Beyond Training: A Personalized Holistic Injury Prediction in Triathletes

Leonardo Rossi, Bruno Rodrigues
Embedded Sensing Group ESG, Institute of Computer Science in Vorarlberg ICV,
University of St. Gallen HSG, Switzerland
E-mail: leonardo.rossi@student.unisg.ch, bruno.rodrigues@unisg.ch

***Abstract*—Triathlon training combines swimming, cycling, and running, often at high volumes, to prepare athletes for long-distance events. The highly intense physical demand puts athletes at significant risk of overuse injuries. While wearable devices provide continuous, high-frequency insights into an athlete's physiological response to training, extracting meaningful, actionable patterns remains a challenge - especially for everyday users. Understanding these metrics and their relationship with injury risk is critical to optimizing training strategies and preventing injuries before they occur.**

This work proposes a learning model to identify patterns that indicate an increased risk of injury, allowing proactive adjustments to training loads. However, building a generalizable model in sports science and healthcare presents a key challenge: the need for large, high-quality labelled datasets, which are often limited by privacy concerns. To address this limitation, this work also explores the generation and application of a highly realistic synthetic dataset that ensures robust model training while mitigating privacy constraints.

***Index Terms*—Machine Learning, Wearable Devices, Synthetic Data, Sports Science**

I. Introduction

Triathlon is a demanding multi-sport discipline combining swimming, cycling, and running, which requires a high-level of endurance and high training volumes. The rigorous nature of triathlon training places athletes at significant risk of overuse injuries, as documented by Andersen *et al* [1], who reported that 56% of 174 participants in the 2011 Norseman Xtreme Triathlon developed overuse injuries while preparing for the event's grueling 3.8km swim, 180km cycle, and 42km run.

Overuse injuries typically stem from overtraining, *i.e.,* a condition where training loads exceed an athlete's recovery capacity, leading to physiological maladaptations, and increased injury susceptibility [2]. By recognizing patterns associated with overtraining and, therefore, injury risk, athletes and coaches can implement data-driven adjustments to training protocols.

Recent advances in wearable technology and machine learning (ML) have revolutionized sports science by enabling large-scale analysis of physiological data and predictive modeling for injury prevention. While ML has shown promise in sports like football [3] and running [6] for detecting workload-related risks based on wearable sensors, triathlon—a multi-sport context—remains underexplored. Also, existing studies often focus on sport-specific metrics, neglecting broader factors such as sleep quality, stress levels, and daily habits that influence recovery and performance. An additional significant limitation in existing research is the reliance on small, sport-specific datasets created for individual studies, which often lack diversity and generalizability. Collection of comprehensive training and injury data is hindered by ethical challenges and strict data privacy regulations in health and sports sciences [7].

This paper proposes a novel ML-based approach to predict injury risk and optimize training in triathlon. The contributions are outlined as follows:

- **Synthetic dataset**: given the scarcity of real-world datasets, this work will build a realistic synthetic dataset that models complex interactions between training load, recovery metrics, and injury risk.
- **Holistic analysis**: provide a multimodal analysis integrating lifestyle factors such as sleep quality, stress levels, and daily habits. This will provide a detailed and personalized understanding of injury risks and allow for tailoring trainings.

The remainder of the paper is outlined as follows. Section II presents selected related work in the field. Section III outlines the approach and synthetic data generation method. Section IV details the ML model development and validation strategy, and Section V presents considerations and future work.

II. Related Work

Prior research in injury prediction spans several interconnected domains that inform our approach. We examine key methodological approaches and limitations our work aims to address.

A. Injury Prediction in Endurance Sports

ML techniques are popular to predict injuries across various sports. Rossi *et al* [3] achieved 80% recall using decision trees on GPS data from professional footballers, while Lövdal *et al* [6] demonstrated that day-to-day monitoring (AUC = 0.724) outperformed weekly models in runners, suggesting higher-frequency data collection improves prediction accuracy.

The physiological foundations are established by Kienstra *et al* [8], who explored the non-linear nature of injury development in triathletes, highlighting the relationship between acute/chronic workload ratios (recommended: 0.8-1.35) and tissue adaptation. Halson [11] established that internal load

monitoring through heart rate measures and subjective assessments is essential for understanding individual responses.

Chen [10] introduced multi-relational clustering to identify distinct risk profiles, while Naglah *et al* [5] developed a hierarchical ML framework integrating load metrics from wearables. These approaches demonstrate the value of multimodal data integration but have not been specifically applied to triathlon's three-discipline context.

A shortcoming of prior work is the focus on single disciplines, failing to address the unique cross-training effects and varied biomechanical demands of triathlon. Also, a finding is that model simplicity often improves performance with physiological data outside traning, which is a principle we incorporate in our modeling strategy.

B. Synthetic Data Generation in Sports Science

The scarcity of large-scale training and injury data is a significant challenge. Hohl *et al* [7] found that TimeGAN demonstrated the best balance between fidelity, diversity, and predictive utility when evaluated using regression models with limited endurance athlete datasets. Lange *et al* [12] explored GAN-based approaches for generating physiological time series, with Conditional GAN emerging as most effective when augmenting real training data.

