UNIVERSITY OF ST.GALLEN

School of Computer Science

Bachelor Thesis

Preventing Injuries in Triathletes with Machine Learning using Wearable Sensor Data

Submitted by:

Leonardo Rossi

22-618-482

Approved on Application by:

Professor:

Prof. Dr. Bruno Rodrigues

Date of Submission:

15/May/2025

Abstract

Wearable devices now allow endurance athletes to collect continuous physiological data that reveals critical insights into training responses and emerging injury risk. However, interpreting this abundant data is not always trivial. Integrating it with predictive machine learning systems presents a major opportunity to advance injury prevention. A major restriction are privacy restrictions in health and sports science domains, impeding large-scale access to athlete data, essential for training high-quality predictive models. This limitation creates a bottleneck in deploying effective early-warning systems for injury prevention. This thesis proposes a novel solution through the creation of a synthetic data generation framework. The objective is to simulate physiologically realistic athlete datasets that reflect true injury progression patterns, enabling the development and initial training of predictive models even when real-world data is scarce. The framework begins by generating a diverse synthetic population of triathletes with physiologically plausible profiles. Each athlete receives a personalized annual training plan incorporating load management and periodization principles. Daily training execution, recovery dynamics, and physiological responses are then simulated, accounting for stochastic variability and the influence of lifestyle stressors. Injury events are modeled as emerging from cumulative physiological strain patterns. Machine learning models (LASSO, Random Forest, and XGBoost) are trained and evaluated using both time-based and athlete-based data partitioning strategies. Models trained on synthetic data successfully identified injury risk patterns, with time-based splits consistently outperforming athlete-based approaches. The findings validate the synthetic data methodology and open broader opportunities for applying this approach to other health domains where privacy and data scarcity remain significant challenges.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Bruno Rodrigues, for his invaluable guidance, insightful feedback, and consistent support throughout this research journey. His experience in academic writing provided a solid foundation for my work, while his initiative to submit my preliminary findings to the IEEE SmartComp conference both improved the quality of this thesis and presented me with meaningful challenges that kept me motivated. Prof. Rodrigues' mentorship combined technical guidance with thoughtful questions that pushed me to refine my ideas and strengthen my research approach.

I also extend my appreciation to B.J. Leeper and Guy Shechter from TriDot for the numerous meetings and discussions that, although not culminating in a formal data collaboration, provided crucial insights into the practical needs of the coaching industry. These conversations helped refine my research objectives and ensured the work remained grounded in real-world applications.

Finally, I wish to acknowledge the support of my family and friends who provided encouragement throughout the process of researching and writing this thesis.

Contents

List of Figures

List of Tables

List of Algorithms

Listings

1 Introduction

Triathlon, a demanding multi-sport discipline combining swimming, cycling, and running, requires exceptional physical resilience and high training volumes across all three disciplines. The intense nature of triathlon training places athletes at significant risk of overuse injuries. This risk is well-documented, as seen in [5], where 56% of 174 participants in the Norseman Xtreme Triathlon reported overuse injuries during their preparation for the event's grueling 3.8 km swim, 180 km cycle, and 42 km 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 [6]. By recognizing patterns associated with overtraining and, therefore, injury risk, athletes and coaches can implement data-driven adjustments to training protocols. These adjustments, when properly timed, have the potential to prevent injuries before they manifest clinically, preserving training continuity and performance development.

Over the past decade, endurance sports training has undergone a *data-driven transformation*. Wearable technology is central to this shift, revolutionizing performance monitoring beyond traditional metrics such as stopwatch times and perceived exertion. Smartwatches, fitness trackers, and sensor-equipped devices now offer continuous monitoring of a broad range of physiological conditions, such as heart rate variability (HRV), resting heart rate, sleep patterns, training load, and recovery metrics—capabilities that were once exclusive to elite athletes and research laboratories. This democratization of data collection presents an unprecedented opportunity for large-scale analysis of training patterns and their relationship to injury risk.

The application of machine learning (ML) to injury prediction represents a promising frontier in sports science. While ML has shown promise in sports like football [7] and running [8] for detecting workload-related risks based on wearable sensors, triathlon—a multi-sport context with unique physiological demands—remains underexplored. Furthermore, existing studies often focus narrowly on sport-specific metrics, neglecting broader factors such as sleep quality, stress levels, and daily habits that significantly influence recovery capacity and performance.

A critical limitation in advancing injury prediction research is the scarcity of comprehensive datasets. Collection of longitudinal training and injury data is hindered by ethical challenges and strict data privacy regulations in health and sports sciences [9]. Even when data access is granted, injury labels—essential for supervised learning approaches—are rare and often incompletely documented. This data limitation has substantially impeded the development of robust, generalizable injury prediction models.

1.1 Research Objectives

This thesis addresses these challenges through a novel approach: the creation of a comprehensive synthetic data simulation framework that generates realistic athlete training and physiological data with labelled injury events. The primary objectives of this research are:

1. To develop a simulation framework capable of generating physiologically plausible athlete data, including training loads, recovery metrics, and injury occurrences preceded by realistic physiological warning patterns
2. To ensure that the simulated training programs adhere to established sports science principles for load management, thereby creating a realistic environment where injury occurrence is predominantly influenced by recovery factors that often lie outside a coach's direct control
3. To implement and evaluate multiple machine learning algorithms for injury prediction using the synthetic dataset, with particular emphasis on identifying early warning signs that precede overuse injuries
4. To demonstrate how synthetic data generation can serve as a foundational methodology to overcome data privacy regulations and missing labels, enabling initial model deployment that facilitates real data collection for subsequent model refinement

1.2 Significance and Contribution

The primary contribution of this research is methodological: using synthetic data generation to overcome the significant obstacle of data privacy regulations and labelled data scarcity that currently hinder advancement in injury prediction. This approach has practical applications in the sports technology industry. For instance, companies focusing on training prescription in triathlon (such as for example TrainingPeaks or TriDot) often have access to extensive training and physiological data through APIs (*e.g.,* Garmin Connect), but typically lack injury labels necessary for supervised learning.

The methodology proposed in this thesis could be adopted by such companies to establish initial injury prediction capabilities based on synthetic but physiologically plausible data. As these companies begin collecting user-reported injury data, they could progressively incorporate this real-world information to refine and retrain the models initially developed using synthetic data. This represents a pragmatic pathway to overcome the ”cold start” problem in injury prediction, where the lack of labelled data prevents the development of models that could ultimately help collect better data.

Additionally, by focusing specifically on triathlon—a multi-sport discipline with complex physiological demands—this research addresses a gap in the current literature, which has predominantly focused on single-sport applications. The insights gained may provide valuable cross-disciplinary understanding applicable to other endurance sports with high training volumes.

1.3 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 establishes the theoretical foundation necessary for this thesis, examining wearable device technology in endurance sports, the physiological basis of the interplay between recovery and performance, core machine learning concepts, and the principles of synthetic data generation.

Chapter 3 reviews the state of the art through a structured analysis of related literature. It begins with studies examining the relationship between training load, recovery metrics, and injury risk in endurance athletes. It then progresses to monitoring frameworks combining internal and external load assessments. Finally, it evaluates studies integrating machine learning with data for injury risk prediction

Chapter 4 presents the methodological approach in two main sections. The first details the design principles and implementation of the synthetic data generation framework, including athlete profile creation, training plan development, physiological response simulation, and injury pattern injection. The second section outlines the machine learning analysis methodology.

Chapter 5 provides a comprehensive evaluation of the synthetic data generation framework and machine learning analysis. It begins by validating the physiological plausibility of the synthetic data through distribution, correlation, and response pattern analyses. It then examines pre-injury patterns and evaluates the performance of multiple machine learning models across different data splitting strategies in recognizing these patterns.

Chapter 6 summarizes the key findings and contributions of this research, reflects on lessons learned throughout the research process, and suggests promising directions for future work to extend and refine this approach.

2 Background

This chapter establishes the theoretical foundation and necessary to understand the research presented in this thesis. It begins by exploring the evolution of wearable technology in sports, examining how these devices transform athlete monitoring through continuous data collection across physiological and performance metrics. Following this, the chapter introduces machine learning concepts and methodologies essential for extracting meaningful insights from the complex, multimodal data produced by wearable sensors.

It then continues with an examination of synthetic data generation techniques, addressing the critical challenges of data privacy and availability that often limit sports science research. This section builds progressively from basic random generation approaches to sophisticated generative models. Next, the chapter reviews critical data preprocessing methodologies that enhance data quality and model performance.

Finally, the discussion examines the complex relationship between performance and recovery in endurance sports, establishing the physiological basis that underlies effective training management. Together, these interconnected domains form the conceptual backbone of the research, providing necessary context for the synthetic data generation framework proposed in subsequent chapters.

2.1 Wearable Technology in Sports

The integration of wearable technology in endurance sports, has revolutionized how athletes and coaches approach training, competition, and recovery [10]. What was once limited to professional athletes with access to specialized laboratories is now continuously available to athletes of all levels through consumer wearable devices, requiring no active participation from medical staff or specialists [10]. Modern smart devices, predominantly worn as wrist-based devices like Garmin watches, continuously capture vast amounts of physiological and performance data throughout an athlete's day, from intensive training sessions to daily activities and sleep [10]. This comprehensive monitoring enables evidence-based decision-making for optimizing both performance and recovery. Following [10], wearable devices can be classified into three main categories:

1. Location-Based Wearables (LBW): These devices rely on sensors like GPS, accelerometers, and gyroscopes to collect data for measuring distance, speed, and pace.

2. Biometric Wearables (BMW): Using sensors such as photoplethysmography, electrocardiography, and bioimpedance sensors, these devices monitor heart rate, breathing rate, blood pressure, and other physiological metrics. Composites of tracked metrics can be used to assess stress levels and track sleep.

3. Performance Wearables (PMW): These devices track athlete's performance data, including cadence, speed, power output, and distance. They rely on similar sensors as LBW but focus specifically on performance metrics.

2.1.1 Core Components and Sensors

2.1.1.1 Microcontroller

At the foundation of wearable technology lies the microcontroller, essentially a computing device that is designed to perform a single functionality and is typically associated (embedded) into a larger product. Associated with sensing capabilities, microcontrollers enables larger cyber-physical ecosystems, such as the several Internet-of-Things (IoT) applications [11]. The microcontroller's architecture comprises several key components [11], as illustrated in Figure 2.1:

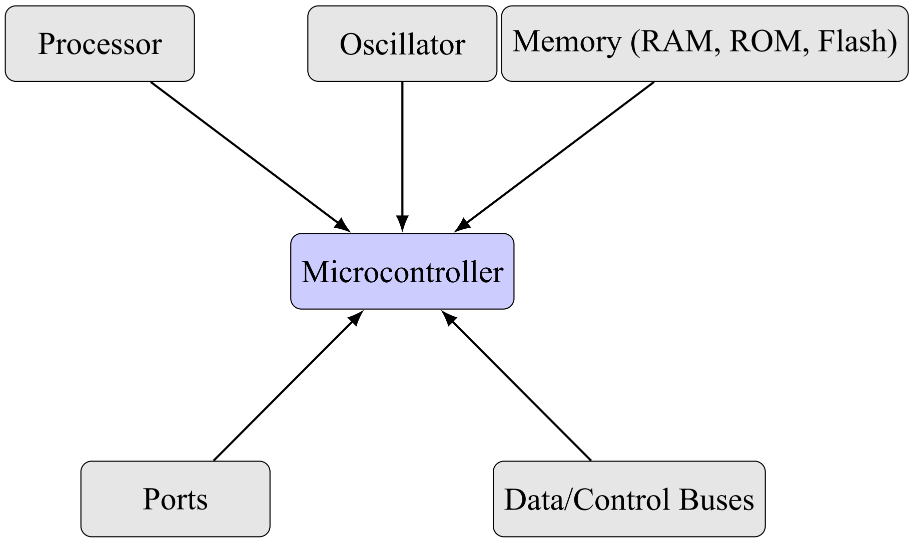

Figure 2.1: Architecture of a Microcontroller in Wearable Devices

- Processor: Reads input data and processes information according to programmed instructions
- Oscillator: Provides timing synchronization for all operations
- Memory: Includes RAM (volatile), ROM (non-volatile), and flash storage
- Ports: Connect the processor to the outside world, by handling input/output connections
- Buses: Transport data and control signals

2.1.1.2 Accelerometer

Accelerometers serve as fundamental sensors in wearable devices, measuring both linear and gravitational acceleration [11]. They operate using two primary mechanisms: the piezoelectric effect and capacitive sensing [11]. In the piezoelectric effect, kinetic movement is converted into electrical signals through stressed crystal structures, which generate a voltage proportional to the applied force [11]. On the other hand, capacitive sensing measures changes in capacitance between microstructures as they move relative to each other, allowing precise detection of acceleration [11]. These sensors enable tracking of movement patterns, speed calculations, and even sleep monitoring, making them versatile tools for both sports and medical applications [11].

2.1.1.3 Gyroscope

Gyroscopes complement accelerometers by specifically measuring angular acceleration. They operate primarily through mechanical sensing, which uses angular momentum principles to detect changes in orientation via a spinning disk. Optical sensing utilizes the Sagnac effect [12], where two fiber optic coils spin in opposite directions; the difference in the distance travelled by light through these coils indicates rotational movement [11]. Vibration sensing detects angular velocity through the Coriolis force, which acts on a vibrating element within the sensor. This force, caused by the earth's rotation, creates a shift in the vibration pattern, generating a potential difference that is converted into an electrical signal for precise motion tracking [11]. The combination of accelerometers and gyroscopes provides comprehensive motion tracking capabilities, essential for analysing complex athletic movements [11].

2.1.1.4 Global Positioning System (GPS)

GPS technology enables precise location tracking through satellite communication, offering valuable insights into various performance metrics [11]. It allows for the measurement of distance covered, the analysis of movement patterns, and positional play assessment, which is particularly useful in team sports for tactical evaluation [11]. Additionally, GPS provides speed and acceleration metrics, aiding in performance monitoring and optimization [11]. Despite its advantages, GPS technology poses challenges related to power consumption, which must be carefully considered in wearable device design [11].

2.1.1.5 Heart Rate Sensors

Heart rate monitoring in wearable devices employs different technologies, each with distinct mechanisms. Capacitive sensing utilizes skin-electrode interfaces to detect electrical signals from the heart [11]. Photoplethysmography (PPG), a widely used optical method, measures blood flow by analyzing light absorption variations in tissue. More advanced forms of PPG leverage multiple wavelengths, such as green and red light, to extract additional metrics like oxygen saturation, providing deeper physiological insights [11]. These technologies play a crucial role in health and fitness tracking, ensuring real-time monitoring of cardiovascular performance [11].

2.1.2 Limitations and Challenges

While wearable technologies offer substantial benefits for sports performance and recovery monitoring, several significant limitations must be considered.

Table 2.1: Limitations and Challenges of Wearable Technologies

Challenge	**Description**	**Examples**
Validity	Accuracy of measurements may vary depending on activity type, conditions, or sensor limitations.	Heart rate monitors may provide unreliable data during high-intensity activities [13].
Reliability	Data consistency can be affected by environmental factors, physiological conditions, or device placement.	GPS signals degrade in urban areas; biometric readings are affected by skin moisture or body movement [14].
Interpretability	Large volumes of data can be difficult to analyze, limiting actionable insights.	Complex datasets can overwhelm users; difficulty in translating data into practical strategies [15].

Table 2.1 highlights the key limitations and challenges faced by wearable technologies in sports performance and recovery monitoring. Each challenge — validity, reliability, and interpretability — is described alongside examples.

Validity refers to the accuracy of data collected by wearable devices. For instance, heart rate monitors may provide less reliable readings during high-intensity activities due to motion artifacts or poor sensor contact [13].

Reliability concerns the consistency of data across various environments and physiological conditions. For example, GPS signals often lose precision in urban environments due to signal reflection from buildings, while biometric sensors may be affected by skin moisture or device movement [14].

Interpretability addresses the challenge of managing and understanding the vast amount of data generated by wearables. Users may find complex datasets overwhelming, making it difficult to derive actionable insights [15]. To overcome this, developers focus on simplified data visualization, creating user-friendly interfaces, and employing automated interpretation tools that translate raw data into meaningful feedback for athletes and coaches.

These limitations collectively highlight the critical need for advanced machine learning systems that can interpret complex wearable data and translate it into actionable injury risk insights for triathletes. While individual sensors provide valuable data points, their isolated use often leads to fragmented insights, and the sheer volume and complexity of multimodal sensor data exceed human analytical capabilities. Athletes and coaches need intelligent systems that can process these diverse inputs—heart rate variability, GPS metrics, sleep quality indicators, and other relevant parameters—and synthesize them into comprehensible risk assessments.

2.2 Machine Learning in Sports Science

The widespread use of wearable devices in sports has led to an exponential increase in data availability. However, this data overload poses a challenge for researchers, coaches, and athletes, making it difficult to extract meaningful insights efficiently [16]. Machine Learning (ML) provides a powerful solution by automating data analysis, recognizing complex patterns, and enabling predictive modelling in sports performance and athlete monitoring.

ML is a subfield of Artificial Intelligence (AI) in which an algorithm learns from past data to make informed predictions about future events [17]. At its core, ML focuses on extracting patterns from existing datasets, transforming raw data into actionable knowledge, and generalizing this knowledge to previously unseen data [18]. By leveraging ML techniques, sports science can move beyond descriptive statistics and traditional heuristics, adopting data-driven decision-making to enhance training, injury prevention, and overall athletic performance [19].

2.2.1 Types of Machine Learning

Machine Learning (ML) methods can be broadly categorized based on the availability of feedback during the learning process. The three primary types of learning are [17]:

- **Unsupervised Learning:** No labelled data. The goal is to uncover patterns or structures in the data, see Figure 2.2.
- **Reinforcement Learning (RL):** An agent learns by interacting with the environment and receiving rewards or penalties, consult Figure 2.3.
- **Supervised Learning:** The model learns from labelled data to predict outputs for new, unseen inputs, this is illustrated in Figure 2.4.

In more detail, **Unsupervised Learning** deals with datasets where no labels are provided. The objective in this type of learning is to find hidden patterns, structures, or relationships in the data without any predefined categories. Techniques commonly used in unsupervised learning include clustering algorithms like k-means and DBSCAN, which group data based on similarities, and dimensionality reduction methods such as PCA, t-SNE, and autoencoders, which aim to reduce the complexity of the data while preserving essential information [17].

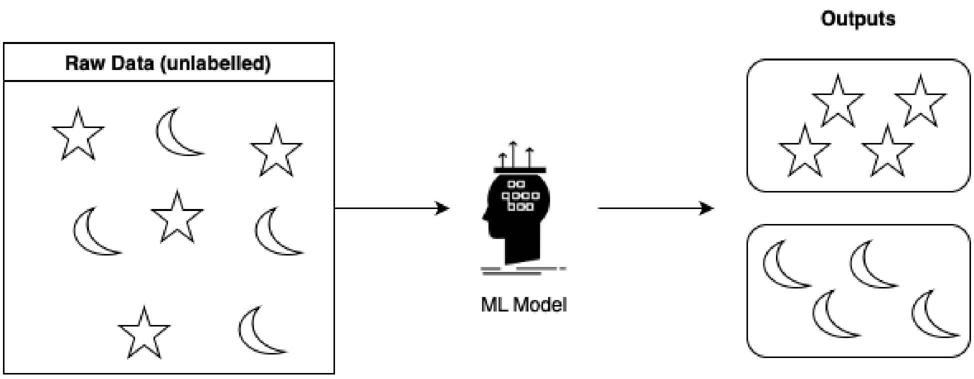

Figure 2.2: Unsupervised Learning

Reinforcement Learning (RL) involves an agent interacting with an environment, making decisions based on its observations. The agent receives feedback in the form of rewards or penalties depending on its actions. The goal of RL is for the agent to learn a policy that maximizes its cumulative reward over time. This problem is typically framed as a Markov Decision Process (MDP), with states, actions, transition probabilities, rewards, and a discount factor. Applications of RL are widespread in areas like game playing, robotics, and autonomous systems, where adaptive decision-making is key. For instance, a chess-playing agent learns by receiving rewards for winning games, gradually discovering strategies that maximize its chances of winning [17].

Supervised Learning requires labelled data, where each input x_j is paired with an associated

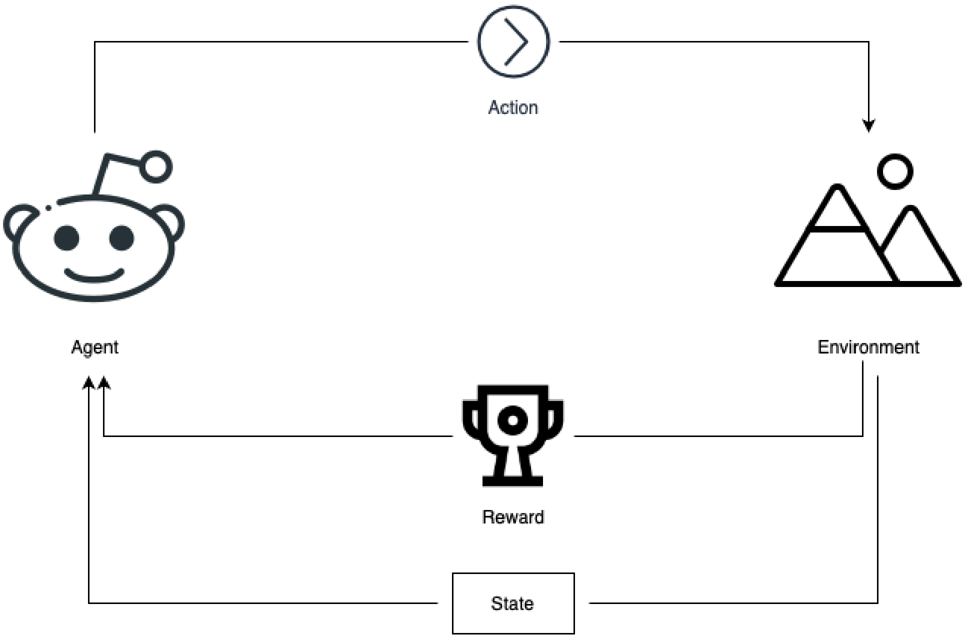

Figure 2.3: Reinforcement Learning

output y_j. The goal is to learn a function f that maps inputs to their corresponding outputs. Specifically, we are given a dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where each pair (x_j, y_j) is drawn from some unknown distribution. The objective is to approximate the target function $f(x)$ by finding a hypothesis function $h(x)$ such that:

$$h(x) \approx f(x) \quad \forall x \in X.$$

This is typically done by minimizing a loss function $L(h(x), y)$, where L measures the difference between the predicted output $h(x)$ and the true output y. Optimization techniques, such as Gradient Descent, are often used to minimize this loss function over the training set [17]. Supervised learning is commonly used for both classification tasks (*e.g.,* spam detection) and regression tasks (*e.g.,* predicting house prices), where the model learns to predict the output based on input data.

