers A, Op De Be´ eck T, Brink MS, et al., "Relationships between the external and internal training load in professional soccer: what can we learn from machine learning?", International Journal of Sports Physiology and Performance, 2018;13(5):625–630. https://doi.org/10.1123/ijspp.2017-0299.

[13] Knobbe A, Orie J, Hofman N, Burgh B, Cachucho R., "Sports analytics for professional speed skating", Data Min Knowl Disc. 2017; 31(6):1872– https://doi.org/10.1007/s10618-017-0512-3.

[14] Naglah A, Khalifa F, Mahmoud A, et al., "Athlete-customized injury prediction using training load statistical records and machine learning", Data Mining and Knowledge Discovery, Volume 31, pages 1872–1902, (2017) https://doi.org/10.1109/ISSPIT.2018.8642739.

[15] Raya-Gonzalez J, Nakamura FY, Castillo D, Yanci J, Fanchini M., "Determining the relationship between internal load markers and noncontact injuries in young elite soccer players", International Journal of Sports Physiology and Performance. 2019;14(4):421–425. https://doi.org/10.1123/ijspp.2018-0466.

[16] Rossi A, Pappalardo L, Cintia P, Iaia FM, Fern` andez J, Medina D., "Effective injury forecasting in soccer with GPS training data and machine learning", PLoS One. 2018;13(7):e0201264. PubMed ID:30044858 https://doi.org/10.1371/journal.pone.0201264.

[17] Lysholm J, Wiklander J., "Injuries in runners", The American Journal of Sports Medicine, 1987;15(2):168–171. PubMed ID: 3578639 https://doi.org/10.1177/036354658701500213.

[18] Van Der Does HTD, Brink MS, Otter RTA, Visscher C, Lemmink KAPM, "Injury risk is increased by changes in perceived recovery of team sport players", Clinical Journal of Sport Medicine 27(1):p 46-51, January 2017. https://doi.org/10.1097/JSM.0000000000000306.

[19] Rogalski B, Dawson B, Heasman J, Gabbett TJ., "Training and game loads and injury risk in elite Australian footballers", Journal of Science and Medicine in Sport, Volume 16, Issue 6, November 2013, Pages 499-503. PubMed ID: 233330 https://doi.org/10.1016/j.jsams.2012.12.004.

[20] Nielsen RO, ParnerET, NohrEA, SørensenH, LindM, RasmussenS., "Excessive progression in weekly running distance and risk of running-related injuries: an association which varies according to type of injury", J Orthop Sports Phys Ther. 2014;44(10):739–747. PubMed ID: 25155475 https://doi.org/10.2519/jospt.2014.5164.

[21] Menaspa P, Sassi A, Impellizzeri F M. "Aerobic fitness variables donot predict the professional career of young cyclists" Medicine and Science in Sports and Exercise, 2010; 42(4):805–812. PubMed ID: 19952851 https://doi.org/10.1249/MSS.0b013e3181ba99bc.

[22] Jayden Lowrie, “Accelerometry in Surfing: the feasibility of employing accelerometers to capture, detect, and analyse the waveform signals of manoeuvres and movements performed during wave-riding”, University of the Sunshine Coast, Queensland, Doctor of Philosophy, University of the Sunshine Coast, Queensland, Dissertations & Theses, 2022

[23] Bamborough, G M, “An Exploration of the Influences on Clinical Decisions, Made by Chartered Physiotherapists in Relation to the Hemiplegic Upper Limb Following Stroke”, University of Northumbria at Newcastle (United Kingdom) ProQuest Dissertations & Theses, 2015. 27747459.