Our work extends these methods by developing a synthetic data generation framework specifically designed for triathlon's unique multi-disciplinary nature, incorporating domain knowledge of the interactions between swimming, cycling, and running.

III. Synthetic Data Generation Approach

Our methodology centers on creating a realistic synthetic data framework that captures the complex interaction between training load, recovery metrics, and injury risk in triathletes. This approach addresses data limitations in current research while enabling robust injury prediction models. Figure 1 provides an overview of our two-phase process.

The synthetic data generation phase is well advanced, with the main framework, outlined below, in place and showing strong potential for producing a realistic dataset. Our current focus is on fine-tuning parameters and in particular refining the correlations between heart rate, power, speed, and altitude time-series for activities to further enhance physiological realism.

A. Synthetic Athlete Profile Generation

We generate diverse, physiologically realistic athlete profiles defined by 24 parameters covering demographic factors (gender, age, height, weight), physiological metrics (genetic predisposition, heart rate variability (HRV) baseline and range, resting, maximal and lactate threshold heart rates, VO_2max), sport specific performance indicators (functional threshold power, critical swim speed, running threshold pace), training history (experience level, weekly hours, recovery rate) and lifestyle factors (sleep patterns, diet, stress, smoking and alcohol consumption)

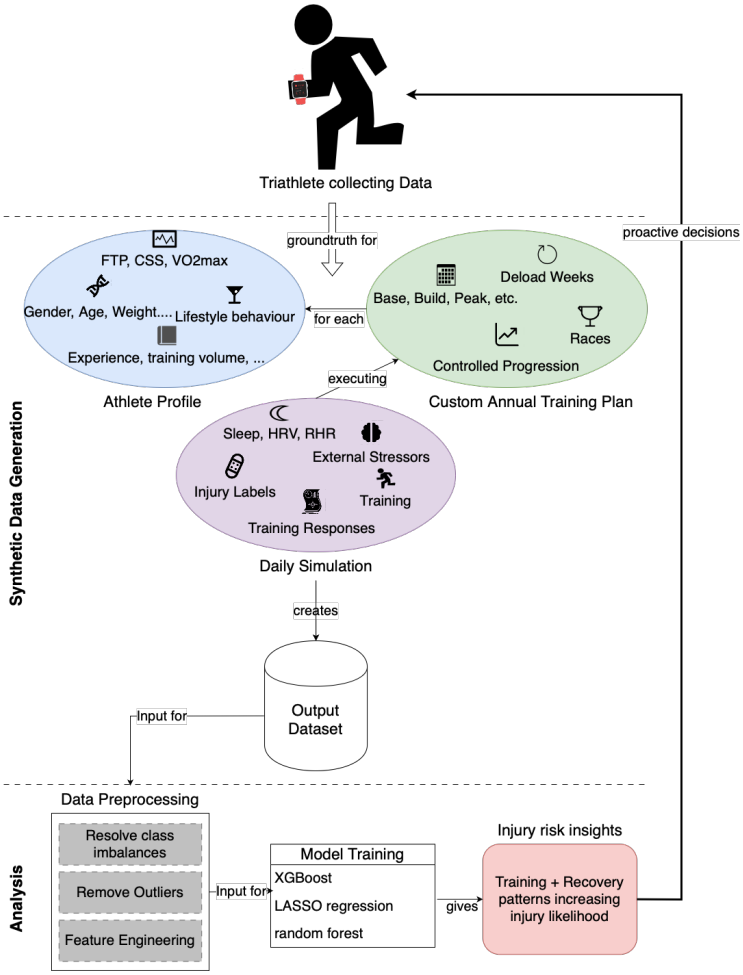

Fig. 1. Methodology overview: Synthetic Dataset Generation and ML model development

Parameter distributions are derived from published population studies of triathletes, with interdependencies modelled using established sport science principles to maintain physiological realism and internal consistency.

B. Training Plan Generation and Daily Simulation

For each synthetic athlete, we generate a customized annual training plan following periodization principles. The plan incorporates random races and structures phases (base, build, peak, recovery) based on race proximity.

The daily simulation process iterates through the year, modelling:

1) **Morning metrics simulation**: Sleep metrics, HRV, resting heart rate, and morning body battery based on previous training load and recovery status
2) **Training execution**: Probabilistic deviations from planned workouts based on morning metrics
3) **Sport-specific wearable data**: Detailed time-series data for each workout including heart rate, power, pace, and stroke rate

4) **Training load updates**: Calculation of acute (7-day) and chronic (42-day) loads across disciplines
5) **Evening metrics**: Stress levels, recovery state, and evening body battery

Critical to our approach is simulating realistic deviations between planned and actual training responses. Although coaches may design theoretically injury-proof programs following sports science principles, our model accounts for how external factors, such as lifestyle behaviors, sleep quality, and psychological stress, significantly alter an athlete's physiological response to training loads, potentially triggering injuries despite sound program design.