In real-world scenarios, a clear distinction between the three types introduced above is not always given [17]. **Semi-supervised learning** leverages both a small subset of labelled and a large subset of unlabelled samples to improve model performance when obtaining labelled data is expensive or time-consuming [20]. This approach is particularly useful in domains where manual annotation is difficult.

A common strategy in semi-supervised learning is to use self-training, where a model trained on labelled data makes predictions on unlabelled samples, which are then iteratively added to

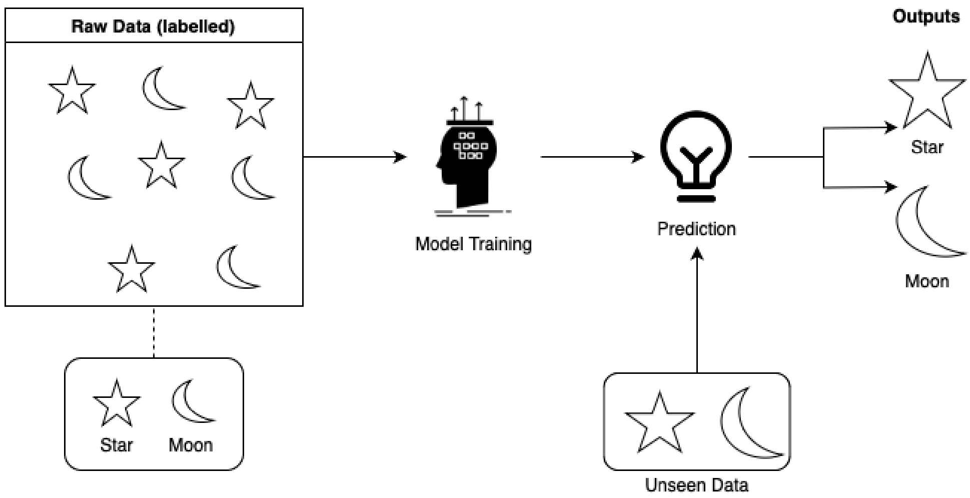

Figure 2.4: Supervised Learning

the training set. Other methods include co-training, where multiple models learn from different feature representations [21], and graph-based learning, which models relationships between data points using graph structures [22].

In sports science, semi-supervised learning can be applied to performance monitoring and fatigue prediction, where a limited number of biometric readings (*e.g.,* lactate threshold tests) are available, but large amounts of wearable sensor data remain unlabelled [19]. By integrating semi-supervised techniques, models can generalize better and make accurate predictions with minimal supervision [20].

2.3 Synthetic Data in Sports Science

As in any domain, developing a ML application in sports science requires large amounts of high quality data to ensure model accuracy and generalizability [9, 23]. However, obtaining such data is a significant challenge. Physiological and biomechanical data collected from wearable sensors often contain sensitive biological information that can reveal an individual's health status [23]. Even after anonymization, the risk of re-identification remains, raising ethical and privacy concerns that hinder the large-scale data collection and sharing necessary for robust research in sports science [9, 23].

To mitigate this challenge, synthetic data generation has emerged as a promising solution. Unlike real-world data, which is inherently constrained by privacy regulations, collection costs, and ethical considerations, synthetic data is artificially generated using mathematical models or algorithms. This approach replicates the statistical properties and structure of real data while ensuring privacy protection [24]. Beyond privacy protection, synthetic data can address other limitations in sports science datasets, such as correcting biases or class imbalances (such as injury incidence) that are often present in real-world collections [24].

There are many techniques to achieve realistic synthetic datasets, ranging from statistical methods to generative ML models [25]. This section aims to provide an overview of these methods, starting with more basic approaches and building up to more sophisticated solutions. Two studies discussed in chapter 3 compare different generative methods and evaluate the quality and performance of the synthetic data they produce.

2.3.1 Random Data Generation

Random data generation is the simplest approach to creating synthetic data. This method involves generating random values, inspired by real-world ranges, based on predefined distributions, such as uniform or normal distributions [25]. While it ensures complete data privacy, its random nature means that it lacks the structural patterns and statistical relationships found in real-world sports sensor data [9]. For example, randomly generated heart rate values may fall within physiologically possible ranges, but would fail to capture the correlation between exercise intensity and cardiac response [1]. Despite these limitations, random data generation serves as a useful baseline for evaluating more advanced synthetic data generation techniques [9].

2.3.2 Rule-Based Generation

Rule-based generation involves creating synthetic data using predefined rules and constraints derived from domain knowledge [27]. In the field of sports science, this approach incorporates sport-specific physiological principles and biomechanical constraints. For instance, rules might enforce that heart rate increases proportionally with power output during cycling, or that running cadence maintains a realistic relationship with speed. This method produces more realistic data than purely random generation by incorporating known relationships between variables. However, it is limited by the assumptions made when defining the rules and may not capture complex, real-world variability or unexpected patterns [27].

2.3.3 Simulation Based Generation

Simulation-based methods generate synthetic data by modelling real-world processes using mathematical or physics-based simulations [27]. In sports science, biomechanical and physiological simulations can generate synthetic sensor readings that reflect authentic athletic perform-

[1] In endurance training, heart rate is used to measure intensity. The three-zone model is a widely applied framework, based on ventilatory thresholds (VT). Zone 1 (low intensity) is exercise below VT1, performed at low heart rates. Zone 2 (moderate intensity) is between VT1 and VT2, where lactate accumulation begins. Zone 3 (high intensity) is above VT2, characterised by elevated heart rates due to increased cardiovascular demand [26].

ance across various conditions and scenarios. OpenSim, for example, is widely used to simulate human movement and predict musculoskeletal dynamics [28]. Researchers have employed OpenSim to generate synthetic motion capture data for analysing injury mechanisms in sports like running and jumping [29]. These methods provide high interpretability and physiological validity but may require extensive domain expertise and can be computationally expensive to develop and run [27].

2.3.4 Generative Models

Generative models, in particular Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have gained considerable attention for their ability to produce highly realistic synthetic data. These models learn the underlying distribution of real data and generate new samples that closely resemble authentic sensor readings [27].

2.3.4.1 Generative Adversarial Networks (GANs)

GANs were introduced by Goodfellow et al. in 2014 [30] as a framework for estimating generative models through an adversarial process. The architecture (cf. Figure 2.5) consists of two neural networks: a *generator* G that learns to produce realistic synthetic data from random noise and a *discriminator* D that learns to discriminate between real and synthetic data. These two networks are trained simultaneously in a minimax game, where G aims to maximize the probability of D misclassifying its outputs, while D attempts to correctly classify real and fake samples. The competition between these networks drives the generator to produce increasingly realistic data. GANs have shown remarkable success in generating synthetic images, text, and time-series data.

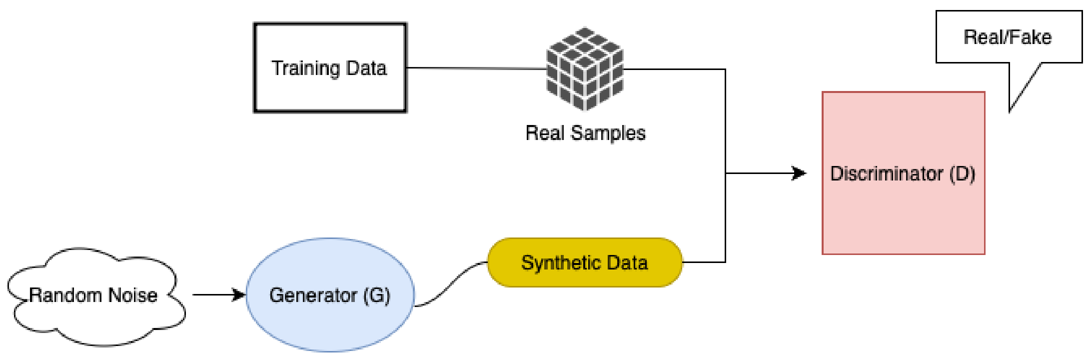

Figure 2.5: Generative Adversarial Networks Architecture

2.3.4.2 Variational Autoencoders (VAEs)

VAEs provide another deep learning-based approach to synthetic data generation. Unlike GANs, which learn via adversarial training, VAEs rely on probabilistic modelling and variational inference to map input data to a structured latent space [31]. The model consists of an *encoder*, that compresses data into a probabilistic latent representation, and a *decoder*, which reconstructs synthetic samples from this latent space (cf. Figure 2.6). A key feature of VAEs is the Stochastic Gradient Variational Bayes (SGVB) estimator, which allows efficient optimization using the reparametrization trick. This technique enables VAEs to sample from a smooth, continuous latent space, making them particularly well-suited for generating diverse yet realistic synthetic data.

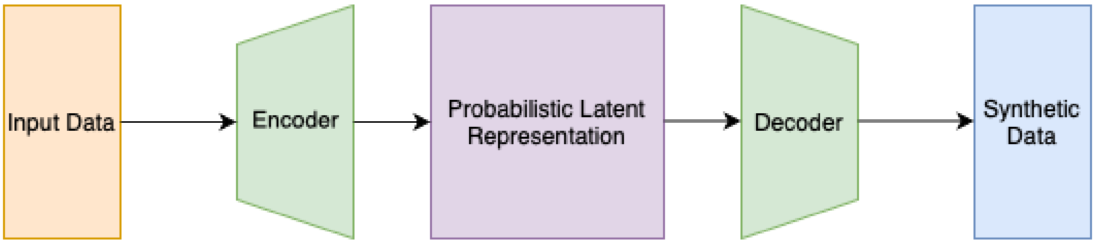

Figure 2.6: Variational Autoencoder Architecture

2.4 Data Preprocessing for Machine Learning

In a nutshell, the main task in ML is to select a learning algorithm and train it on some data, hence a major success indicator next to the applied algorithm is the quality of the data being used [32]. Data preprocessing is a critical step in the ML pipeline that significantly impacts the generalization performance of models [33]. Without proper preprocessing, models may struggle with irrelevant features, redundant information, and noisy data, leading to poor performance and unreliable predictions. To understand the importance of data preprocessing, it is essential to recognize the potential risks associated with poor data quality, such as:

- **Insufficient Quantity of Training Data:** To make accurate predictions, ML algorithms need thousands of examples [32]. It has been shown that different ML algorithms performed similarly well on natural language disambiguation, once they were fed enough data [1], cf. Figure 2.7.

- **Non-representive Training Data:** For a machine learning model to generalize effectively, the training data must accurately reflect the range of cases it will encounter in real-world scenarios [32]. When the training set is non-representative – whether due to sampling bias or sampling noise – the model may learn patterns that do not hold true for new data, leading to poor predictive performance [32]. For example, training a linear model on incomplete or biased data can result in misleading conclusions, such as incorrectly assuming that wealth consistently correlates with happiness across all countries [32].

- **Poor-Quality Data:** If the data used to train a machine learning algorithm contains errors,

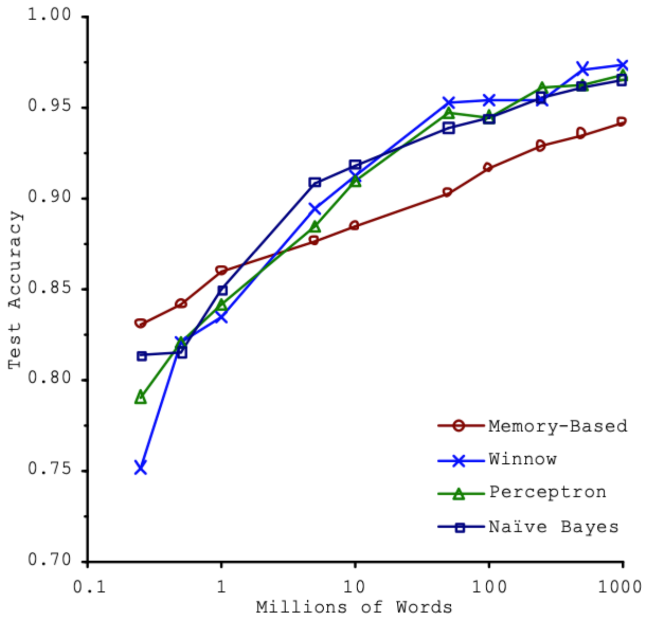

Figure 2.7: The importance of data versus algorithms [1]

outliers, or noise, it becomes significantly more challenging for the model to identify accurate patterns, increasing the likelihood of learning incorrect or misleading relationships [32]. Poor-quality data can stem from various sources, including inaccurate measurements, data entry errors, or malfunctioning sensors. Additionally, missing features can degrade model performance, as incomplete data may lead to biased or inconsistent predictions. For instance, if 5% of participants fail to wear their wearable devices overnight, resulting in missing sleep data, this gap could introduce variability and noise into the model, potentially skewing outcomes [32].

- **Irrelevant Features:** A machine learning model can only perform well if the training data contains relevant and informative features [32]. If the dataset includes too many irrelevant features or lacks critical information, the model may struggle to identify meaningful patterns, leading to poor generalization and inaccurate predictions [32].

Data preprocessing encompasses several critical techniques that transform raw data into a format suitable for machine learning algorithms. These steps help enhance data quality, reduce noise, and improve model accuracy [33].

2.4.1 Data Cleaning

Raw data often contains errors, noise, irrelevant features, and inconsistencies, all of which can undermine the reliability of a model. Figure 2.8 highlights key aspects of data cleaning, including identifying and correcting missing data, unwanted observations, duplicates, and outliers [2]. Unwanted features, which do not contribute to the analysis, should be removed to reduce noise and enhance learning efficiency [2]. While duplicates and irrelevant observations are easily addressed, handling missing data and outliers requires more advanced methods for optimal results [2].

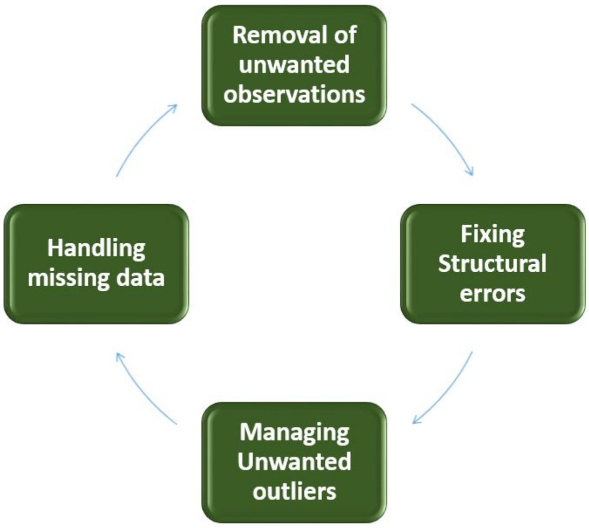

Figure 2.8: Components of Data Cleaning [2]

2.4.1.1 Missing Data

When handling missing data, it is crucial to consider the underlying reasons for its absence. Missing values may result from forgetfulness, non-existence for certain instances, or deliberate omission during dataset creation [33]. Depending on the situation, the following strategies can be employed [32]:

- Discard instances with missing values
- Ignore the missing attribute across all instances
- Impute missing values with zeros, mean, median, or the most frequent value

2.4.1.2 Outliers

Outliers are data points that deviate significantly from the mean value. It must be decided whether affected instances are ignored or transformed to minimize their impact on the analysis [2]. Outliers can be detected using variable-by-variable cleaning, where values deviating from expected distributions are flagged as suspicious [33]. For instance a feature follows a normal distribution with a mean of 5 and a standard deviation of 3, then 10 is a suspicious value [33].

However, some outliers, known as inliers, fall within the distribution but are still erroneous and harder to detect. Multivariate methods like RT (distance-based) and LOF (density-based) help in identifying these subtle outliers [33].

2.4.2 Discretization

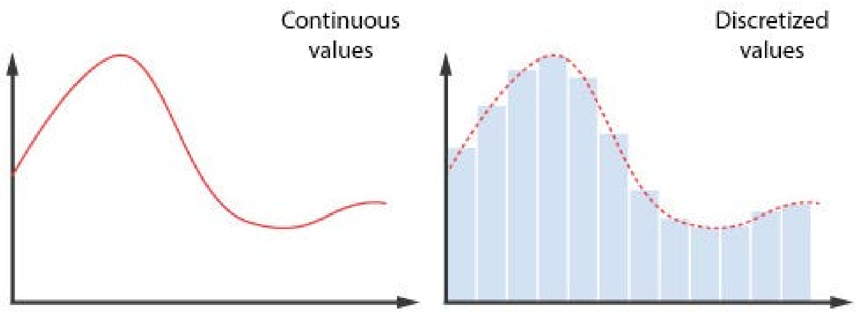

Figure 2.9: Data Discretization [3]

Some features in datasets can take an infinite number of continuous values, which can slow down machine learning processes and reduce efficiency [33]. Discretization addresses this by converting continuous values into a finite set of intervals while minimizing information loss [34]. This transformation is illustrated in Figure 2.9. However, determining the optimal interval boundaries to preserve data integrity remains a challenge. To tackle this, various algorithms exist, broadly categorized into supervised methods (which incorporate class labels) and unsupervised methods (which do not) [33].

Additionally, most discretization techniques follow either a bottom-up or top-down approach. In the bottom-up approach, the process begins with individual data points treated as single-value intervals, which are then progressively merged into broader intervals. Conversely, the top-down approach starts with the full data range and recursively splits it into smaller intervals [33].

An example of a supervised, top-down approach is equal-width discretization. This method calculates the minimum and maximum values of a feature and partitions this range into k equal intervals. For instance, if a feature ranges from 0 to 100, it can be divided into five intervals: 0–20, 21–40, 41–60, 61–80, and 81–100. While straightforward, this method may lead to unbalanced intervals if the original data distribution is skewed, potentially reducing the effectiveness of the model [33].

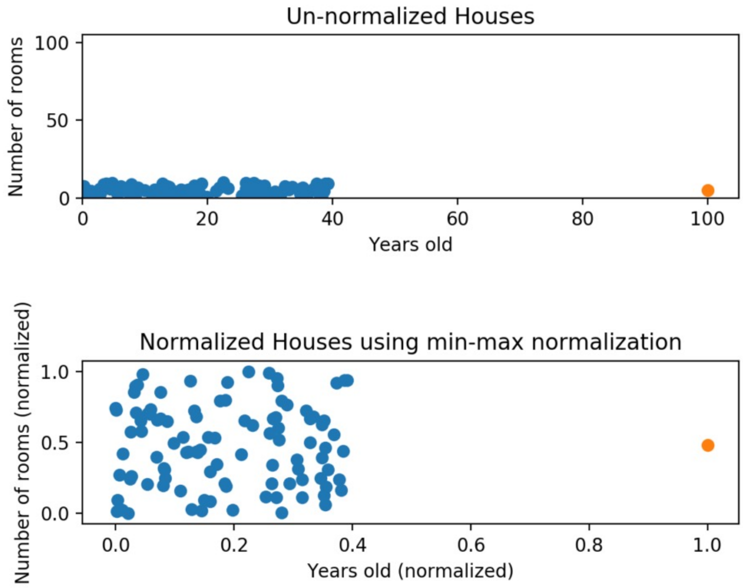

Figure 2.10: Data Normalization using Min-Max scaling [4]

2.4.3 Data Normalization

Most machine learning (ML) algorithms struggle to perform optimally when input features have vastly different numerical scales [32]. Features with larger scales can disproportionately influence the model, overshadowing smaller-scale features and leading to biased results [35]. To address this, normalization is employed-a preprocessing technique that scales features to ensure each contributes equally to the learning process. This transformation, illustrated in Figure 2.10, is critical for improving both model performance and convergence speed [33, 35].

The two most widely used normalization techniques are Min-Max Scaling and Z-Score Normalization (also known as standardization) [32, 33]:

- **Min-Max scaling:** features are rescaled to a fixed range, typically [0, 1]:

$$X' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

  where:

  - X is the original value,
  - X_{min} and X_{max} are the minimum and maximum values of the feature,
  - X' is the normalized value.

  This technique is sensitive to outliers, as extreme values can compress the majority of data points into a narrow range.

- **Z-Score scaling:** standardizes features to have a mean of 0 and standard deviation of 1:

$$X' = \frac{X - \mu}{\sigma}$$

  where:

  - X is the original value,
  - μ is the mean of the feature,
  - σ is the standard deviation of the feature,
  - X' is the standardized value.

  Z-Score scaling does not strictly confine values to a specific range, which can pose challenges for certain algorithms (*e.g.,* neural networks). However, this technique is generally less sensitive to the influence of outliers.

2.4.4 Feature Engineering

Feature engineering is a critical step in the machine learning (ML) pipeline, as the quality and relevance of the features significantly influence the model's performance. It involves transforming raw data into meaningful inputs that enhance the learning process. A well-engineered set of features can often outperform more complex algorithms trained on poorly selected data [32]. Feature engineering consists of several key processes [32]:

- *Feature Selection:* This process involves identifying and retaining the most relevant features while removing redundant or irrelevant ones.
- *Feature Extraction:* Feature extraction transforms raw data into new feature spaces, often reducing dimensionality while preserving critical information. This is particularly useful in scenarios with large datasets or highly correlated variables.
- *Feature Creation:* This step involves generating new features from existing data to better represent the underlying patterns. Effective feature creation can reveal relationships that are not immediately apparent in the raw data.

Effective feature engineering not only improves model accuracy but also contributes to better generalization on unseen data. It requires a combination of domain knowledge, statistical analysis, and iterative experimentation to identify the features that contribute most significantly to predictive performance [32].

2.5 Recovery and Performance in Endurance Sports

The intricate relationship between recovery and performance has been a focal point in sports science for many years [36]. While increased training intensity and volume are essential for physiological adaptations, failure to incorporate proper recovery strategies can lead to fatigue accumulation, injuries, immune suppression, and psychological burnout [37]. Monitoring and managing the stress-recovery balance is therefore crucial in endurance sports, particularly in triathlon, where athletes train across multiple disciplines (swimming, cycling, and running) [37].

2.5.1 The Role of Recovery in Endurance Performance

Recovery in endurance sports is a multifaceted process encompassing both physiological and psychological components. Physiologically, recovery aims to restore the body's homeostasis following intense training or competition [36]. Key strategies include:

- **Proactive Recovery:** Focuses on choosing activities that help managing mental fatigue, such as social activities or relaxation exercises [36].
- **Active Recovery:** Involves low-intensity exercises like cool-down jogging or swimming to enhance blood flow, facilitating the removal of metabolic byproducts such as lactic acid [36].
- **Passive Recovery:** Includes complete rest or external interventions like compression garments or massage to promote muscle relaxation without physical exertion [36].

2.5.2 Fatigue and Overtraining Risks

While a certain degree of fatigue is necessary for performance improvements (functional overreaching), chronic fatigue without adequate recovery leads to detrimental effects such as underrecovery, nonfunctional overreaching (NFO), and eventually overtraining syndrome (OTS) [36]. NFO is characterized by persistent fatigue, hormonal imbalances, and declining performance, and if left unaddressed, can evolve into OTS, marked by prolonged physical symptoms, psychological disturbances, and reduced athletic output [36].

2.5.3 Monitoring Recovery in Endurance Athletes

Monitoring recovery is essential to optimize performance and prevent overtraining. A comprehensive approach includes both physiological and psychological assessments [36]:

- **Physiological Markers:**

- *Heart Rate Variability (HRV):* Non-invasive monitoring of the autonomic nervous system to assess recovery status.
- *Blood-Based Markers:* Measuring biomarkers such as creatine kinase, urea nitrogen, and cortisol to evaluate muscle damage and stress responses.
- *Performance Metrics:* Sub maximal performance tests, like time-to-exhaustion trials, serve as practical indicators of recovery.

- **Psychological Assessments:**
  - *Rating of Perceived Exertion (RPE):* Subjective self-reports on training intensity and fatigue.
  - *Mood State Questionnaires:* Tools like the Profile of Mood States (POMS) and the Recovery-Stress Questionnaire for Athletes provide insights into the athlete's mental recovery [36].

Wearable devices now allow athletes and coaches to monitor recovery in real time. HRV as well as sleep and other metrics can be tracked by these devices facilitating data-driven decision-making [38]. For example, the Dutch women's field hockey team has successfully implemented heart rate and GPS monitoring to optimise recovery and training adjustments [39].

2.5.4 Practical Implications

Coaches and athletes need to take a holistic approach to recovery management, integrating physiological, psychological and lifestyle factors. This includes [36]:

- Implementing regular monitoring routines to identify early signs of under-recovery.
- Educate athletes on the importance of sleep, nutrition and mental health for recovery.
- Adjust training loads based on individual recovery responses to prevent overtraining and ensure sustained performance gains.

Recovery is a dynamic and individual process that plays a central role in endurance sport performance. Traditional recovery assessment methods, while valuable, are increasingly being complemented by wearable technology. These innovations provide real-time, personalised insights that enable athletes and coaches to make evidence-based decisions that improve performance, prevent overtraining and reduce the risk of injury.

3 Related Work

Injury prevention and maintaining an optimal balance between training stress and recovery have been central topics in sports science for decades. Researchers have extensively studied the interplay between training load, recovery, and injury risk, seeking to identify patterns that lead to both peak performance and injury prevention. Monitoring these factors has evolved from simple self-reported training logs to sophisticated real-time tracking systems. The emergence of wearable technology and machine learning (ML) has opened new avenues for more automated and less invasive athlete monitoring. These technologies enable continuous data collection of physiological and biomechanical parameters, allowing for deeper insights into athletes' training responses and recovery states. By integrating ML algorithms with detailed biometric data, researchers have developed predictive models aimed at assessing injury risk and optimising recovery.

In this chapter, we review key contributions in this area, organized into three main areas. First, we discuss foundational studies that examine the relationship between training load, biomechanics and injury risk. Next, we review research that introduces monitoring frameworks that combine internal and external training load assessments. Finally, we review studies that integrate ML techniques with physiological and biomechanical data to improve injury risk prediction, leading to the gap that this work aims to address. Additionally, section 3.1 delves into two related studies that compare different synthetic data generation methods.

A summary of these contributions is presented in Table 3.1, which provides a concise overview of each study's methodology, dataset, key findings, and limitations. This table serves as a comparative reference, highlighting the strengths and gaps in existing research relevant to injury prediction in endurance sports.

Table 3.1: Summary of Related Work

Study	Methodology	Data	Findings	Limitations
[40]	Review of risk factors for injuries in triathletes and theories on training volume assessment and injury risk	Literature review	Highlights the role of external and internal load in injury development. Suggests monitoring acute-to-chronic workload ratios (0.8–1.35) to reduce injury risk.	Workload ratio validation is limited to other sports; no direct empirical validation for triathletes.
[41]	Framework for monitoring training load and fatigue in athletes	Literature review and survey data from high-performance programs	Emphasizes the importance of both external and internal load monitoring. Identifies six key monitoring tools and highlights the varying effectiveness of RPE and heart rate.	Does not provide new empirical data; focuses on existing monitoring tools without integrating predictive modelling.
[42]	Machine learning analysis of athlete recovery states	Medical data from 3,661 athletes, including blood and urine biomarkers	Identifies key biochemical markers for recovery effectiveness using random forest and multinomial logistic regression. Demonstrates that machine learning can classify metabolic states.	Does not incorporate biomechanical or training load data, limiting its use for real-time injury prediction.

Continued on next page

Table 3.1 Continued

Study	Methodology	Data	Findings	Limitations
[43]	Machine learning analysis of recovery prediction in endurance athletes	3,572 days of tracking data from 43 athletes (67.3% triathletes) over 12 weeks, collected via wearables and mobile apps	Found that LASSO models outperformed complex algorithms. Introduced Markov unfolding with a 7-day look-back window to handle time-series dependencies.	Unable to standardize external load metrics across sports. Limited to a single internal training load measure (RPE). Self-reported data may introduce quality issues.
[7]	Decision tree classifier with recursive feature elimination (RFECV) and adaptive synthetic sampling (ADASYN) for class imbalance correction	952 training sessions from 26 professional football players over 23 weeks, capturing 55 workload variables (kinematic, metabolic, and mechanical)	Decision tree achieved 80% recall and 50% precision. Identified prior injury, high-speed running distance, and total distance monotony as key predictors	Specific to football's movement patterns and injury mechanisms, limiting transferability to multidisciplinary endurance sports like triathlon
[44]	Hierarchical three-phase ML framework: (1) feature selection, (2) K-Means/KNN clustering for injury-related feature identification, and (3) SVM classifier trained on balanced datasets	21 football players tracked over 16 months, with 36 non-contact injuries; 65 features from GPS sensors, accelerometers, gyroscopes, and wellness questionnaires	SVM classifier achieved 76% prediction accuracy; generated personalised probabilistic injury maps enabling athlete-specific risk assessment and training load optimization	Football-specific movement patterns and team sport dynamics limit applicability to individual endurance sports

Continued on next page

Table 3.1 Continued

Study	Methodology	Data	Findings	Limitations
[45]	Machine learning-based hamstring injury prediction using multi-factorial risk assessment	96 professional football players tracked over a full season, with 229 pre-season screening variables and 18 recorded hamstring injuries	The SmoteBoostM1 ensemble with an Alternating Decision Tree (ADTree) classifier achieved an AUC of 0.837, identifying prior injury history, sleep quality, and hip flexion ROM as key predictors	High model complexity (10 classifiers, 66 predictors) may limit practical application. Generalizability to other sports and competition levels remains uncertain
[8]	Bagged XGBoost model for injury risk prediction in competitive runners	Dutch elite running team (74 athletes over 7 years)	Daily models outperform weekly composite models	Lacks objective recovery data, requires invasive perception reports, potential overfitting (single team, same coach, training methodology, nutrition, etc.)
[46]	Noninvasive Data Mining Model with Multi-relational and Multi-dimensional Clustering (NiDMMMC) algorithm, employing hierarchical clustering with dynamic similarity measures	Biomechanical, physiological, training load, and environmental data from 1000 athletes across multiple sports, collected over 18 months	Successfully categorized athletes into distinct risk profiles based on multidimensional factors, improving injury prediction accuracy by 23% compared to traditional methods	Limited transparency regarding data collection protocols and feature extraction methods

Kienstra et al. [40] explored the complex and non-linear nature of injury prediction in triathletes, emphasizing the dynamic interaction of multiple risk factors over time. Their study highlights the importance of both external load (*e.g.,* distance, time) and internal responses (*e.g.,* perceived exertion, fatigue, heart rate) in injury development. They identify biomechanical adaptations to fatigue, such as reduced heel lift, increased ground contact time and altered load transfer, which correlate with common injuries such as patellofemoral pain and stress fractures. A key takeaway from their work is the relationship between training load and tissue adaptation: injuries occur either when acute loads exceed tissue capacity or when chronic loads surpass the body's ability to recover. They suggest monitoring the acute/chronic workload ratio and recommend that athletes maintain a ratio between 0.8 and 1.35 to optimise performance while minimising the risk of injury. While this ratio has been validated in other sports, its specific applicability to triathletes remains unconfirmed. Additionally, they review prior research on injury prevalence, offering insights into relevant data features for injury prediction. Their findings underscore the potential of continuous monitoring systems capable of tracking both external loads and biomechanical adaptations to predict injury risk.

Building on this foundation, Halson [41] provided a comprehensive framework for monitoring training load and fatigue in athletes, emphasizing that while external load metrics are valuable, internal load monitoring is essential for understanding individual training responses. The study categorizes six primary monitoring tools: perception of effort, heart rate measures, biochemical markers, hormonal assessments, immunological parameters, and psychological questionnaires. Notably, the research demonstrates strong correlations between rating of perceived exertion (RPE) and heart rate during steady-state exercise but weaker correlations in high-intensity interval training, suggesting that certain training loads require different monitoring approaches. Additionally, [41] reports that 91% of high-performance programs surveyed implemented some form of training monitoring, with self-report questionnaires being the most common (84%), followed by performance tests (61%). This study reinforces [40]'s findings by showing that monitoring both external and internal loads can reveal early signs of maladaptation.

With the advent of ML in sports science, researchers have explored data-driven approaches to enhance injury prediction models. Petrovsky et al. [42] applied ML techniques to analyse medical data from 3,661 athletes, focusing on classifying them into catabolic and anabolic phenotypes using biochemical markers from blood and urine samples. By employing random forest and multinomial logistic regression models, they identified key indicators such as creatine kinase, lactate dehydrogenase, and aspartate aminotransferase as significant predictors of recovery effectiveness. Their results demonstrated that ML can accurately distinguish between different metabolic states, providing valuable insights into post-competition recovery strategies. This research highlights the potential of data-driven approaches in optimizing athlete performance and injury prevention, but it does not incorporate biomechanical or training load data for real-time injury risk prediction.

Rothschild et al. [43] conducted a comprehensive ML study on recovery prediction in endurance athletes (67.3% triathletes), which provides valuable insights for injury risk prediction in triathletes. Their work evaluated nine different algorithms, ranging from regularized linear models to neural networks, on multimodal data collected from wearables and mobile applications. Notably, they found that simpler LASSO (Least Absolute Shrinkage and Selection Operator) models outperformed more complex algorithms like XGBoost and neural networks, achieving a Root Mean Square Error (RMSE) of 11.8 for group-level recovery predictions. Their methodological contribution of using Markov unfolding with a 7-day look back window to transform time-series data into independent observations is particularly relevant for handling temporal dependencies in wearable sensor data. The study's finding that individual-level models showed wide variation in accuracy (RMSE range: 5.5-23.6) underscores the importance of personalised approaches in athlete monitoring systems. While their work focused on recovery prediction rather than injury risk, their technical approach to handling missing data through multiple linear regression and their feature selection methodology using permutation-based importance provides valuable insights for processing and analysing wearable sensor data in endurance sports.

There has been much research into injury prediction through ML in football [7, 44, 47, 48]. One prominent study has been done by Rossi et al. [7], which developed a ML-based injury prediction model for professional football, using wearable GPS tracking devices to collect detailed training workload metrics. In this study, 26 professional players were monitored for 23 weeks using a GPS sensor integrated with a 3D accelerometer, gyroscope, and digital compass. These devices recorded kinematic (*e.g.,* total distance covered), metabolic (*e.g.,* high metabolic load distance) and mechanical (*e.g.,* number of accelerations and decelerations) characteristics, which were then used to construct a dataset of 952 individual training sessions with 55 workload-related variables. The injury prediction model used a decision tree classifier with recursive feature elimination (RFECV) for feature selection and adaptive synthetic sampling (ADASYN) for class imbalance between injury and non-injury sessions. Validation was performed using stratified cross-validation, comparing the model with traditional workload heuristics such as the Acute:Chronic Workload Ratio (ACWR) and monotony-based predictors. The decision tree achieved 80% recall and 50% precision in injury classification, significantly outperforming baseline methods. Feature selection identified prior injury history (PI), high speed running distance (dHSR) and total distance monotony (dTOT MSWR) as key predictors. This study highlights the potential of integrating wearable sensor data with ML to improve injury prediction, while maintaining model interpretability for practical application.

Naglah et al. [44] propose a hierarchical ML framework for early injury prediction in football players by integrating internal and external training load data from wearable sensors and questionnaires. The study collected data from 21 football players over a 16-month period, tracking 36 non-contact injuries. Internal load metrics, such as heart rate, and external load data, including workout duration and number of jumps, were captured using wearable GPS sensors, accelerometers, and gyroscopes. The dataset was processed using feature selection techniques to extract 65 relevant attributes, which were then analysed using K-Nearest Neighbors (KNN) and K-Means clustering to identify injury-related features. The final injury prediction model used a Support Vector Machine (SVM) classifier trained on balanced datasets, with injury samples defined as training records within one week of an injury event. Model validation was performed using k-fold cross-validation, demonstrating that the SVM provided robust classification accuracy. The framework also generated athlete-specific probabilistic injury maps, allowing coaches to tailor training loads based on individual risk factors. The study concludes that the integration of wearable sensor data, statistical workload models and ML techniques enables early injury detection and supports proactive intervention strategies for injury prevention in team sports.

A further research related to football has been made by Ayala et al. [45]. They propose a ML-based model to predict hamstring strain injuries (HSIs) in professional football players using a combination of individual, psychological, and neuromuscular risk factors. The study followed 96 professional football players for an entire season, collecting pre-season screening data on 229 variables, including range of motion (ROM), muscle strength, core stability, sleep quality, and previous injury history. Injury surveillance recorded 18 hamstring injuries during the season. To develop an optimal predictive model, the authors evaluated several ML techniques, ultimately selecting an Alternating Decision Tree (ADTree) classifier within a SmoteBoostM1 ensemble as the most effective approach. This model achieved an AUC of 0.837, a true positive rate of 77.8% and a true negative rate of 83.8%, demonstrating high accuracy in classifying injury risk. In particular, previous injury history, sleep quality and hip flexion ROM were identified as important predictors of HSI. The study highlights the importance of integrating multifactorial risk assessment with ML techniques to improve injury prevention strategies in professional football. However, its generalisability to other sports remains uncertain and the complexity of the final model (10 classifiers, 66 predictors) may limit its practical implementation.

Moving away from football and into endurance sports, Lövdal et al. [8] applied bagged XGBoost ML models to predict injuries in competitive runners. Their dataset consisted of 74 high-level middle- and long-distance runners from a Dutch elite running team, collected over a period of seven years. The study integrated both objective data from wearable devices (*e.g.,* training duration, distance, and intensity) and subjective data, including perceived recovery before training, perceived exertion post training, and perceived training success. Injuries were flagged based on indicators such as the inability to complete a scheduled session.

Two different analytical approaches were employed: a ”daily” approach, which considered training data from up to one week prior to the injury event, and a ”weekly” approach, which aggregated data from the three weeks leading up to the injury and compressed each week into a composite score. The results showed that the day-to-day approach, produced more accurate injury predictions (AUC = 0.724) than the aggregated weekly composite approach (AUC = 0.678). However, the study did not incorporate objective recovery metrics such as heart rate variability (HRV) or sleep quality, and its reliance on a highly specific athlete population may limit generalizability to broader endurance sports contexts.

A more recent approach was presented by Chen [46], who introduced the Noninvasive Data Mining Model with Multi-relational and Multi-dimensional Clustering (Ni-DMMMC) algorithm to predict injury risk in athletes. This method leverages advanced clustering techniques to analyse biomechanical, physiological, training load, and environmental data, identifying hidden patterns in injury susceptibility. By categorizing athletes into distinct risk profiles, the study enables targeted interventions and personalised training modifications to minimize injury occurrences. The results demonstrate that the Ni-DMMMC algorithm effectively groups athletes based on multi-dimensional risk factors, improving prediction accuracy and aiding in injury prevention strategies.

Current limitations in injury prediction research include insufficient focus on multidisciplinary endurance sports such as triathlon [7, 44, 45], over-reliance on subjective data, and failure to integrate physiological parameters outside of training [8]. Furthermore all the works applying ML models for athlete monitoring, rely on limited datasets [43, 7, 44, 45, 8], for example Lövdal et al. [8] used a dataset originating from a single running team trained by a single coach.

Methodological advances worth incorporating include Rothschild’s Markov unfolding technique [43] for time series analysis, Lövdal’s finding [8] supporting daily over weekly modelling, and the integration of both external load metrics and internal response indicators [40, 41]. Rothschild’s observation [43] that simpler LASSO models outperformed complex algorithms suggests that parsimony may be beneficial with wearable data, while Naglah’s hierarchical framework [44] demonstrates the value of integrating data from multiple sources.

No existing study has comprehensively addressed injury prediction in triathletes using commercially available wearables with modern ML techniques. This thesis aims to fill this gap by developing a triathlon-specific ML model for injury prediction using multimodal sensor data from off-the-shelf sports watches, making advanced injury prevention accessible to all triathletes through non-invasive, already collected metrics.

3.1 Synthetic Data Generation Applications

Hohl et al. [9] conducted a comparative study of synthetic data generation methods in sport science, focusing on addressing the challenge of limited biological data availability. Their study used a limited real dataset of five endurance athletes (three ski mountaineers, one triathlete and one trail runner) who provided daily self-reported metrics over several months. These included sleep quality, mood and perceived exercise intensity (all rated on a 1-10 scale), as well as exercise duration and O_2Score, a biomarker of oxidative stress and recovery obtained from blood samples. The training workload was calculated as the product of session duration and perceived intensity. The number of days tracked per athlete ranged from 85 to 153, with the data structured into 8-day time series sequences to serve as ground truth for synthetic data generation.

The collected data was synthetically extended to 1000 multidimensional sequences of 8 days, using each of the following methods:

- **Random Sampling**: Data points were drawn from a standard normal distribution using the same mean and standard deviation as the real data to serve as a baseline comparison.
- **Noise Injection**: Real data sequences were slightly distorted with Gaussian noise to create variation while maintaining realism.
- **k-Medoids Clustering with Transition Probabilities**: Real sequences were clustered and synthetic sequences were created by sampling transitions between clusters based on learned probabilities.
- **Variational Autoencoder (VAE)**: A neural network was trained to encode and decode 8-day sequences, generating new sequences by sampling from the learned latent space.
- **TimeGAN (Generative Adversarial Network for Time-Series)**: A deep learning-based approach that uses adversarial training to generate realistic synthetic sequences by learning both temporal dependencies and feature correlations.
- **TimeGrad (Denoising Diffusion Model)**: A generative diffusion-based model trained to predict the next time step from previous values, iteratively generating complete synthetic sequences.

Then, the authors assessed the quality of the synthetic data generated using two key metrics: data fidelity (similarity between real and synthetic data) and data utility (usefulness of the synthetic data for a specific task). To measure utility, regression models were trained on synthetic data and tested on real data to predict O_2Score values, using a leave-one-out policy (one athlete was not included for training and later used for testing). Predicting this score should allow training adjustments to be made to avoid excessive fatigue and risk of injury, which is comparable to the aims of this thesis.

The results showed significant differences between the methods. Random sampling achieved high fidelity but produced unrealistic distributions and poor predictive performance, leading to a bad utility. K-Medoids clustering and noise injection achieved both high fidelity and correlations close to the ones obtained when the model was trained with real data, but posed privacy risks as the generated sequences closely resembled real data. VAE produced plausible but repetitive sequences, indicating a lack of diversity. TimeGrad produced diverse samples but struggled with fidelity, occasionally producing unrealistic sequences. TimeGAN emerged as the best performing method, striking a balance between fidelity, diversity and predictive utility, and demonstrating strong generalisation when models trained on synthetic data were tested on real data.

Lange et al. [23] explored the use of synthetic data generation to improve stress detection from smartwatch sensor data while ensuring privacy, in scenarios where real data availability is limited. Their study used open-source data from the WESAD (Wearable Stress and Affect Detection) dataset, which consists of physiological signals recorded under different emotional states (neutral, stress and amusement). The data was collected from wrist-worn devices.