C. Injury Risk Modeling and Label Generation

We implement a probabilistic injury model based on established risk factors. Daily injury probability is calculated as:

$$P(injury) = f(ACWR, fatigue, recovery, athlete\_factors) \tag{1}$$

Where ACWR represents the acute:chronic workload ratio, fatigue measures accumulated training stress, recovery integrates sleep and cardiac metrics, and athlete factors include individual risk modifiers. The function incorporates non-linear relationships between risk factors, with exponentially increasing risk when multiple factors align.

IV. Model Development and Validation

Our injury prediction framework uses a three-stage modelling approach, which includes pre-processing the synthetic data, training the ML models and evaluating them. The preprocessing is crucial because there will be an imbalance between injury occurrences and non-injury days that needs to be addressed in order to avoid bias towards non-signalling risks. Model development will include the comparison of several algorithms, following the findings of related work, including LASSO regression, XGBoost and random forest with class-balanced sampling. Models will be evaluated using stratified 5-fold cross-validation with AUPRC as the primary metric given the class imbalance.

In order to assess the realism of our synthetic data, we will validate our model using real wearable device data and examine its ability to predict injury events. Our current efforts are focused on refining the synthetic data generation framework before moving on to full model development, so this stage may be subject to refinement.

V. Considerations and Future Work

This paper presents a novel framework for injury prediction in triathletes by integrating wearable sensor data, lifestyle factors, and synthetic data generation. Unlike prior models, which focus mainly on training workload, our approach will provide a holistic analysis incorporating stress, sleep, and recovery patterns, improving the personalization and generalizability of injury risk predictions.

At this stage, we have established the foundational components for synthetic data generation, including athlete profile parametrization, training periodization modeling, and probabilistic injury risk simulation. These components allow for the creation of realistic datasets that mimic real-world athlete responses and deviations from planned training regimens.

The next phase of our research will focus on refining the synthetic data model to enhance physiological realism and validate its effectiveness against real-world wearable device data. We will systematically calibrate the synthetic dataset based on published triathlete performance distributions and assess model robustness through cross-validation techniques. Furthermore, we will explore optimal ML algorithms for injury prediction, comparing the effectiveness of LASSO regression, XGBoost, and random forest models.

In addition to methodological refinements in our approach, future work will address practical implementation challenges, such as real-time injury risk monitoring and integration with existing athlete tracking platforms. By bridging the gap between theoretical modeling and applied sports science, our research aims to contribute toward data-driven coaching strategies that enhance athlete performance while minimizing injury risk.

References

[1] Andersen, C. A., Clarsen, B., Johansen, T. V., & Engebretsen, L. (2013). High prevalence of overuse injury among iron-distance triathletes. British Journal of Sports Medicine, 47 (12), 857–861. https://doi.org/10.1136/bjsports-2013-092397

[2] Soligard, T., Schwellnus, M., Alonso, J.-M., Bahr, R., Clarsen, B., Dijkstra, H. P., ... & Engebretsen, L. (2016). 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, 50 (17), 1030–1041. https://doi.org/10.1136/bjsports-2016-096581

[3] Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. PloS One, 13 (7), e0201264.

[4] Ayala, F., López-Valenciano, A., Gámez-Martín, J. A., De Ste Croix, M., Vera-Garcia, F. J., Garcia-Vaquero, M. P., Ruiz-Pérez, I., & Myer, G. D. (2019). A preventive model for hamstring injuries in professional soccer: Learning algorithms. International Journal of Sports Medicine, 40 (05), 344–353.

[5] Naglah, A., Khalifa, F., Mahmoud, A., Ghazal, M., Jones, P., Murray, T., Elmaghraby, A. S., & El-Baz, A. (2018). Athlete-customized injury prediction using training load statistical records and machine learning. In 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (pp. 459–464). IEEE.

[6] Lövdal, S. S., Den Hartigh, R. J. R., & Azzopardi, G. (2021). Injury prediction in competitive runners with machine learning. International Journal of Sports Physiology and Performance, 16 (10), 1522–1531.

[7] Hohl, B., Satizábal, H. F., & Perez-Uribe, A. (2024). Unveiling the potential of synthetic data in sports science: A comparative study of generative methods. In International Conference on Artificial Neural Networks (pp. 162–175). Springer.

[8] Kienstra, C. M., Asken, T. R., Garcia, J. D., Lara, V., & Best, T. M. (2017). Triathlon injuries: Transitioning from prevalence to prediction and prevention. Current Sports Medicine Reports, 16 (6), 397–403.

[9] Rothschild, J. A., Stewart, T., Kilding, A. E., & Plews, D. J. (2024). Predicting daily recovery during long-term endurance training using machine learning analysis. European Journal of Applied Physiology, 1–12. Springer.

[10] Chen, Q. (2024). Identification of potential injury risk factors and prediction model construction of athletes using data mining algorithm. Journal of Electrical Systems, 20 (6s), 1048–2058.

[11] Halson, S. L. (2014). Monitoring training load to understand fatigue in athletes. Sports Medicine, 44 (Suppl 2), 139–147. Springer.

[12] Lange, L., Wenzlitschke, N., & Rahm, E. (2024). Generating synthetic health sensor data for privacy-preserving wearable stress detection. Sensors, 24 (10), 3052. MDPI.