To extend this dataset, they generated synthetic physiological time series data using three different GANs:

- **Conditional GAN (CGAN)**: A variant of GAN where both the generator and discriminator receive class labels as input, allowing it to generate synthetic samples specific to stress or non-stress conditions.
- **DoppelGANger (DGAN)**: Unlike standard GANs, DGAN decouples the generation of auxiliary metadata (e.g., labels) from the continuous time-series signals, using separate generators for each. Additionally, it includes an auxiliary discriminator that specifically evaluates metadata accuracy. To address mode collapse, a third generator is introduced to treat min and max signal values as metadata, enhancing diversity.
- **Differentially Private CGAN (DP-CGAN)**: A privacy-enhanced CGAN that incorporates differential privacy (DP) [1] to prevent the storage of individual real samples. The generator remains the same as in the standard CGAN, while the discriminator implements the noise injection.

They explicitly decided to neglect TimeGAN, used by Hohl *et al.* [9], as a related work [50] suggested that TimeGAN performed worse than CGAN for their use-case.

[1] a mathematical approach that guarantees that the addition/removal of a single data point in a data set does not significantly affect the results of statistical computations on that data set, while enforcing additional privacy [49]

The metrics used to evaluate the synthetic data quality were similar as previously and included diversity, fidelity and utility. Two different methods were compared for utility, augmentation-based training (AUGM) and TSTR. In AUGM, real training data was augmented with synthetic samples to determine whether stress detection performance improved.

Finally they found that DGAN achieved high fidelity but was prone to overfitting real data distributions, increasing privacy risks. CGAN produced high-quality synthetic samples, closely matching real data distributions and yielding the best classification performance. DP-CGAN effectively preserved privacy but slightly degraded fidelity and classification performance due to noise introduced by differential privacy mechanisms. Overall, CGAN emerged as the best-performing model, offering a good balance between fidelity, diversity, and utility. In non-private settings, models trained on CGAN-augmented data outperformed state-of-the-art stress detection models trained on real data alone. However, in private training scenarios, DP-CGAN maintained reasonable classification performance while ensuring strong privacy guarantees.

Both the Hohl *et al.* [9] and Lange *et al.* [23] studies examined synthetic data generation in contexts comparable to the wearable sensor data required for this thesis. In both cases, the studies focused on time series data, where observations at different time intervals exhibit temporal dependencies, with later values partially determined by earlier ones. Furthermore, Lange *et al.*[23] specifically worked with data from wrist-worn wearable devices, making their findings particularly relevant.

The results of these studies suggest that TimeGAN and CGAN are among the most promising approaches for synthetic data generation. Lange *et al.* [23], based on related work, suggested that CGAN would outperform TimeGAN for their specific use case. However, careful evaluation is required to determine which of these models is better suited to the objectives of this work. If either model is to be used, it will be essential to assess their ability to capture the complex relationships in triathlete sensor data.

Certain limitations of these studies must also be considered. All of the methods evaluated required real data for training, making them unsuitable for scenarios where no real data set is available. In addition, both studies relied on small datasets, raising concerns about potential overfitting and lack of generalisability. Hohl *et al.* [9] specifically validated their approach using a leave-one-out strategy, where only one athlete was excluded from training and used for testing. This could result in synthetic data that does not adequately represent inter-individual variability, a limitation that was not explicitly addressed in their paper.

Despite the promise of ML-based generative models, research on generating synthetic datasets from scratch - without relying on real data - remains scarce in sports science and health applications. More traditional approaches, such as rule-based generation (discussed in section 2.3), have not been widely explored for wearable sensor data. Investigation of such methods, possibly in combination with ML-based models, could provide valuable alternatives for synthetic data generation when real-world datasets are not available.

4 Design and Implementation

This chapter presents the comprehensive design and implementation used to develop a machine learning approach to injury prediction in triathletes. As shown in Figure 4.1, the design consists of two main steps: synthetic data generation and ML analysis, with the aim of predicting the likelihood of injury within a specific forward looking window.

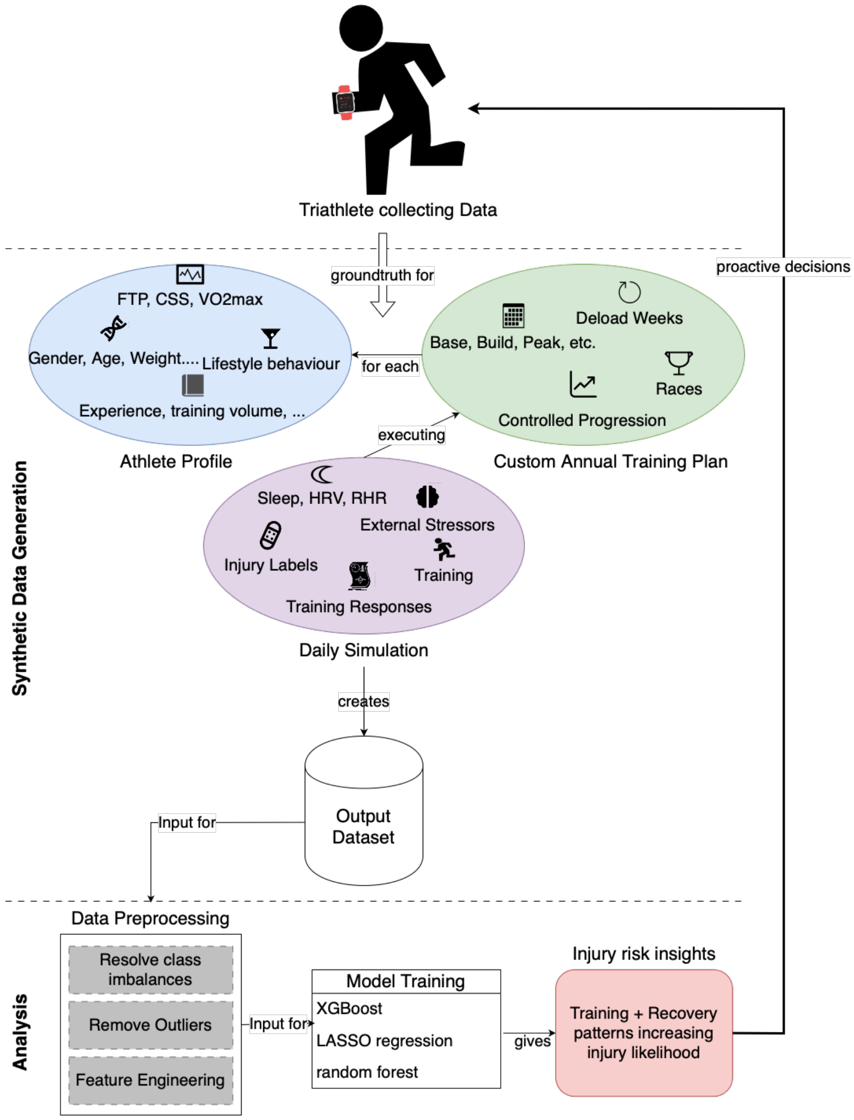

Figure 4.1: Methodology overview: Synthetic Dataset Generation and ML Analysis

In general, the limited availability of comprehensive triathlete monitoring data is a significant challenge for injury prediction research. Although Generative Adversarial Networks (GANs) can effectively augment limited real datasets, this thesis addresses the more challenging scenario of the complete absence of real training data. For this reason, it proposes a novel synthetic data generation framework that simulates comprehensive athlete profiles, including daily training executions and physiological responses throughout a complete annual cycle. This approach enables robust model training without relying on actual athlete data, addressing critical privacy concerns while maintaining sufficient diversity and realism in the synthetic population.

The framework begins with the creation of a diverse population of synthetic triathletes with physiologically consistent profiles. Each athlete is given a personalised annual training plan that incorporates periodization principles and race goals. Next, their daily physiological responses are simulated, including training execution, recovery metrics, and potential injury events. This simulation takes into account the stochastic nature of training responses and the influence of lifestyle factors and external stressors.

The second step of the design involves the training and evaluation of machine learning models using the synthetic dataset. This process includes data pre-processing to deal with class imbalances and outliers, feature engineering to capture relevant training patterns, and the application of various modelling techniques including XGBoost [51], LASSO regression [52] and random forest algorithms [53], following the insights gained in the related work Chapter 3. These models are used to identify significant training and recovery patterns that increase the likelihood of injury.

The following sections detail each component of the design and implementation, starting with the synthetic data generation framework, followed by the machine learning model development approach.

4.1 Synthetic Data Generation

This section introduces a comprehensive synthetic data generation framework that simulates the training-monitoring ecosystem of modern triathletes to produce realistic health and performance data. One central objective of this thesis is to develop a methodology that overcomes data limitations arising from privacy concerns in athletic monitoring. Consequently, no real athlete data is utilized, and the ML-based data augmentation methods evaluated in Section 3.1 are inadequate for this specialized domain.

Instead, the proposed methodology meticulously replicates the bidirectional relationship between athletes and wearable technology, capturing the continuous feedback loop of training stimulus, physiological response, and adaptation. This approach combines rule-based mechanisms, derived from domain knowledge, with sophisticated simulation techniques.

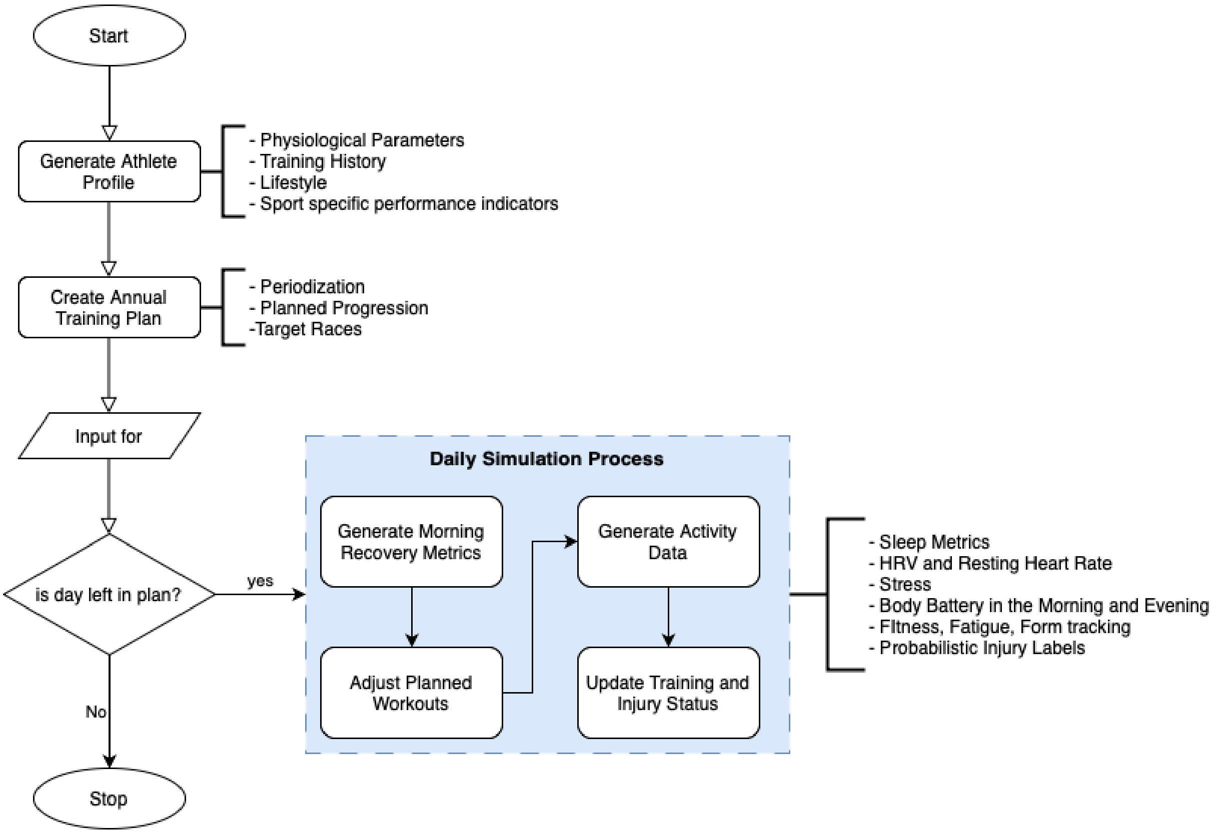

Figure 4.2: Data Generation Process

A substantial part of the category of endurance athletes in the generated dataset (see Section 4.1.1) is assumed to follow structured training plans that apply sports science principles to balance progressive overload with recovery. Despite careful planning, the unpredictable nature of human physiology, external stressors, and deviations from the plan introduce complexity, making injury risk a product of dynamic interactions rather than isolated training errors [54].

An example of this complexity is when a triathlete meticulously follows a structured training plan, but experiences an unexpected spike in work-related stress. This stress can lead to poor sleep quality and elevated cortisol levels, which in turn impair recovery and increase fatigue. As a result, even a routine training session can push the athlete beyond their recovery capacity and increase the risk of injury - not because of a direct training error, but because of the interplay between physiological stress, recovery limitations and cumulative workload.

Wearable devices have transformed training by providing daily physiological metrics – such as heart rate variability, sleep quality, and resting heart rate – that reflect the body's response to both training and external stressors [10]. These metrics influence subsequent training decisions, enabling a responsive system in which training sessions are adjusted based on physiological readiness.

The data generation process includes (*cf.* Figure 4.2):

- Creating diverse athlete profiles with physiologically plausible characteristics
- Generating periodized annual training plans incorporating key races
- Simulating daily training dynamics including:
  - Morning recovery metrics from wearable devices
  - Realistic deviations from planned workouts based on physiological readiness
  - sport-specific wearable device activity data
  - Training responses and potential injury events

This approach captures the complex, bidirectional relationship between training prescription and execution, taking into account how external stressors and recovery limitations interact with cumulative workload to influence injury risk.

4.1.1 Target Population Definition

The synthetic data generation focused specifically on competitive age-group triathletes to create a population with meaningful homogeneity while preserving realistic individual variations. Precise inclusion parameters were established to define this population, ensuring physiological coherence whilst allowing for the individual differences that characterise real world athlete cohorts. Inclusion parameters are outlined as follows (*cf.* Figure 4.3).

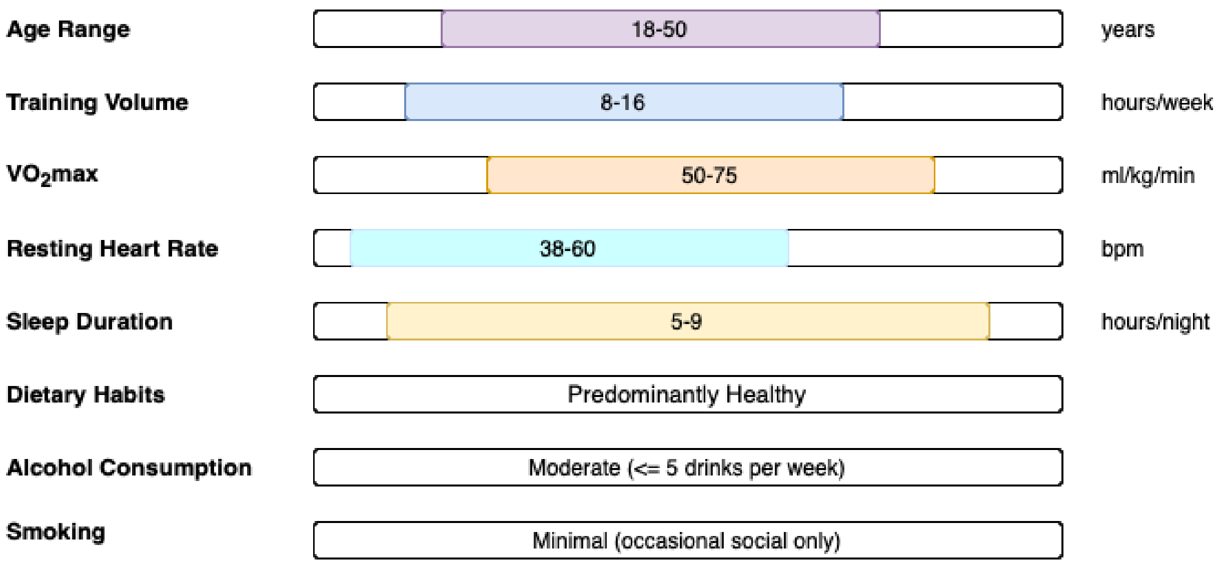

Figure 4.3: Inclusion Parameters

The target population represents dedicated amateur triathletes who train systematically toward competitive goals without reaching professional status. These athletes maintain substantial training volumes of 8-16 hours per week, distributed across swimming, cycling, and running [55]. Physiologically, they exhibit strong cardiovascular fitness with VO_2max values ranging from 50-75 ml/kg/min [56], reflecting their advanced but non-elite aerobic capacity. The age range was restricted to 18-50 years to capture the primary demographic of competitive age group athletes while excluding masters athletes who may have different physiological characteristics.

Lifestyle parameters consistent with health-conscious competitive athletes were implemented, based on assumptions. These include predominantly nutritionally optimized dietary practices (appropriate macronutrient distribution tailored to training phases, emphasis on whole foods rich in micronutrients, strategic nutrient timing around workouts, and minimal consumption of highly processed foods), minimal smoking behaviour (limited to occasional social consumption) and moderate alcohol consumption not exceeding five standard drinks per week. Cardiovascular fitness markers were limited to a resting heart rate of 38-60 beats per minute, reflecting the increased parasympathetic tone typical of endurance trained individuals [57]. Sleep patterns were realistically modelled with durations of 5-9 hours per night, recognising the variability in sleep habits even among dedicated athletes balancing training with career and family responsibilities.

These boundaries create a well-defined population that represents the core demographic of Olympic and half-Ironman distance triathlon participants at a competitive but non-professional level. By constraining these parameters, a synthetic population is created with sufficient homogeneity to allow meaningful pattern recognition in subsequent machine learning analysis, while preserving the natural variability that characterises real-world athlete populations.

4.1.2 Athlete Profile Generation

Each synthetic athlete profile includes 24 parameters that cover demographic characteristics, physiological metrics, sport-specific performance indicators, training history, and lifestyle factors relevant to triathlon performance (see Table 4.1).

Lifestyle factors are determined by six different lifestyle profiles designed to reflect the competitive age-group triathlete population, incorporating weighted random variation between individuals. These profiles include ”Highly Disciplined Athlete” (30%), ”Balanced Competitor” (25%), ”Weekend Socialiser” (12%), ”Sleep-deprived Workaholic” (12%), ”Under-Recovered Athlete” (11%) and ”Health-Conscious Athlete” (10%). Each profile identifies key lifestyle parameters such as sleep duration and quality, dietary habits, stress levels and alcohol consumption - all of which influence physiological metrics such as heart rate variability, recovery rate and performance markers [58, 59, 60].

To maintain physiological realism and internal consistency, established principles of sport science were used to model interdependencies between parameters. This approach ensured that synthetic profiles represented plausible athlete characteristics rather than random combinations of values. For example, a multi-factorial approach was used to estimate lactate threshold heart rate (LTHR).

$$\text{LTHR}_{\text{est}} = \max\left(160, \min\left(HR_{\text{max}} \times P_{\text{base}} \times M_{\text{final}} \times (1+\epsilon), 190\right)\right) \tag{4.1}$$

where:

- LTHR_{est} is the estimated lactate threshold heart rate.
- HR_{max} is the athlete's maximum heart rate.
- $P_{\text{base}} = 0.87$ is the base LTHR percentage for competitive triathletes.
- M_{final} is the final modifier, computed as:

$$M_{\text{final}} = M_{\text{age}} \times M_{\text{training}} \times M_{\text{VO2max}} \times M_{\text{HRR}} \times M_{\text{gender}} \tag{4.2}$$

The model begins with a baseline assumption that LTHR occurs at approximately 87% of maximal heart rate in well-trained endurance athletes, consistent with the findings of Bentley et al. [61], who demonstrated that lactate threshold typically manifests within 85-90% of maximal heart rate in this population. This baseline is then modified by a series of evidence-based adjustments to account for key physiological determinants. For example, age-related modifiers are consistent with research by Tanaka and Seals [62], who found that while well-trained older athletes can maintain high lactate thresholds, age-related declines become more pronounced after the age of 50. Gender differences are accounted for with a slight upward adjustment for women (3%), supported by Billaut and Bishop's [63] findings that women generally have higher fat oxidation rates and potentially superior lactate clearance at submaximal intensities.

The model also incorporates VO_2max as a key determinant, following Coyle's [64] research demonstrating that highly trained endurance athletes with superior maximal oxygen uptake can sustain work rates closer to 90% of their VO_2max power. Finally, Heart Rate Reserve (HRR), the difference between maximal and resting heart rate, serves as an additional modifier, building on Karvonen's [65] foundational work establishing HRR as an effective predictor of cardiovascular efficiency.

Table 4.1: Parameter Ranges and Distributions for Synthetic Athlete Profiles

Category	Parameter	Range	Distribution	Source
Demographic	Gender	Binary	60% male, 40% female	[66]
	Age	18–50 years	Normal ($\mu = 33$, $\sigma = 6$)	
	Height	Males: 164–192 cm	Normal ($\mu = 178$, $\sigma = 7$)	
		Females: 153–177 cm	Normal ($\mu = 165$, $\sigma = 6$)	
	Weight	Variable	Gender-specific, height-adjusted	
Physiological	Genetic predisposition	0.8–1.2 scale	Truncated normal ($\mu = 1$, $\sigma = 0.1$)	
	HRV baseline	40–95 ms	Age-adjusted, scaled by VO_2max, resting HR	[67]
	HRV range	±15% of baseline	Derived from baseline	
	Resting HR	38–60 bpm	Derived from VO_2max with lifestyle adjustments	[68]
	Maximum HR	165–205 bpm	Tanaka formula (208 - 0.7×age) with random variation ($\sigma = 5$)	[62]
	Lactate threshold HR	160–190 bpm	8̃7% of max HR, adjustments for age, training, VO_2max	[61, 69]
	VO_2max	Males: 50–75 ml/kg/min	Base: Normal ($\mu = 49$, $\sigma = 4$) with modifiers	[70]
		Females: 50–70 ml/kg/min	Base: Normal ($\mu = 45$, $\sigma = 4$) with modifiers	
Performance	Functional threshold power	2.5–5.5 W/kg	Gender-specific, experience and specialization adjusted	[71]
	Critical swim speed	0.8–1.55 m/s	Converted to seconds per 100m (64–125 s/100m)	
	Running threshold pace	3:00–5:30 min/km	VO_2max-correlated with multiple adjustments	
Training	Experience level	2–20 years	Right-skewed, age-constrained	
	Weekly training hours	8–16 hours	Normal ($\mu = 12$, $\sigma = 2$) with lifestyle adjustments	[72]
	Recovery rate	0.5–1.3 scale	Complex function of age, VO_2max, and lifestyle factors	[36]
Lifestyle	Sleep duration	5–9 hours	Profile-dependent	[73]
	Sleep quality	0.3–1.0 scale	Profile-dependent	[58]
	Nutrition quality	0.4–1.0 scale	Profile-dependent	
	Stress level	0–0.9 scale	Profile-dependent	[58]
	Training adherence	0.6–1.0 scale	Profile-dependent	
	Smoking	0–0.1 scale	5% occasional smokers, 95% non-smokers	
	Alcohol consumption	0–0.6 scale	Profile-dependent, limited range	

4.1.3 Annual Training Plan Construction

Following the synthetic athlete profile generation, a personalised annual training plan for each athlete is created. This process transforms the athlete's characteristics into a structured day-by-day training schedule using a series of algorithmic steps to ensure appropriate training load, progression, and specificity (*cf.* Table 4.2).

Table 4.2: Example One-Week Training Plan

Day	Phase	Total TSS	Swim TSS	Bike TSS	Run TSS	Strength TSS	Workout Details
Mon	Base	0	0	0	0	0	**REST DAY**
Tue	Base	65	15	30	20	0	Swim: Technique drills (30 min) Bike: Zone 2 endurance (1 hr) Run: Easy run (30 min)
Wed	Base	45	0	0	45	0	Run: Tempo intervals (45 min)
Thu	Base	90	20	50	0	20	Swim: Speed work (45 min) Bike: Sweet Spot (1.5 hr) Strength: Core strength (30 min)
Fri	Base	0	0	0	0	0	**REST DAY**
Sat	Base	130	0	80	50	0	Bike: Long ride (2.5 hr) Run: Endurance run (40 min)
Sun	Base	75	25	0	50	0	Swim: Easy Swim (45 min) Run: Long run (1 hr)

4.1.3.1 Training Stress Score

The training plan is based on the concept of Training Stress Score (TSS), a metric originally developed by Dr. Andrew Coggan and former professional cyclist Hunter Allen [74] to quantify training load. The TSS is a dimensionless number that represents the total training load accumulated during a training session, taking into account both intensity and duration. A TSS of 100 is equivalent to an athlete performing at their maximum sustainable effort (*e.g.,* FTP for cycling or running threshold pace) for one hour. Lower intensity workouts accumulate TSS more slowly, while higher intensity workouts accumulate TSS more quickly.

TSS is widely used in cycling, running, and swimming, but it is rarely applied to strength training due to differences in physiological stress. This limitation was disregarded in this model.

Garmin devices estimate training load using Excess Post-Exercise Oxygen Consumption (EPOC) [75], which measures the elevated oxygen consumption following exercise as an indicator of energy expenditure. While EPOC provides a valid alternative for assessing load, synthetically modelling it proved more complex than using TSS. Therefore, TSS was chosen as the primary load metric in this thesis.

4.1.3.2 Periodization Framework

A race schedule, that scales appropriately with experience level (calculated according to the function shown in Listing 4.1), is generated. Beginners are allocated 1-2 races per year, intermediate athletes 2-4 races, advanced athletes 3-6 races, and elite athletes 5-8 races. Races are placed primarily in typical racing season months (May-September) with adequate recovery intervals between competitions (minimum 21 days) [76].

```python
def calculate_athlete_ability(years_experience, age, vo2max, weekly_hours,
    ftp_kg):
    # Define scoring ranges
    score_ranges = {
        "experience": [(10, 20), (float("inf"), lambda x: min(20, x * 2))],
        "age": [(25, 8), (35, 10), (45, 7), (55, 6), (float("inf"), 3)],
        "vo2max": [(40, 5), (50, 10), (60, 15), (70, 20), (float("inf"),
            25)],
        "weekly_hours": [(7, 5), (9, 10), (12, 15), (float("inf"), 20)],
        "ftp_kg": [(3, 5), (3.5, 10), (4.5, 15), (5.5, 20), (float("inf"),
            25)],
    }
    # Helper function to get score from ranges
    def get_score(value, category):
        for threshold, score in score_ranges[category]:
            if value < threshold:
                return score(value) if callable(score) else score
    # Calculate individual scores
    experience_score = get_score(years_experience, "experience")
    age_score = get_score(age, "age")
    vo2max_score = get_score(vo2max, "vo2max")
    weekly_score = get_score(weekly_hours, "weekly_hours")
    ftp_score = get_score(ftp_kg, "ftp_kg")

    # Total Score Calculation
    total_score = sum([experience_score, age_score, vo2max_score,
        weekly_score, ftp_score])
    # Ability Level Categorization
    ability_levels = [(30, "Beginner"), (50, "Intermediate"), (75, "
        Advanced"), (float("inf"), "Elite")]
    ability_level = next(label for threshold, label in ability_levels if
        total_score <= threshold)

    return ability_level
```

Listing 4.1: Python function to calculate athlete ability

The training plan follows a traditional periodization model structured around the races and divided into distinct training phases (*cf.* Figure 4.4), following principles found in [77, 78, 79, 80]. The base phase focuses on building aerobic capacity and endurance with predominantly low intensity training (approximately 90% low intensity, 10% high intensity), typically comprising

40-50% of the pre-race preparation period. As athletes progress to the Build phase, training specificity and intensity increase (shifting to approximately 80% low intensity, 20% high intensity) while maintaining aerobic development. Build phases include systematic progression with deload weeks every fourth week to prevent overtraining and constitute 30-40% of the pre-race preparation period.

The peak phase emphasizes race-specific intensities and reduces volume to sharpen performance (70% low intensity, 30% high intensity), representing 10-15% of the preparation period. This is followed by a taper phase that reduces training volume while maintaining intensity to promote recovery and peak performance for competition, typically spanning the last two weeks before a race. The Race Phase accommodates the competition itself, with the training stress score (TSS) adjusted according to the expected race demands based on the level of athlete experience.

Post-competition recovery phases promote active recovery with substantially reduced volume and intensity, varying from 7-14 days depending on the athlete's recovery rate attribute. During periods without races, the Off-season Phase provides structured downtime while maintaining baseline fitness.

Each phase is assigned a TSS factor that modifies the base weekly training load (*e.g.,* recovery phases use a factor of 0.5, while peak phases may use a factor of 1.1).

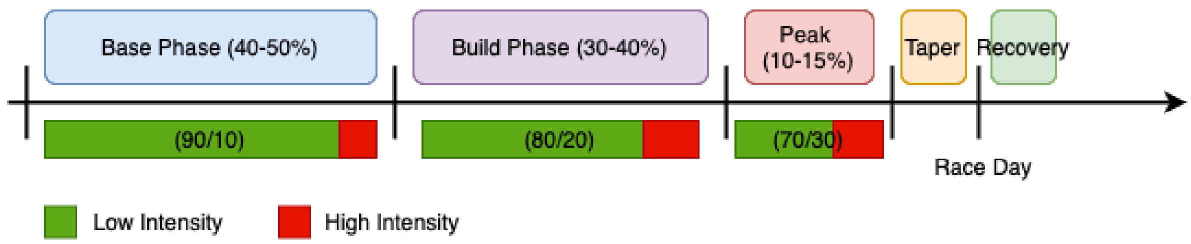

Figure 4.4: Training Periodization Model

4.1.3.3 Training Load Management

The framework implements several evidence-based principles for managing training load. Progressive overload is maintained by limiting the increase in weekly training stress to 10% [81, 82], reducing the risk of injury while promoting adaptation. Daily and weekly TSS values are calculated based on the athlete's weekly training hour availability, experience level, current training phase, and day-of-week patterns that allocate more training on weekends and less on Mondays and Fridays.

```python
# Decide if this is a rest day (more likely on Mondays and Fridays)
rest_probability = 0.05  # Base probability
if day_of_week == 0 or day_of_week == 4:  # Monday or Friday
    rest_probability = 0.2

# Recovery weeks have more rest days
if current_phase['name'] == 'Recovery' or current_phase['name'] == 'Taper':
    rest_probability *= 2

# Determine if it's a rest day
is_rest_day = random.random() < rest_probability
```

Listing 4.2: Python code snippet to assign rest days

Rest and recovery are algorithmically determined (*cf.* Listing 4.2) with higher probability on specific weekdays and increased frequency during recovery and taper phases. Every fourth week during build and peak phases incorporates a systematic reduction in training load (approximately 30% lower TSS) to facilitate recovery and adaptation [83, 84, 85].

4.1.3.4 Sport Distribution and Workout Assignment

Training load is distributed across swimming, cycling, running and strength training based on a combination of factors [86, 87]. For example, identified strengths and weaknesses of the athlete, principles of injury prevention, requirements of the training phase, distribution on race day, and practical limitations. The default distribution allocates approximately 20% to swimming, 50% to cycling, 25% to running, and 5% to strength training.

Each day's training is further specified with workout types and durations, based on TSS allocation, training phase, and the 80/20 polarized training principle [88, 89, 90]. Workout durations are calculated from TSS values using sport-specific TSS-per-hour estimates that vary by type of workout. The model supports a variety of types of workouts in all sports, which are summarized in Table 4.3.

4.1.3.5 Training Plan Output Specification

For each athlete, the system produces a comprehensive annual training plan comprising 365 days (or user-specified duration), where each day specifies:

- Date and day of the week
- Training phase (Base, Build, Peak, etc.)
- Total planned TSS

Table 4.3: Workout Library with Descriptions and TSS Per Hour

Discipline	Workout Name	TSS/Hour	Description
Swim	Easy Swim	25	Easy technique-focused swim with drills
	Endurance Swim	50	Steady-paced endurance swim with some drill sets
	Swim Intervals	90	Mixed intervals focusing on speed and technique
	Threshold Swim	80	Sustained effort at or near threshold pace
	Speed Work	100	Short, high-intensity repeats with full recovery
Bike	Recovery Ride	30	Very easy spinning to promote recovery
	Endurance Ride	40	Steady effort to build aerobic endurance
	Tempo Ride	60	Sustained moderate effort with some harder efforts
	Sweet Spot Intervals	70	Intervals at 88-93% of FTP
	Threshold Intervals	80	Intervals at or just below FTP
	VO2max Intervals	100	Short, high-intensity intervals
Run	Recovery Run	30	Very easy pace to promote recovery
	Endurance Run	50	Steady effort to build aerobic endurance
	Long Run	50	Extended duration at easy to moderate pace
	Tempo Run	70	Sustained effort at moderate intensity
	Threshold Intervals	95	Intervals at or near threshold pace
	Speed Intervals	100	Short, high-intensity repeats with recovery
Strength	Core Strength	40	Core-focused exercises to improve stability
	General Strength	50	Full-body strength routine
	Sport-Specific	60	Strength exercises targeting swim/bike/run muscles
	Plyometric Training	70	Explosive exercises to build power

- Sport-specific TSS allocation (swim, bike, run, strength)
- Specific workout name for each discipline (*e.g.,* ”Threshold Intervals”, ”Recovery Ride”)
- Planned duration in minutes for each workout
- Rest day and race day indicators
- Week number for periodization and progressive overload tracking

(*cf.* Table 4.2)

4.1.4 Daily Training Plan Execution Simulation

Next, the framework implements a detailed daily simulation process to model realistic training execution and a physiological response recorded using wearable devices. This simulation accounts for the inherent variability between planned training and actual execution, incorporating numerous factors that influence athletic performance, adaptation, and injury development in real-world scenarios.

The daily simulation iterates through each day of the annual plan, executing a sequential process that models both physiological states and training behaviors:

Algorithm 1 Daily Activity and Health Data Simulation Loop

1: Morning Physiological State Simulation
2: Training Execution with Realistic Deviations and Wearable Activity Data Simulation
3: Training Load Metrics Calculation
4: Evening Recovery Status Simulation
5: Injury Risk Assessment and Simulation
6: State Transition to Next Day

4.1.4.1 Morning Physiological State Simulation

Each simulation day begins with the generation of morning physiological markers that reflect recovery status and readiness to train by modelling several interrelated physiological markers:

Algorithm 2 Morning Physiological State Simulation

1: Calculate recovery parameters based on previous day metrics and training history
2: Simulate sleep quantity and quality metrics
3: Calculate recovery heart rate, heart rate variability and morning body battery based on recovery parameters and sleep metrics

Recovery Parameter Calculation The simulation first evaluates the foundational recovery parameters from the previous day's metrics and training history. These parameters include total fatigue accumulation (incorporating both immediate and delayed fatigue effects with a 72-hour window), recovery score, injury effects, and training load patterns.

$$\text{TotalFatigue} = \frac{\text{Fatigue}_{\text{immediate}}}{\text{RecoveryRate}} + \sum_{i=1}^{3} \frac{\text{TSS}_{t-i} \cdot \text{DelayFactor}_i}{\text{RecoveryRate}} \tag{4.3}$$

where DelayFactor_i represents the impact coefficients for training loads from previous days (0.3, 0.15, and 0.05 for days $t-1$, $t-2$, and $t-3$ respectively).

The simulation also accounts for chronic adaptations by analyzing the preceding 28-day training history [91]:

$$\text{ChronicAdaptation} = \begin{cases} 0.2 \cdot \frac{\text{AvgMonthlyTSS}}{\text{MaxDailyTSS}}, & \text{if AvgMonthlyTSS} > 0.7 \cdot \text{MaxDailyTSS} \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

Sleep Metrics Simulation Sleep quality and quantity are simulated, with the framework generating [92, 93, 94]:

- Sleep duration as a function of normal sleep patterns indicated in an athlete profile, modified by fatigue, injury, and stress factors (*cf.* Listing 4.4)
- Sleep architecture distribution (deep, REM, and light sleep percentages) (*cf.* Listing 4.3)
- Overall sleep quality score derived from both duration and sleep stage distribution (0-1 scale)

The sleep quality calculation employs a two-factor model that weighs both duration and sleep architecture:

$$\text{SleepQuality} = w_d \cdot \text{DurationScore} + w_s \cdot \text{StageQualityScore} \tag{4.5}$$

where w_d and w_s are weighting factors that shift based on sleep duration, with stage quality becoming more influential when sleep duration exceeds 6 hours.

```python
def _calculate_sleep_distribution(self, sleep_hours, fatigue_factor,
    injury_effect, stress_factor):
    # Adjust sleep stages based on fatigue, injury, and stress
    deep_sleep_pct = self.IDEAL_SLEEP_PROPORTIONS['deep'] - (0.05 *
        fatigue_factor) - (0.07 * injury_effect) - (0.03 * stress_factor)
    rem_sleep_pct = self.IDEAL_SLEEP_PROPORTIONS['rem'] - (0.03 *
        fatigue_factor) - (0.05 * injury_effect) - (0.02 * stress_factor)
    # Ensure sleep percentages are physiologically plausible
    deep_sleep_pct = max(0.08, min(deep_sleep_pct, 0.25))
    rem_sleep_pct = max(0.15, min(rem_sleep_pct, 0.25))
    light_sleep_pct = 1 - deep_sleep_pct - rem_sleep_pct
    # Calculate actual sleep times
    deep_sleep = sleep_hours * deep_sleep_pct
    rem_sleep = sleep_hours * rem_sleep_pct
    light_sleep = sleep_hours * light_sleep_pct
    return deep_sleep, rem_sleep, light_sleep
```

Listing 4.3: Calculation of the Distribution of Deep, REM, and Light Sleep

```python
def _calculate_sleep_hours(self, fatigue_factor, injury_effect,
    stress_factor, sleep_norm):

    fatigue_sleep_effect = 0.1 * fatigue_factor - 0.2 * injury_effect
    stress_effect = 0.1 * stress_factor
    sleep_hours = sleep_norm + fatigue_sleep_effect - stress_effect +
        random.normalvariate(0, 0.5)

    return max(sleep_hours, self.MIN_SLEEP_HOURS)
```

Listing 4.4: Sleep Duration Calculation

4.1.4.2 Training Execution with Realistic Deviations

The framework models the execution of planned training with probabilistic deviations based on the athlete's physiological state and fatigue level. When physiological markers indicate suboptimal recovery (*e.g.,* suppressed HRV, poor sleep quality), the simulation can reduce the probability of completion of the exercise, adjust the intensity of training, or modify the duration [95]. This approach reflects real-world athlete behaviour, where training adherence is influenced by multiple physiological feedback mechanisms.

Training execution is modeled through a logistic function that transforms fatigue levels into completion probabilities:

$$P_{\text{base}} = \frac{1}{1 + e^{(\text{fatigue}-75)/10}} + \frac{\text{sleep\_quality} \cdot 100 - 50}{200} \tag{4.6}$$

This base probability is further adjusted by HRV status, with workout modifications following these principles:

$$P_{\text{completion}} = \begin{cases} P_{\text{base}} \cdot 0.5, & \text{if HRV Status is very low} \\ P_{\text{base}} \cdot 0.8, & \text{if HRV Status is slightly low} \\ P_{\text{base}} \cdot 1.0, & \text{if HRV Status is normal} \\ P_{\text{base}} \cdot 1.2, & \text{if HRV Status is high and fatigue} < 40 \end{cases} \tag{4.7}$$

$$\text{HRV Status} = \begin{cases} \text{Very Low}, & \text{if } HRV_{\text{current}} < 0.75 \times HRV_{\text{baseline}} \\ \text{Slightly Low}, & \text{if } 0.75 \times HRV_{\text{baseline}} \leq HRV_{\text{current}} < 0.85 \times HRV_{\text{baseline}} \\ \text{Normal}, & \text{if } 0.85 \times HRV_{\text{baseline}} \leq HRV_{\text{current}} \leq 1.15 \times HRV_{\text{baseline}} \\ \text{High}, & \text{if } HRV_{\text{current}} > 1.15 \times HRV_{\text{baseline}} \end{cases} \tag{4.8}$$

For executed workouts, the actual training stress and duration are calculated as:

$$TSS_{\text{actual}} = 100 \cdot (\text{IF}_{\text{actual}})^2 \cdot \frac{\text{duration}_{\text{actual}}}{60} \tag{4.9}$$

where $\text{IF}_{\text{actual}}$ is the intensity factor, potentially adjusted based on fatigue and HRV status. The framework also accounts for unplanned additional workouts, which occur with higher probability when fatigue is low and on days where 0 or 1 session is planned.

The usage of HRV as indicator for training readiness is widely adopted in wearable device technology. For example, Garmin uses HRV trends to assess daily training readiness and training status [96].

4.1.4.3 Wearable Activity Data Simulation

For each executed training session, the framework generates realistic, sport-specific wearable device data that mimics the physiological responses an athlete would experience during actual workouts.

```
for sport, activity in actual_activities.items():
    if activity['workout'] == "Unplanned extra session":
        duration_hours = activity['actual_duration'] / 60
        extra_tss_per_hour = activity['actual_tss'] / duration_hours
        sport_workouts = workout_libraries[sport]
        closest_match = min(sport_workouts.items(),
                        key=lambda item: abs(item[1]['tss_per_hour'] -
                            extra_tss_per_hour))
        workout_type = closest_match[0]
        activity['workout'] = f"Unplanned Extra Workout ({workout_type.
            capitalize()})"
```

Listing 4.5: Assigning an existing workout to an unplanned training session

The simulation process (*cf.* Figure 4.5) begins by determining the appropriate workout type. For planned activities, the framework uses the predefined workout parameters. For unplanned activities (those not present in the original training plan), the system identifies the closest matching

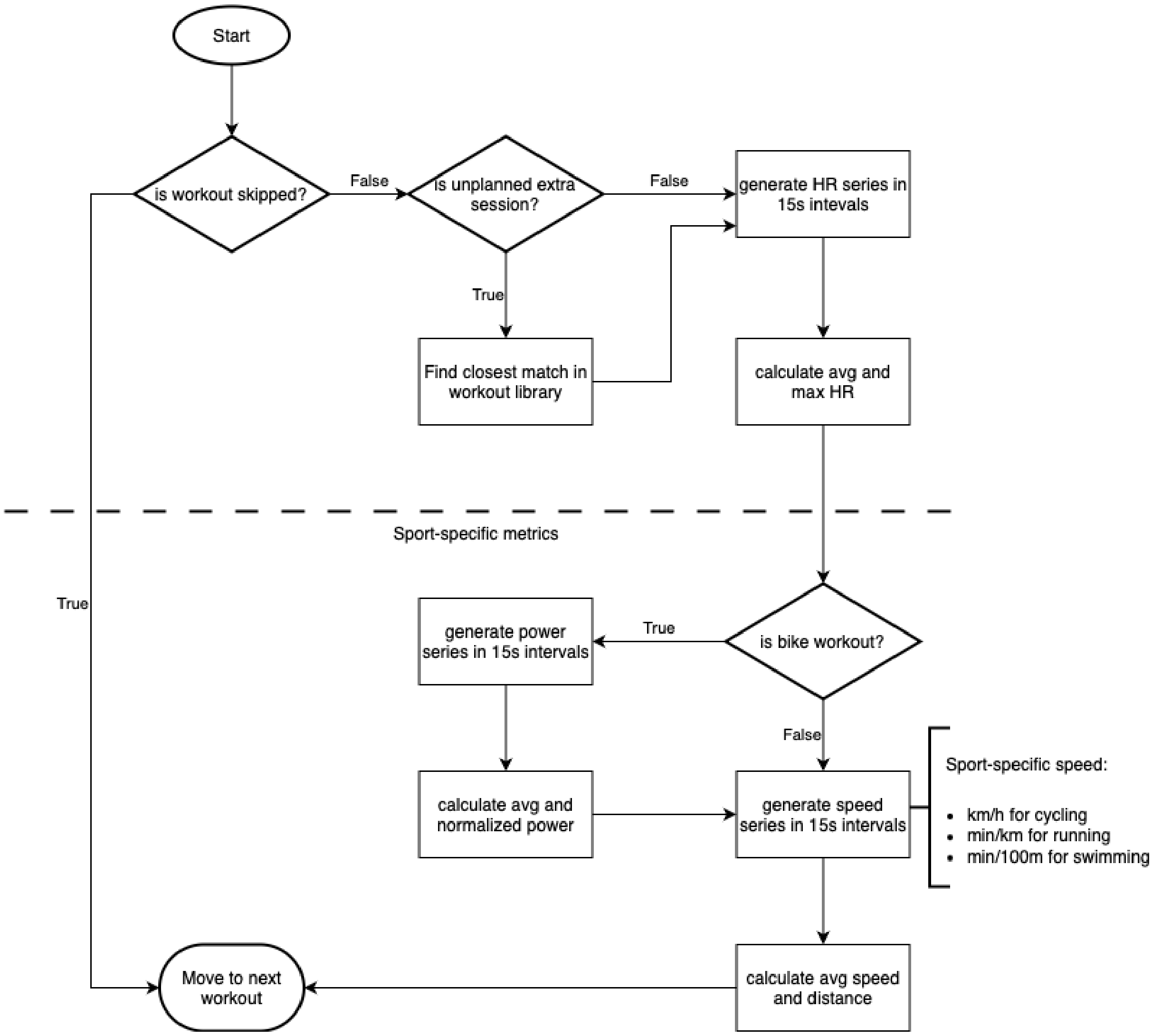

Figure 4.5: Sport-specific Activity Data Generation Process

workout from the library based on TSS per hour (Listing 4.7).

Once the workout type is established, the framework generates time series data at 15-second intervals for the duration of the session. The frequency in commercially available wearable devices is significantly higher (1-50 Hz), but for simplicity reasons the specified time interval was considered as sufficient. This includes heart rate data for all activities, as well as sport-specific metrics such as speed (cycling: km/h, running: min/km, swimming: min/100m), and power output for cycling.

`generate_hr_series()` The `generate_hr_series()` function creates realistic cardiovascular response patterns that account for workout type, intensity distribution, duration, and various physiological factors. The general relationship can be expressed as:

$$\mathrm{HR}_t = f(w, z, i, d, f) + \varepsilon_t \tag{4.10}$$

where HR_t is the heart rate at time t, w represents workout type parameters, z denotes athlete-specific heart rate zones, i is the intensity factor, d accounts for cardiac drift, f is the athlete's fatigue level, and ε_t is a stochastic component representing natural heart rate variability. The function implements the following procedural workflow:

1. Calculate intensity adjustment factors based on the ratios between actual and planned TSS and duration
2. Determine target heart rate zone distributions using the `determine_target_zones()` function, which maps workout names and sport types to physiologically appropriate zone distributions
3. Apply intensity-based adjustments to these zone distributions using `adjust_zones_for_tss()`
4. Select and execute the appropriate pattern generator based on workout classification:
   - `generate_interval_hr()` for interval, and VO2max workouts
   - `generate_strength_hr_series()` for strength training sessions
   - `generate_steady_pattern()` for endurance, recovery, and tempo workouts

The model incorporates several sophisticated physiological elements to enhance realism:

- **Zone-specific intensity distributions**: Each workout type has a characteristic pattern of time spent in different heart rate zones [97, 98, 99]. For example, recovery workouts emphasize Zone 1 (95%) and Zone 2 (5%), while VO_2max sessions distribute effort across Zone 3 (10%), Zone 4 (35%), and Zone 5 (55%).
- **Cardiac drift modelling**: The simulation accounts for cardiovascular drift–the progressive increase in heart rate during sustained exercise despite constant workload [100]–using sport- and duration-specific parameters. This effect is modeled as:

  $$\mathrm{HR}_{\text{drift}} = \mathrm{HR}_{\text{base}} \cdot (1 + k_d \cdot t) \tag{4.11}$$

  where k_d is the drift factor and t is exercise duration in hours.
- **Fatigue influence**: Accumulated fatigue elevates heart rate response [101], modeled as a multiplicative factor:

  $$\mathrm{HR}_{\text{fatigued}} = \mathrm{HR}_{\text{base}} \cdot \left(1 + \frac{f}{400}\right) \tag{4.12}$$

where f represents the current fatigue level of the athlete.

- **Heart rate kinetics**: For interval workouts, the system models the non-linear cardiovascular response, including the delayed rise in heart rate at interval onset and the exponential decay during recovery periods [102].
- **Natural variability**: Gaussian noise ($\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$) is added to simulate beat-to-beat variability, with σ values tuned to different workout phases.
- **Sport-specific adjustments**: Swimming heart rates are typically lower than running or cycling at equivalent intensities [103]. The system accounts for this by adjusting the zone distributions accordingly.

For interval-based workouts, the heart rate pattern exhibits characteristic features including:

- Rapid initial rise (approximately 20% of interval duration), followed by a plateau or slight increase [102].
- Two-phase recovery kinetics: an initial rapid decline (approximately 30% of recovery period), followed by slower return toward baseline [104].

For strength training, the model simulates the cardiovascular response to resistance exercise, with distinct set-and-rest cycles producing a sawtooth pattern in heart rate data. The simulation accounts for cumulative fatigue across multiple sets, resulting in progressively higher peak heart rates despite equivalent work [105].

`generate_power_series()` For cycling activities, the framework generates realistic power output data through the `generate_power_series()` function. This function simulates the mechanical work produced during different types of cycling exercise, accounting for physiological capabilities, fatigue, and specific power distributions for the workout. The power output at time t can be represented as:

$$\text{Power}_t = g(w, z, i, f) \cdot (1 - k_f \cdot f) + \eta_t \tag{4.13}$$

where g represents the power generation function based on workout type w, power zones z, and intensity factor i; k_f is the fatigue coefficient; f is the athlete's fatigue level; and η_t is the stochastic component that models natural power variations.

The power generation process follows these methodical steps:

1. Calculate intensity adjustment factors based on the ratios of actual to planned TSS and duration.

2. Determine target power zones using `determine_power_target_zones()`, which maps workout types to appropriate power zone configurations.

3. Adjust the target zones according to the intensity ratios with `adjust_power_zones_for_tss()`.

4. Generate the power series using either:

   - `generate_interval_power()` for interval, VO_2max, and sprint workouts.

   - `generate_steady_power()` for endurance, tempo, and recovery rides.

The model incorporates several cycling-specific physiological elements:

- **Zone-specific power targets**: Each power zone corresponds to a percentage of the athlete's Functional Threshold Power (FTP), creating physiologically appropriate intensities for different workout types [106].

- **Fatigue modelling**: Unlike heart rate, which increases with fatigue, power output capability decreases with accumulated fatigue [107], modelled as:

$$\text{Power}_{\text{fatigued}} = \text{Power}_{\text{base}} \cdot \left(1 - \frac{f}{500}\right) \tag{4.14}$$

- **Power drift**: For sustained efforts, the model includes a gradual decline in sustainable power over time, reflecting glycogen depletion and neuromuscular fatigue [108]:

$$\text{Power}_{\text{drift}} = \text{Power}_{\text{base}} \cdot (1 + k_d \cdot t) \tag{4.15}$$

  where k_d is the drift factor and t is the duration of the exercise in hours.

- **Natural variability**: The simulation incorporates higher variability ($\eta_t \sim \mathcal{N}(0, \sigma^2)$) for power output compared to heart rate, with σ values of 5–10 watts depending on the training phase, reflecting the inherent variability in human power production [109].

For interval workouts, the power profile exhibits rectangular patterns with [110]:

- Sharp transitions between work and recovery intervals.

- Slight power increases during intervals (1–2%) modelling initial ATP-PC system contribution.

- Recovery intervals at consistent lower power zones.

- Higher stochastic variability during high-intensity efforts.

Speed Series Generation The framework generates sport-specific speed data that complements heart rate and power simulations through specialized functions for each sport.

`generate_swim_speed_profile()` For swimming activities, the `generate_swim_speed_profile()` function creates realistic pace variations (in seconds per 100m) based on Critical Swim Speed (CSS) and workout type. The swim pace at time t can be represented as:

$$\text{Pace}_t = h(w, \text{CSS}, t) + \delta_t \tag{4.16}$$

where h is the pace generation function based on the type of workout w, the Critical Swim Speed (CSS) and the time t; and δ_t is the stochastic component that models natural pace variations.

The swim pace generation follows these steps:

1. Use CSS as the base pace reference (measured in s/100m)
2. Apply workout-specific pacing strategies:
    - **Interval workouts**: Alternating fast/slow cycles (15% variance from base pace)
    - **Threshold workouts**: Sustained effort at slightly faster than base pace (3% improvement)
    - **Recovery workouts**: Consistent swimming at slower than base pace (10% slower)
    - **Endurance workouts**: Moderate effort with small oscillations (5% slower than base)

`cycling_speed_profile()` For cycling activities, the `cycling_speed_profile()` function translates power output data into realistic speed estimations (km/h) by applying physics-based modelling of cycling dynamics. The function solves the speed using the power equation proposed by Martin *et al.* [111], while neglecting frictional losses in the wheel bearing and drive chain:

$$\text{Power} = F_{\text{aero}} + F_{\text{rolling}} + F_{\text{gravity}} \tag{4.17}$$

where:

- $F_{\text{aero}} = \frac{1}{2}\rho C_d A V^3$ (aerodynamic drag)
- $F_{\text{rolling}} = mgC_{rr}V$ (rolling resistance)
- $F_{\text{gravity}} = mg\,\sin(\alpha)V$ (gravitational resistance on slopes)

with ρ as air density, $C_d A$ as drag area, m as combined mass (rider + bike), g as gravitational acceleration, C_{rr} as coefficient of rolling resistance, V as velocity and α as angle of gradient of the road.

The implementation uses numerical methods (specifically, `fsolve`) to solve this equation for speed at each time point.

`running_pace()` For running activities, the `running_pace()` function converts heart rate data into realistic running pace values (min/km) based on the athlete's FTPace (functional threshold pace) and heart rate zones. The pace estimation can be represented as:

$$\text{Pace}_t = \text{FTPace} \cdot f(\text{HR}_t, \text{zones}) \cdot (1 + k_f \cdot f) \tag{4.18}$$

where f is the zone-based pace adjustment function, HR_t is the heart rate at time t, zones represents the limits of the heart rate zone of the athlete, k_f is the fatigue coefficient and f is the level of fatigue of the athlete.

The pace determination process follows these steps:

1. Identify the current heart rate zone based on the athlete's zone structure
2. Apply zone-specific pace factors relative to FTPace [112]:
   - Zone 1–2 (Recovery/Easy): 25% slower than threshold pace
   - Zone 3 (Aerobic): 15% slower than threshold pace
   - Zone 4 (Threshold): At threshold pace
   - Zone 5 (VO_2max): 15% faster than threshold pace
   - Above Zone 5 (Sprint): 30% faster than threshold pace
3. Apply fatigue adjustment factor based on the accumulated training load.

Finally, each completed activity is characterized by the parameters specified in Table 4.4, following Garmin [113].

Table 4.4: Wearable Device Activity Data Parameters

Parameter	**Type**	**Description**
Identification Variables		
athlete_id	Categorical	Unique identifier for each athlete

Continued on next page

Table 4.4 – continued from previous page

Parameter	Type	Description
date	Date	Date of activity record
Workout Characteristics		
sport	Categorical	Type of sport (bike, run, swim, etc.)
workout_type	Categorical	Classification of workout (Endurance Ride, Recovery Run, etc.)
duration_minutes	Numeric	Total duration of workout in minutes
tss	Numeric	Training Stress Score
intensity_factor	Numeric	Measure of workout intensity relative to athlete's threshold
intensity_variability	Numeric	Variation in intensity throughout the workout
work_kilojoules	Numeric	Total work done during activity measured in kilojoules
elevation_gain	Numeric	Total elevation gained during activity
Physiological Measures		
avg_hr	Numeric	Average heart rate during activity
max_hr	Numeric	Maximum heart rate achieved during activity
hr_zones	Categorical	Distribution of time spent in different heart rate zones
power_zones	Categorical	Distribution of time spent in different power output zones
Performance Metrics		
distance_km	Numeric	Total distance covered in kilometers
distance_m	Numeric	Total distance covered in meters
avg_speed_kph	Numeric	Average speed in kilometers per hour (km/h)
avg_power	Numeric	Average power output during activity
normalized_power	Numeric	Normalized power accounting for intensity variations
avg_pace_min_km	Numeric	Average pace in minutes per kilometer
avg_pace_min_100m	Numeric	Average pace in minutes per 100 meters (swimming)
Training Effect Indicators		
training_effect_aerobic	Numeric	Impact of workout on aerobic fitness development

Continued on next page

Table 4.4 – continued from previous page

Parameter	Type	Description
training_effect_anaerobic	Numeric	Impact of workout on anaerobic fitness development

4.1.4.4 Training Load Metrics Calculation

Following training execution, the simulation updates several training load metrics using an HRV-adjusted TSS model, following the insight that HRV is a sensitive indicator of training stress and recovery [68].

$$\text{Adjusted TSS} = \text{Raw TSS} \cdot \frac{\text{HRV}}{\text{Baseline HRV}} \tag{4.19}$$

- Fitness (Chronic Training Load): 28-day exponentially weighted moving average (EWMA) [114] of HRV-adjusted TSS

$$\text{Fitness} = \text{EWMA}(\text{Adjusted TSS}last28days, \lambda_{\text{chronic}}), \tag{4.20}$$

$$\text{where } \lambda_{\text{chronic}} = \frac{2}{28+1} \tag{4.21}$$

- Fatigue (Acute Training Load): 7-day EWMA of HRV-adjusted TSS

$$\text{Fatigue} = \text{EWMA}(\text{Adjusted TSS}last7days, \lambda\text{acute}), \tag{4.22}$$

$$\text{where } \lambda_{\text{acute}} = \frac{2}{7+1} \tag{4.23}$$

- Form: The difference between fitness and fatigue, indicating training readiness

$$\text{Form} = \text{Fitness} - \text{Fatigue} \tag{4.24}$$

- Acute Workload Ratio (ACWR): The ratio of fatigue to fitness

$$\text{ACWR} = \frac{\text{Fatigue}}{\text{Fitness}} \tag{4.25}$$

The fitness, fatigue and form model is adapted from the TrainingPeaks Performance Management Chart [115], a widely used platform to track and prescribe endurance training. The primary modification in this implementation is the use of a 28-day window to calculate fitness, rather than the standard 42-day period. However, this is not a novel approach; other platforms, such as Garmin, also use a 28-day window in their chronic training load calculation [116].

4.1.4.5 Evening Recovery Status Simulation

Next, the framework simulates recovery metrics for the evening that reflect the cumulative impact of training and daily stressors.

Algorithm 3 Evening Recovery Status Simulation

1: Calculate physiological and lifestyle stress factors
2: Quantify daily stress score based on weighted factor combination
3: Calculate energy depletion and evening body battery level
4: Update athlete state with evening recovery metrics

Stress Factor Calculation The simulation computes a comprehensive stress score on a 0-100 scale by integrating multiple physiological and lifestyle factors:

- **Biometric Factors:** Deviations in HRV and resting heart rate from baseline values
- **Recovery Factors:** Sleep quality and morning body battery values
- **Training Factors:** Accumulated fatigue from training load
- **Lifestyle Factors:** Athlete-specific stress tendencies, alcohol consumption, and other behavioral patterns

Each factor is normalized to a 0-1 scale and weighted according to its relative importance:

$$\text{Stress} = \sum_i w_i \cdot \text{Factor}_i + \epsilon \tag{4.26}$$

where w_i represents the weight assigned to factor i and ϵ represents day-to-day variability.

The model incorporates non-linear scaling for critical factors, applying exponential weighting when physiological markers deviate significantly from baseline values:

$$\text{Factor}_{\text{HRV}}^{\text{adjusted}} = \begin{cases} \text{Factor}_{\text{HRV}}^{1.5}, & \text{if HRV} < 0.8 \cdot \text{HRV}_{\text{baseline}} \\ \text{Factor}_{\text{HRV}}, & \text{otherwise} \end{cases} \tag{4.27}$$

4.1.4.6 Injury Pattern Injection

The final component of the simulation framework concerns the assignment of realistic injury labels to each day. The injury label is binary, where 1 indicates an injury on a given day and 0 indicates no injury. This component represents one of the most significant challenges in the simulation process, as positive injury labels must be preceded by realistic physiological patterns rather than occurring randomly.

To address this challenge, the framework evolved from a probabilistic model that assigned injury probabilities based on previous recovery and performance data, to a system of scheduled injuries, that once reached lead to retrospective pattern injection (*cf.* Figure 4.6.)

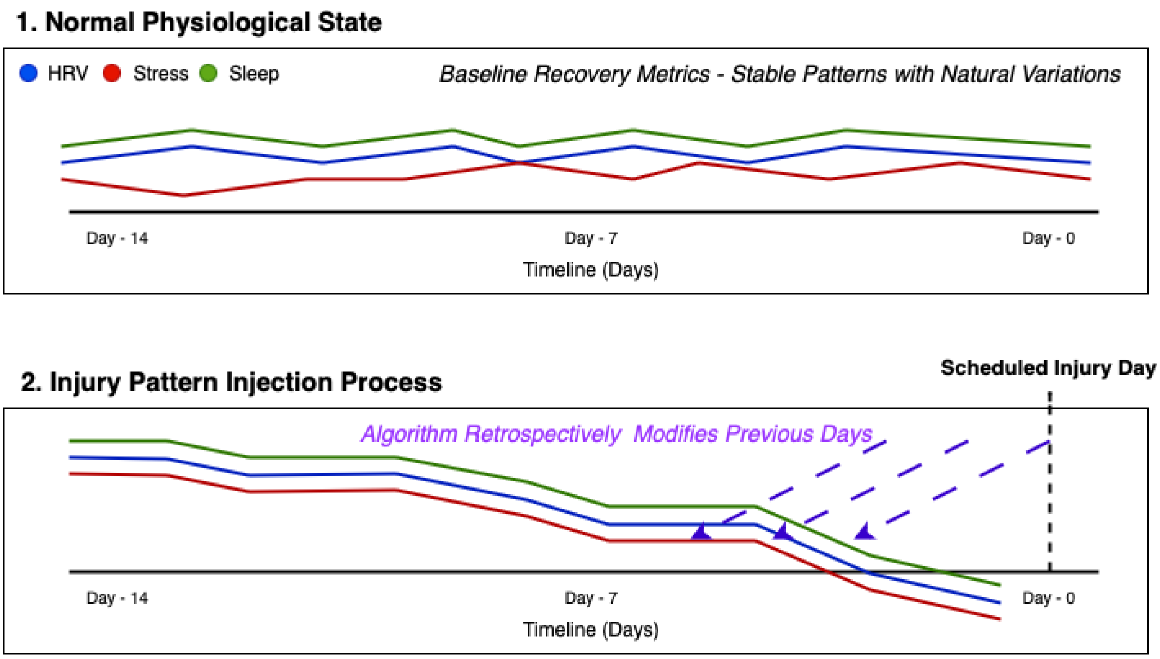

Figure 4.6: Injury Pattern Injection Visualization

These patterns are based on established physiological warning signs that typically precede athletic injuries [117, 67, 118, 68, 119, 120, 121]. Recent research has shown that measurable changes in several biomarkers often occur days or weeks before an injury becomes symptomatic.

The primary physiological metrics that exhibit pre-injury patterns include:

- Heart Rate Variability (HRV) – typically showing a progressive decline [67, 117]
- Resting Heart Rate (RHR) – often demonstrating a gradual increase [118, 68]
- Sleep quality metrics – including disruptions to REM and deep sleep phases [119, 120]
- Stress markers – generally increasing as the body struggles to adapt [121]

Injury Pattern Framework The framework operates on multiple temporal scales:

- **Pattern-Based Injuries:** The system schedules approximately 4.5–5 pattern-based injuries per year, with a variable warning period of 7–14 days.
- **Acute Random Injuries:** Approximately 1 injury per athlete per year occurs without warning signs (with 0.3% daily probability), simulating acute trauma.

- **False Alarms:** The system regularly introduces physiological warning patterns (every 20–35 days) that resemble injury precursors but do not result in actual injuries, challenging any predictive model to distinguish between these cases.

Pattern Injection Methodology When a pattern-based injury is scheduled, the system injects realistic physiological changes through the `inject_realistic_injury_patterns` function:

$$\text{daily\_data} = \text{inject\_realistic\_injury\_patterns}(\text{athlete}, \text{daily\_data}, \text{current\_day\_index}, \text{lookback\_days})$$

This function is called on the scheduled injury day and retrospectively modifies the physiological data of the preceding days to create realistic deterioration patterns. Two fundamental principles enhance the biological validity of these patterns:

1. Individual Variation in Injury Responses A cornerstone of the simulation is the recognition that athletes exhibit substantial heterogeneity in pre-injury physiological signatures. To capture this reality, the system assigns each athlete a unique ”recovery profile” that determines their specific physiological vulnerability signature:

```
# Recovery profile characteristics
recovery_profile = random.choice([
    "hrv_dominant",       # Shows strong HRV changes with fatigue
    "sleep_dominant",     # Shows strong sleep disruption with fatigue
    "rhr_dominant",       # Shows strong RHR changes with fatigue
    "stress_dominant",    # Shows strong stress response with fatigue
    "balanced"            # Shows balanced changes across metrics
])

# Customize physiological sensitivity for each athlete
recovery_signature = {
    "hrv_sensitivity": random.uniform(0.8, 1.2),
    "sleep_sensitivity": random.uniform(0.8, 1.2),
    "rhr_sensitivity": random.uniform(0.8, 1.2),
    "stress_sensitivity": random.uniform(0.8, 1.2)
}

# Adjust sensitivities based on recovery profile
if recovery_profile == "hrv_dominant":
    recovery_signature["hrv_sensitivity"] *= 1.6
```

Listing 4.6: Individual Recovery Profile Generation

This implementation enables personalised injury patterns where athletes exhibit differential sensitivity across physiological metrics. For example:

- An ”HRV-dominant” athlete demonstrates pronounced HRV deterioration before injury while maintaining relatively stable sleep metrics
- A ”sleep-dominant” athlete may show minimal HRV changes but experience significant sleep architecture disruptions
- A ”stress-dominant” athlete exhibits marked elevations in stress biomarkers as the primary warning signal

This approach avoids the oversimplification of assuming universal pre-injury patterns across all athletes.

2. Cross-Stressor Interactions The simulation incorporates the critical phenomenon of multiplicative effects that emerge when multiple stressors coincide. Research shows that combinations of stressors (*e.g.*, poor sleep with high psychological stress) produce effects exceeding the sum of their individual contributions [122, 6]. The system models these interactions through specific multiplier effects:

```python
def calculate_cross_stress_effects(metrics, history=None):
    """Calculate multiplicative effects between different stressors."""
    multipliers = {
        'hrv': 1.0,
        'rhr': 1.0,
        'sleep': 1.0,
        'stress': 1.0,
        'body_battery': 1.0
    }

    # Sleep and stress interaction
    if metrics['sleep_quality'] < 0.6 and metrics['stress'] > 70:
        multipliers['hrv'] *= 1.4  # 40% stronger HRV impact
        multipliers['rhr'] *= 1.3  # 30% stronger RHR impact

    # Temporal sequence effects (if history available)
    if history and len(history) >= 3:
        # High stress followed by high training load
        if (history[-3]['stress'] > 70 and
            history[-2]['stress'] > 70 and
            history[-1]['actual_tss'] > history[-1]['planned_tss'] * 1.1):
            multipliers['hrv'] *= 1.6
            multipliers['sleep'] *= 1.3

    return multipliers
```

Listing 4.7: Cross-Stressor Interaction Calculation

Post-Injury Recovery Simulation Following injury occurrence, the system assigns individualized recovery trajectories:

- **Pattern-based injuries:** Recovery periods of 3–10 days
- **Acute random injuries:** Recovery periods of 3–7 days

Throughout the recovery period:

- Actual training load (`actual_tss`) is substantially reduced relative to planned values, reflecting training modifications
- Physiological markers demonstrate gradual normalization consistent with recovery processes
- The athlete maintains an injured status (`injury = 1`) throughout the designated recovery period
- The probability of new injuries is significantly reduced, implementing a realistic ”injury protection” period

False Alarm Implementation The system regularly introduces false alarm patterns to challenge predictive models and enhance data realism:

$$\text{daily\_data} = \text{create\_false\_alarm\_patterns}(\text{athlete}, \text{daily\_data}, \text{current\_day\_index}, \text{pattern\_days})$$

These patterns exhibit several distinguishing characteristics:

- Duration spans 7–12 days, similar to genuine pre-injury patterns
- Pattern intensity is moderated (typically 0.4–0.8 on a normalized scale), with 30% classified as ”strong” false alarms (0.8–1.1 intensity)
- Fewer concurrent physiological warning signs compared to true injury patterns
- Distinctive resolution trajectory where metrics gradually normalize during the latter half of the pattern duration

This implementation creates the essential challenge of distinguishing between transient physiological fluctuations and genuine injury precursors, a central difficulty in real-world injury prediction.

4.1.4.7 State Transition to Next Day

At the conclusion of each simulated day, the algorithm updates the athlete's state variables, including fatigue level, form value, resting heart rate, HRV, stress accumulation, and energy reserves. These updated values become the initial conditions for the following day's simulation, creating a continuous and interdependent model of training response throughout the annual plan.

4.2 Machine Learning Analysis

This section outlines the design and implementation of the machine learning analysis, following the goal of early injury risk assessment to enable proactive preventive measures.

4.2.1 Data Overview

The final output from the data generation framework discussed in the previous section consists of three interrelated CSV files representing a comprehensive dataset of 1000 athletes over the full calendar year of 2024:

- **athletes.csv**: Contains the foundational athlete profiles with all parameters specified in Table 4.1, including physiological baselines, training experience, and recovery characteristics. Each athlete is assigned a unique identifier that serves as a primary key connecting all datasets.
- **daily_data.csv**: Records day-by-day physiological signals captured by simulated wearable devices. Each record includes metrics such as HRV, resting heart rate, sleep quality, stress levels, recovery indicators and a binary injury label. Every record is linked to its corresponding athlete through the athlete_id field.
- **activity_data.csv**: Documents every completed training activity with detailed metrics including duration, training load, and sport type as well as sport-specific metrics. Similar to the daily data, each activity record contains an athlete_id field establishing the relationship with both the athlete profile and, in combination with the date field, to the corresponding daily physiological data.

For ML analysis, the activity_data.csv file was aggregated by athlete_id and date, such that there was at most a single line per athlete per day. After that, the three files were merged.

4.2.2 Injury Label Engineering

The raw injury data, represented as a binary indicator (1 = injured, 0 = not injured), poses several limitations for predictive modelling. Most notably, this simplistic representation combines both the initial onset of injuries and the subsequent recovery period into a single category, obscuring the distinctive physiological patterns that precede injuries.

To address this limitation, a comprehensive label engineering approach was implemented, (Listing 4.8) that transforms the raw injury indicators into more nuanced and meaningful representations.

```python
def engineer_injury_labels(df, prediction_window=7):
    """
    Transform raw injury labels into more meaningful labels for machine
        learning.
    """
    # Process each athlete separately
    for athlete_id in result_df['athlete_id'].unique():
        # Find injury onset days (where injury=1 but previous day injury=0)
        # Mark injury onset days
        # Label previous 'prediction_window' days as pre-injury
        # Mark recovery days (injured but not onset)
        # Create multi-class injury state
    return result_df

def create_prediction_targets(df, predict_window=7):
    """
    Create binary target column for predicting if injury will occur within
        the next X days
    """
    # For each day, check if injury will occur within prediction window
    return result_df

def create_time_to_injury_feature(df, max_days=28):
    """
    Create a feature representing days until next injury
    """
    # Calculate days until next injury for each record
    return result_df
```

Listing 4.8: Injury label engineering

This approach creates several refined representations of injury status:

- **Injury Onset (Binary):** Identifies only the first day of each injury occurrence, distinguishing initial injury events from the recovery process.
- **Pre-Injury State (Binary):** Labels the 7 days preceding an injury, capturing the critical window when physiological warning signs may be present but before the athlete becomes symptomatic.

- **Recovery Day (Binary):** Identifies days during the recovery period after initial injury onset.
- **Injury State (Multi-class):** Combines the above into a single categorical variable: 0=healthy, 1=pre-injury, 2=onset, 3=recovery.
- **Will Get Injured (Binary):** Forward-looking target variable indicating whether an injury will occur within the next 7 days.
- **Time to Injury (Numeric):** Represents the number of days until the next injury (up to 28 days), offering a continuous measure of injury imminence.

This detailed label engineering approach addresses several methodological challenges. First, it distinguishes between injury onset and recovery, preventing the model from learning patterns associated with the recovery process rather than prediction. Second, the forward-looking ”will_get_injured” target directly aligns with the practical goal of early intervention. Third, by creating a multi-class representation, it enables more nuanced analysis of the model's predictive capabilities across different phases of the injury cycle.

4.2.3 Problem Formulation

The injury prediction task was formulated as a binary classification problem to predict whether an athlete would experience an injury within the next 'x' (predict_window) days (the newly engineered field ”will_get_injured” defined as target). This forward-looking window approach transforms the data from simple historical records into a predictive framework that can provide actionable insights before injuries occur.

4.2.4 Feature Engineering

The engineered features were designed to capture the critical patterns that precede the injury events. As synthetic data was designed to simulate athletes following well-designed training plans that should avoid injury, a special emphasis was placed on recovery metrics, while still including certain key load features suggested by sport science literature.

4.2.4.1 Temporal Features

To capture the time-dependent nature of injury risk, several categories of temporal features were implemented. Rolling window statistics were calculated over periods of 3, 7, and 14 days for key metrics like HRV, training load, and sleep quality to capture both acute and chronic trends. Rate of change metrics, including day-to-day changes and acceleration values, were computed to identify rapid shifts in physiological status that often precede injuries. In addition, consecutive-day counters were created to track persistent patterns such as consecutive days of low HRV or high stress.

4.2.4.2 Normalized Individual Baseline Features

Recognizing the high degree of athlete-to-athlete variability in physiological responses, metrics were normalized to each individual's personal baseline. Ratio features expressed current values relative to the athlete's established baseline (*e.g.,* HRV ratio = current HRV / baseline HRV). Athlete-specific z-scores transformed measurements into standard deviations from the athlete's own average, enabling meaningful comparison of relative deviations across different individuals with varying physiological ranges. Deviation tracking counted days when metrics remained below personal thresholds.

4.2.4.3 Training Load and Recovery Features

Building on established sports science principles, features quantifying both training load and recovery capacity were constructed. The Acute Workload Ratio (ACWR) [123] was calculated using both standard and exponentially weighted variations. Training monotony and strain metrics captured the variability and overall impact of training schedules, while load progression metrics identified potentially risky training progressions. Most importantly, recovery adequacy measures evaluated whether recovery metrics were proportional to the imposed training stress, addressing the fundamental question of whether the athlete's body was successfully adapting to training demands.

4.2.4.4 Cross-Metric Interaction Features

To capture complex interactions between physiological systems, compound features were developed. The features of the sleep-HRV compound identified days when both sleep quality and autonomic nervous system regulation were compromised simultaneously. Load-recovery interactions highlighted imbalances between training stress and recovery capacity. Stress-recovery ratios quantified the relationship between psychological stress levels and physiological recovery metrics, capturing how external stressors might amplify training-related fatigue.

4.2.4.5 Post-Extreme Event Monitoring

The following feature engineering strategies were implemented to monitor the aftermath of high-intensity training sessions and races:

- **Event Classification:** Training sessions were categorized into discrete TSS bands using daily TSS values: *Normal (0-300 TSS)*, *High (300-400 TSS)*, *Very High (400-500 TSS)*, and *Extreme (500+ TSS)*.

- **Post-Event Tracking Features:**
  - `days_since_extreme_tss`: A continuous variable tracking the exact number of days since the athlete's last Very High or Extreme TSS event.
  - `post_extreme_period`: A binary indicator marking the critical 7-day window following extreme events.
  - `post_extreme_hrv_deficit`: Quantifies the degree of HRV suppression relative to the athlete's baseline during the post-event window, calculated as the product of the post-extreme period indicator and the HRV ratio deficit (1 - HRV ratio).
  - `post_extreme_sleep_deficit`: Measures sleep quality impairments following high TSS events, implemented as an interaction between the post-extreme period indicator and the poor sleep binary flag.

4.2.4.6 Composite Risk Scores

To synthesize information from multiple features and create more interpretable indicators, several composite scores were developed. A recovery composite combined weighted values of HRV ratio, sleep quality, and resting heart rate metrics. An acute risk estimate aggregated various risk factors using weights determined through correlation analysis with injury outcomes. Critical window alerts were triggered when multiple recovery metrics showed concerning patterns simultaneously, providing a simplified way to identify high-risk periods that might warrant intervention.

4.2.5 Data Preprocessing

Given the synthetic nature of the data set, the pre-processing requirements were minimal compared to those encountered in real world data. However, several pre-processing techniques were applied to transform the synthetic dataset into a format suitable for machine learning algorithms.

Temporal Feature Extraction: Date-based features were transformed to extract cyclical patterns relevant to training and injury prediction. For each *datetime* column, day-of-week (0-6) was extracted to capture weekly training cycles. Also, seasonal information was derived by converting month values into quarterly seasons (1-4), which may influence training environments and injury susceptibility. Original *datetime* columns were subsequently removed to avoid redundancy and prevent models from learning spurious correlations.

Dictionary Flattening: The dataset contained nested dictionary structures for various training zone metrics, including:

- Heart rate zones for different activities (bike, run, swim, strength)
- Power zones for cycling activities

These nested structures were systematically flattened through a process that:

- Safely converted string representations to Python dictionaries
- Identified all possible keys across the dataset
- Created individual columns for each dictionary key with appropriate naming conventions
- Populated missing values with zeros to maintain data consistency

The original dictionary columns were removed after extraction to maintain a clean tabular format.

Categorical Feature Encoding: Categorical variables were encoded according to their semantic properties:

- **Binary categorical variables** (*e.g.,* gender): Encoded using one-hot encoding with the first category dropped to avoid multicollinearity
- **Nominal categorical variables** (e.g., lifestyle): Transformed using one-hot encoding with the first category dropped
- **Ordinal categorical variables**: Encoded using ordinal encoding to preserve inherent order:
  - Heart rate variability zones (below normal, normal, above normal)
  - Training adherence (undertraining, on target, overtraining)
  - Acute Workload Ratio risk levels (too low, optimal, danger zone, high risk)
  - Ramp rate risk (decreasing, optimal, warning, high risk)
  - Training load classification (normal, high, very high, extreme)

This approach ensured that the categorical information was appropriately represented while maintaining the semantic relationships within the data.

Feature Cleanup and Missing Value Handling: Several cleanup procedures were implemented to ensure data quality:

- Non-numeric identifiers (athlete_id) were removed as they provide no predictive value
- Null values in engineered features were filled with zeros

- Consistently null columns resulting from the dictionary flattening and activity_data.csv aggregation (primarily related to sport-specific metrics *e.g.,* elevation gain for swimming activities) were identified and removed from the dataset

4.2.6 Data Splitting Methodology

Two different approaches were defined for splitting the data into training and testing sets:

- **Athlete-based splitting:** Use 800 athletes' whole year data for training and 200 for testing.
- **Time-based splitting:** Use data from January through October for training, data from November and December for testing between all athletes.

The selection between these two splitting strategies represents a fundamental decision regarding the primary purpose of the predictive model: should it generalize to entirely new athletes or should it predict future injuries for athletes whose past data are already known?

Athlete-Based Splitting - Testing Generalizability: The athlete-based split strategy allocates complete annual profiles of 800 athletes (80%) to the training set, while reserving data from 200 athletes (20%) for the testing set. This approach deliberately tests the model's ability to generalize injury patterns to completely unseen individuals.

Time-Based Splitting - Testing Future Prediction: In contrast, the time-based approach divides the data set temporally, using data from January to October (approximately 83%) for training and data from November to December (approximately 17%) for testing between all athletes. This strategy evaluates the model's capability to predict future injuries for athletes whose historical data has already been analyzed.

4.2.7 Model Selection and Implementation

Model Selection: The model selection was informed by related work, and limited to 3 algorithms:

- **Random Forest [53]:** An ensemble learning method that constructs multiple decision trees and aggregates their predictions [124]. The implementation follows Petrovsky et al. [42], who applied this algorithm to predict recovery using physiological indicators.
- **LASSO [52]:** Least Absolute Shrinkage and Selection Operator regression performs simultaneous variable selection and regularization by imposing an L1 penalty that forces certain coefficients to zero. In the injury prediction context, [43] found that LASSO reduced the complexity of the model while maintaining predictive performance.

- **XGBoost [51]:** eXtreme Gradient Boosting implements gradient boosting [125] with optimizations for computational efficiency. The algorithm sequentially builds models to correct previous errors and combines them for final prediction. XGBoost has been validated in injury prediction [8].

Implementation Details: All models were implemented in Python 3.8 using scikit-learn (1.6.1) [126] for the LASSO and Random Forest algorithms, and XGBoost (3.0.0) [127] for the gradient boosting implementation. The LASSO model employed logistic regression with L1 regularization using the 'liblinear' solver, configured with a convergence tolerance of 0.001 and maximum 1000 iterations. For the Random Forest classifier, an ensemble of 200 trees was constructed with a maximum depth of 8 and minimum samples per leaf set to 7 to balance complexity and generalization. The XGBoost implementation utilized 400 estimators with shallow trees (maximum depth of 2), a conservative learning rate of 0.03, and subsampling parameters (0.8 for observations, 0.7 for features) to mitigate overfitting. All models incorporated class weighting to address the substantial imbalance between injury and non-injury observations, with weights inversely proportional to class frequencies.

5 Evaluation

This chapter presents a comprehensive evaluation of the synthetic data generation framework and machine learning analysis for injury prediction in triathletes. The evaluation examines the quality of the generated data, the patterns discovered through exploratory data analysis, and the predictive performance of the machine learning models.

5.1 Evaluation Setup

5.1.1 Alignment with Research Objectives

The evaluation is structured to address the four primary research objectives established in Chapter 1:

1. To develop a simulation framework capable of generating physiologically plausible athlete data, including training loads, recovery metrics, and injury occurrences preceded by realistic physiological warning patterns
2. To ensure that the simulated training programs adhere to established sports science principles for load management, thereby creating a realistic environment where injury occurrence is predominantly influenced by recovery factors
3. To implement and evaluate multiple machine learning algorithms for injury prediction using the synthetic dataset, with particular emphasis on identifying early warning signs that precede overuse injuries
4. To demonstrate how synthetic data generation can serve as a foundational methodology to overcome data privacy regulations and missing labels

For each objective, the evaluation provides specific analyses and metrics to assess the success of the achievement.

5.1.2 Evaluation Methodology

The evaluation follows a three-stage approach:

1. **Exploratory Data Analysis (EDA)**: Examines the synthetic data to validate its physiological plausibility, and the presence of realistic pre-injury patterns
2. **Machine Learning Experiments**: Tests different model architectures and data splitting strategies to assess predictive performance for injury risk
3. **Critical Assessment**: Evaluates the limitations of the synthetic data approach and its potential for real-world applications

5.2 Exploratory Data Analysis

EDA serves dual purposes in this thesis: validating the quality of the synthetic data and discovering physiological patterns that precede injuries. This section presents key findings from this analysis.

5.2.1 Validation of Physiological Plausibility

This subsection addresses the part of the first research objective of this thesis, concerned with the physiological plausibility of the generated data:

To develop a simulation framework capable of generating physiologically plausible athlete data, including training loads, recovery metrics, [...]

Demonstrating physiological plausibility requires establishing that the synthetic data not only falls within appropriate numerical ranges but also exhibits the complex interdependencies and response patterns observed in actual athlete populations. To comprehensively validate the biological validity of the simulated data, three complementary analyses were conducted:

1. Distribution analysis of key physiological metrics to verify that values fall within clinically validated ranges for competitive age-group triathletes
2. Correlation analysis between related physiological parameters to confirm that the simulated relationships reflect known physiological mechanisms rather than random associations
3. Response pattern analysis to determine whether the simulated physiological systems react to training stimuli in a manner consistent with established exercise physiology principles

Together, these analyses form a multifaceted validation framework that examines the synthetic data from both static (distributions), relational (correlations), and dynamic (response patterns) perspectives, providing robust evidence for its physiological plausibility.

5.2.1.1 Distribution of Physiological Metrics

The frequency distribution analysis of key physiological metrics confirmed that the synthetic data generated values within clinically validated ranges for competitive age-group triathletes, closely mirroring metrics typically captured by wearable devices. Figure 5.1 illustrates the distribution patterns of heart rate variability (HRV), resting heart rate (RHR), sleep quality, and stress scores across the athlete population.

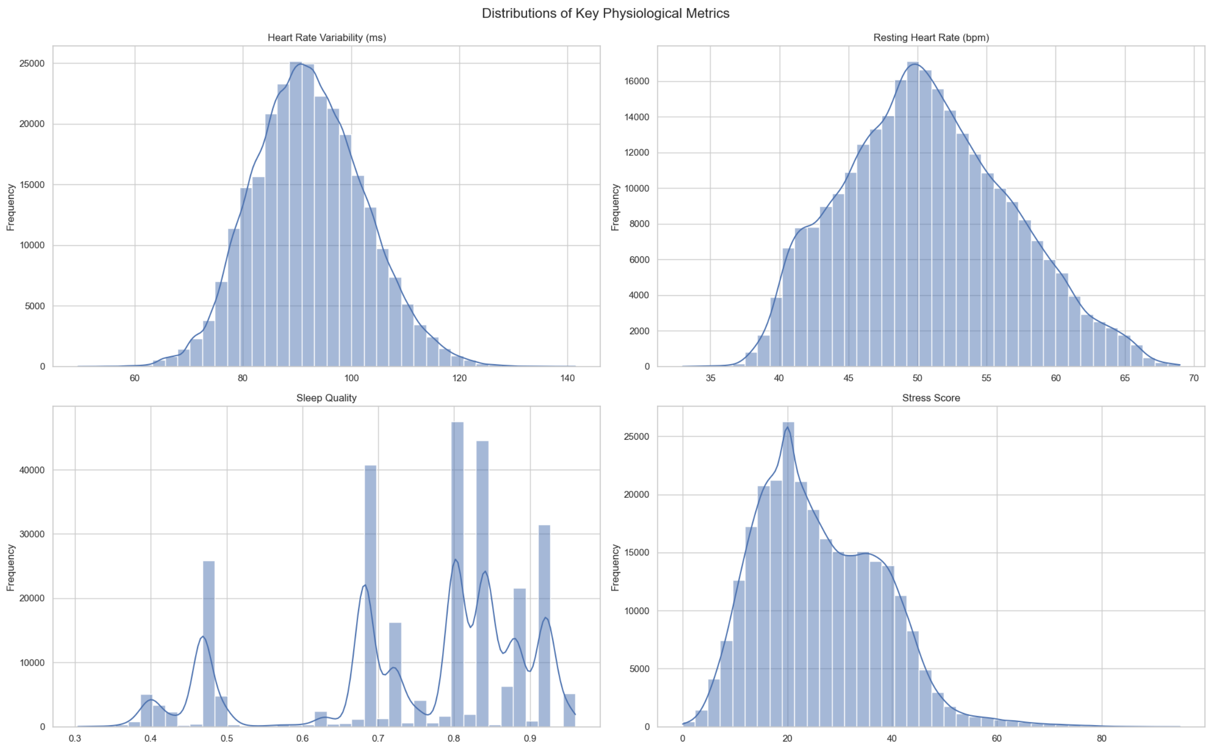

Figure 5.1: Distribution of Key Physiological Metrics in Competitive Age-Group Triathlete Population

Heart Rate Variability (HRV) The findings revealed a normal distribution of 92.3ms (SD = 10.17ms) ranging from 49-141ms. This is in line with reported reference values in endurance-trained athletes by Plews et al. [67], which reported that increased HRV (>80ms) typically indicates increased parasympathetic modulation and improved cardiovascular fitness. The moderate positive skewness (0.19) also reflects the anticipated physiological profile of competitive age-group triathletes and suggests that, although most values are concentrated near the mean, higher values occur more frequently than lower values.

Resting Heart Rate (RHR) The right skewness of 0.24 with a mean of 48-50 beats per minute (bpm) accurately reflects the bradycardic adaptation typical of endurance athletes. The observed range (35-65 bpm) spanned the expected range, and 92.46% of the values were less than the clinical threshold for bradycardia in the general population (<60 bpm). This pattern is consistent with longitudinal endurance athlete research [128], in which higher parasympathetic tone and higher stroke volume contribute to normally lower resting heart rate.

Sleep Quality The distribution of sleep quality is trimodal, with sharp clusters around 0.5, 0.7 and 0.82 on a normalised scale (sleep scores of 50, 70 and 82). These modes represent three discrete states of recovery: poor, adequate and optimal, respectively. The mean is slightly below 0.75, valid considering that Garmin reported an average sleep score of 71 across their users in 2024 [129]. Quantitative analysis showed that 59.13% of the observations were in the optimal recovery range (>75), 24.46% were in the adequate range (60-75) and 16.40% were in the poor recovery range (<60). This is consistent with the lifestyle profiles introduced in the previous chapter (subsection 4.1.2), where ”Highly Disciplined Athlete”, ”Balanced Competitor”, and ”Health-Conscious Athlete” are likely to fall mostly in the optimal or sparsely in the adequate clusters, making up 65% of the population. The ”Weekend Socialiser” (12%), characterised by poor sleep on weekends and good sleep on week-days, might fall into the adequate cluster, and the ”Sleep-deprived Workaholic” and ”Under-Recovered Athlete” (12% and 11% respectively) would fall into the poor to adequate cluster.

Stress Score The bimodal profile with peaks predominantly at 20 and 35 (on Garmin’s 0-100 scale), and a mean of 26.17 (SD = 11.86), close to Garmin’s reported user average of 30 in 2024 [129]. Also, the positive skewness (0.46) accurately captures the physiological reality that chronic conditions of high stress (scores above 50) are impossible to maintain and hence relatively infrequent (2.6% of measurements).

5.2.1.2 Correlations Between Metrics

Having confirmed that the simulated physiological metrics are realistically distributed, it is necessary to ensure that they are related to each other in a realistic rather than random way. The correlation matrix shown in Figure 5.2 confirms this by visualising how strongly heart rate variability (HRV), resting heart rate (RHR), sleep quality, actual training stress score (TSS), stress and morning body battery are correlated.

Heart Rate Variability (HRV) Exhibits a moderate negative correlation with RHR (r = -0.23), which shows that a decline in HRV is correlated with an increased RHR. This relationship is plausible considering that both are indicators of suppressed recovery [118]. Additionally, HRV shows a positive correlation with body battery morning, a composite revealing energy levels in the morning (r = 0.31), also a sound correlation [67].

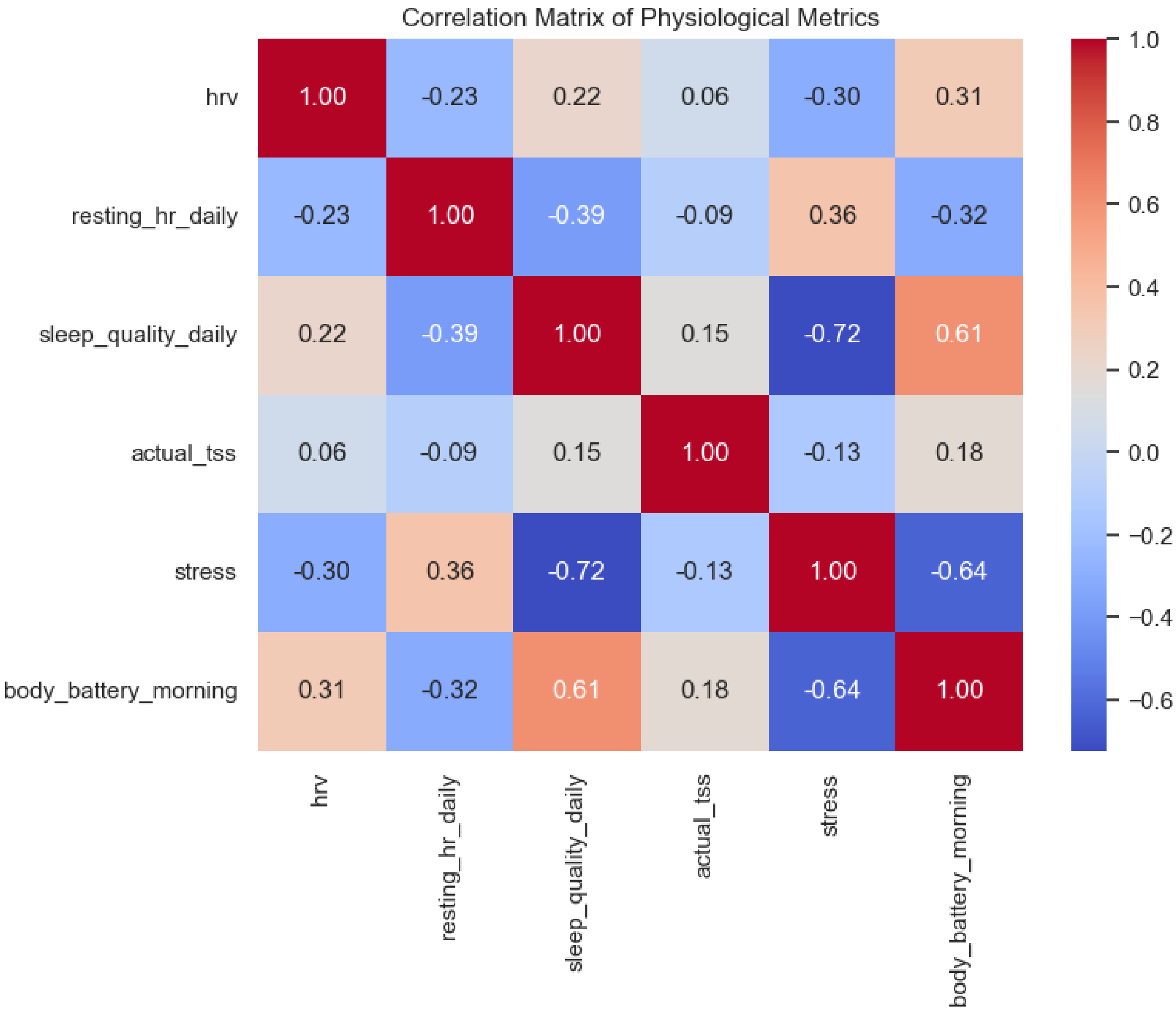

Figure 5.2: Correlation Matrix of Physiological Metrics

Resting Heart Rate (RHR) Demonstrates expected correlations with other recovery metrics, showing positive association with stress levels (r = 0.36) and negative associations with sleep quality (r = -0.39) and morning body battery (r = -0.32). These relationships accurately capture how sympathetic nervous system activation during periods of psychological stress or inadequate recovery leads to elevated resting heart rates, a well-documented phenomenon in athletic populations [130].

Sleep Quality The strongest correlation in the matrix appears between sleep quality and stress (r = -0.72). This strong negative correlation is physiologically sound, as elevated stress hormones are known to disrupt sleep architecture, while poor sleep quality contributes to increased physiological and psychological stress [131]. Similarly, sleep quality shows a robust positive correlation with morning body battery (r = 0.61), appropriately modelling how morning energy is determined by the quality of the night's sleep [120].

Actual Training Stress Score (TSS) Exhibits relatively modest correlations with the physiological metrics, showing slight positive correlations with sleep quality (r = 0.15) and body battery morning (r = 0.18), and a slight negative correlation with stress (r = -0.13). This indicates that better recovery metrics are correlated with higher training loads, which is likely a direct consequence of how training execution probabilities were modeled.

Stress Score Shows consistent correlations across all variables, with strong negative relationships with sleep quality (r = -0.72) and body battery morning (r = -0.64), and a moderate positive correlation with resting heart rate (r = 0.36). These patterns appropriately capture how stress serves as both a contributor to and outcome of recovery status, creating physiologically realistic feedback loops in the synthetic data [120].

Collectively, these correlation patterns confirm that the synthetic data generation incorporates physiologically plausible relationships between metrics rather than treating each variable as an independent parameter.

5.2.1.3 Training Response Patterns

The final step in ensuring the quality and realism of the synthetic dataset is to evaluate the physiological response to training loads so that the synthetic data accurately reflects real fatigue and recovery cycles. Figure 5.3 illustrates how key physiological measures respond to different training stimuli by comparing the percentage change from baseline after high-load training sessions (top 10% of TSS values) versus normal-load sessions (40-60th percentile of TSS).

Three observed physiological indicators - heart rate variability (HRV), resting heart rate (RHR) and sleep quality - show different load-dependent trends, with one commonality: severe and prolonged disturbances are induced by high training loads compared to normal loads.

Heart Rate Variability (HRV): HRV shows a sharp decrease of about 3% the day after high intensity exercise. This is accompanied by a gradual recovery, but levels remain below baseline (approximately -0.5%) even on day 6. However, HRV after normal exercise remains fairly stable, always within 0.5% of baseline, reflecting less autonomic dysfunction. These patterns align with established research showing that HRV suppression is both intensity- and duration-dependent, with greater disturbances following more demanding sessions [132, 133].

Resting Heart Rate (RHR): RHR increases significantly - by more than 1.5% - on the first day after heavy exercise. It decreases steadily over the next few days, stabilising at around 0.4% above baseline on day 5. Under normal exercise conditions, RHR remains close to baseline, with small fluctuations not exceeding 0.25%. This load-dependent response mirrors findings that

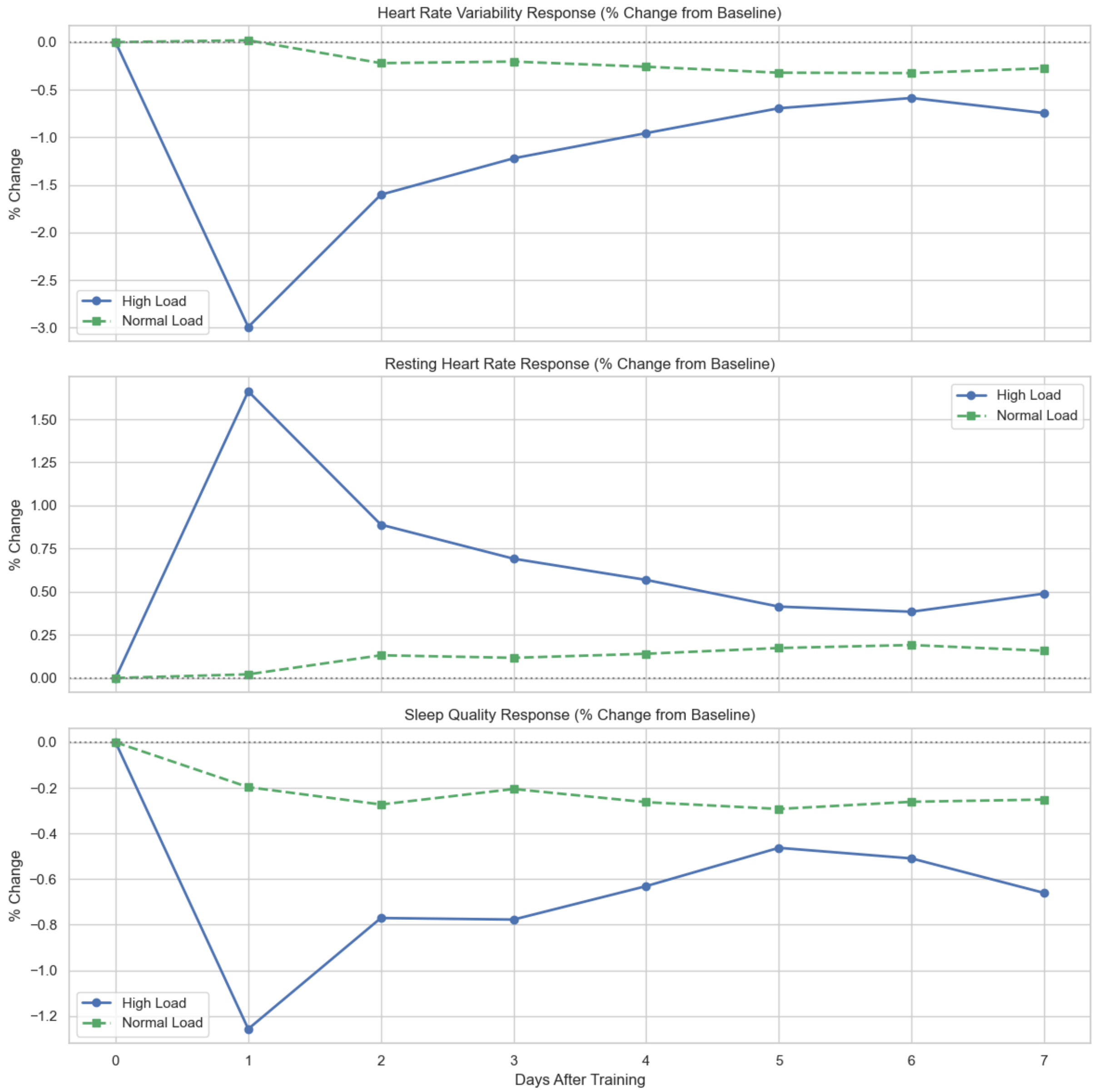

Figure 5.3: Physiological Response (% Change from Baseline) After Heavy vs Normal Training

resting heart rate elevations can serve as a marker of accumulated training stress and incomplete recovery [134, 135].

Sleep quality: Sleep quality scores drop by about 1.25% shortly after high load sessions, and a moderate deficit is evident until days 2 and 3. Recovery occurs by days 4-5 to a level close to -0.4% of baseline. Normal load sessions result in a small decrease in sleep quality (about -0.2%), with negligible disturbance. With an average normalised score of 0.75, a difference of 0.2% corresponds to a score of 0.7485. These findings are consistent with research demonstrating that high training loads can disrupt sleep architecture and quality, while moderate training has minimal impact [136, 137]

These trends confirm that the synthetic data accurately reflects physiological stress and recovery cycles according to different training loads.

5.2.2 Pre-Injury Physiological Patterns

In this subsection, the second part of the first research objective of this thesis is evaluated: *[...], and injury occurrences preceded by realistic physiological warning patterns.*

Injuries in endurance sports rarely occur without warning. Rather, they typically emerge following a cascade of subtle physiological changes that may manifest days or even weeks before the injury becomes clinically apparent. The presence of these pre-injury patterns in the synthetic dataset, is crucial for developing effective early warning systems that could be used in real-world applications to allow athletes and coaches to intervene before tissue damage occurs.

This analysis aims to identify pre-injury physiological patterns in the synthetic data and assess their ecological validity by comparing them with patterns documented in sports medicine literature. By examining how various metrics change in the days and weeks preceding an injury event, we can validate that the simulation framework captures the complex physiological deterioration that typically precedes tissue failure. This validation is essential for ensuring that machine learning models trained on this data will generalize effectively to real-world scenarios.

5.2.2.1 Feature Correlation with Injury

Figure 5.4 illustrates the point-biserial correlation coefficients between various physiological metrics and injury occurrence in the synthetic dataset. Point-biserial correlation is particularly appropriate for this analysis as it measures the relationship between a continuous variable and a binary outcome (in this case, injury occurrence).

The results reveal a clear pattern where recovery metrics demonstrate stronger associations with injury risk than training load metrics. Deep sleep exhibits the strongest negative correlation (approximately -0.19), followed by REM sleep (approximately -0.12) heart rate variability (HRV) (approximately -0.11) and sleep quality (approximately -0.07). Body battery in the morning, a composite expressing the energy levels in the morning, also shows a moderate negative correlation (approximately -0.05), which is positively correlated with sleep metrics (as seen in 5.2.1.2).

These findings are consistent with research by Copenhaver and Diamond [138], who found that sleep disturbance was significantly associated with injury risk in athletes, Plews et al. [67], who demonstrated that decreased HRV often precedes overreaching states in endurance athletes, and [73], who identified disturbed sleep patterns as one of the earliest indicators of functional overreaching that often precedes injury.

In contrast, metrics associated with increased physiological load show positive correlations with

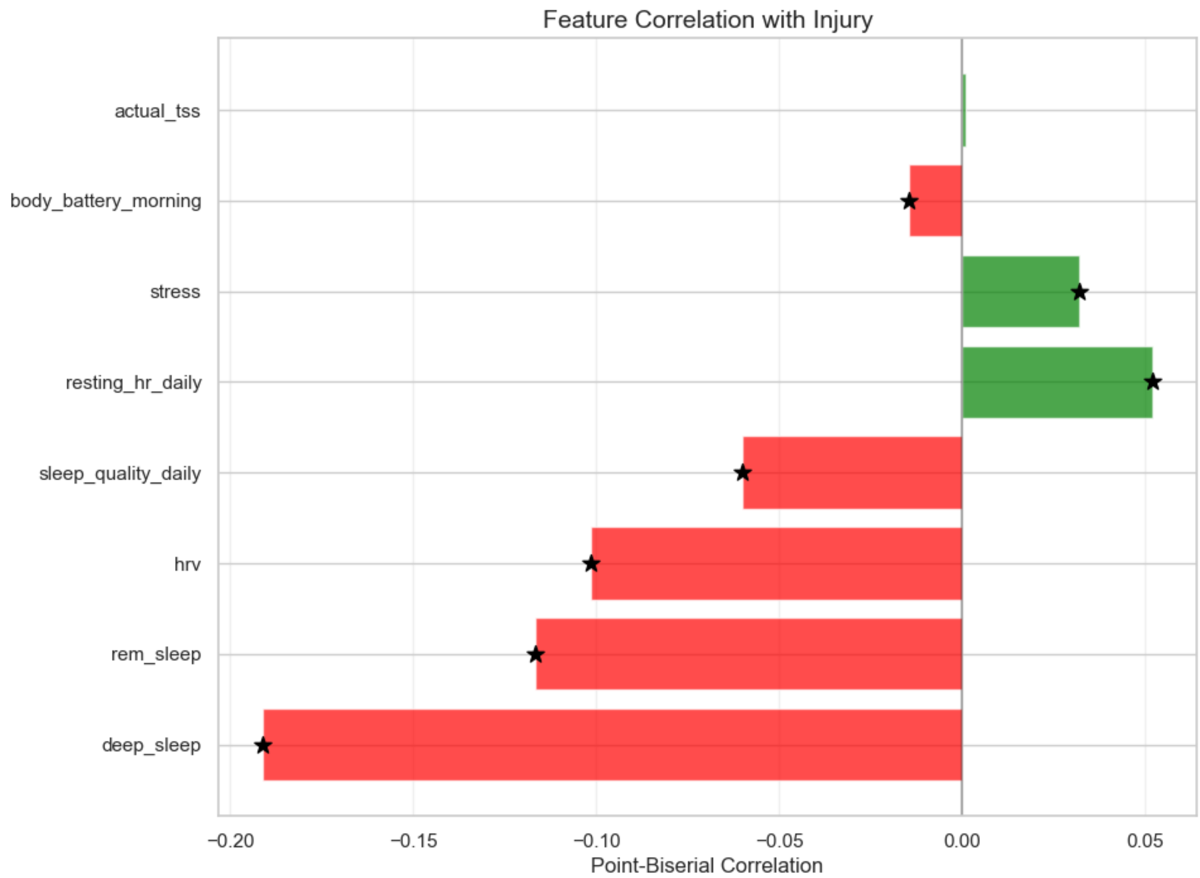

Figure 5.4: Correlation of various Features with Injury

injury risk. Resting heart rate demonstrates the strongest positive correlation (approximately 0.06), followed by stress levels (approximately 0.04). These positive correlations support work by Hellard et al. [139], who found elevated resting heart rate to be a significant predictor of illness in swimmers, and Mann et al. [140], who established relationships between psychological stress and injury in athletes.

Notably, actual training load (actual_TSS) shows only a minimal positive correlation with injury occurrence, an indication that the second research objective of *creating a realistic environment where injury occurrence is predominantly influenced by recovery factors* was met (deeper analysis in subsection 5.2.3).

5.2.2.2 Temporal Progression of Warning Signs

Figure 5.5 illustrates the temporal progression of key recovery metrics (identified in 5.2.2.1), in the two-week period preceding injury occurrences across the synthetic dataset. Only sleep quality is displayed in the plot, as this is directly determined by sleep stage distribution and duration. This time series analysis provides valuable insights into how physiological warning signs develop and escalate prior to injury manifestation.

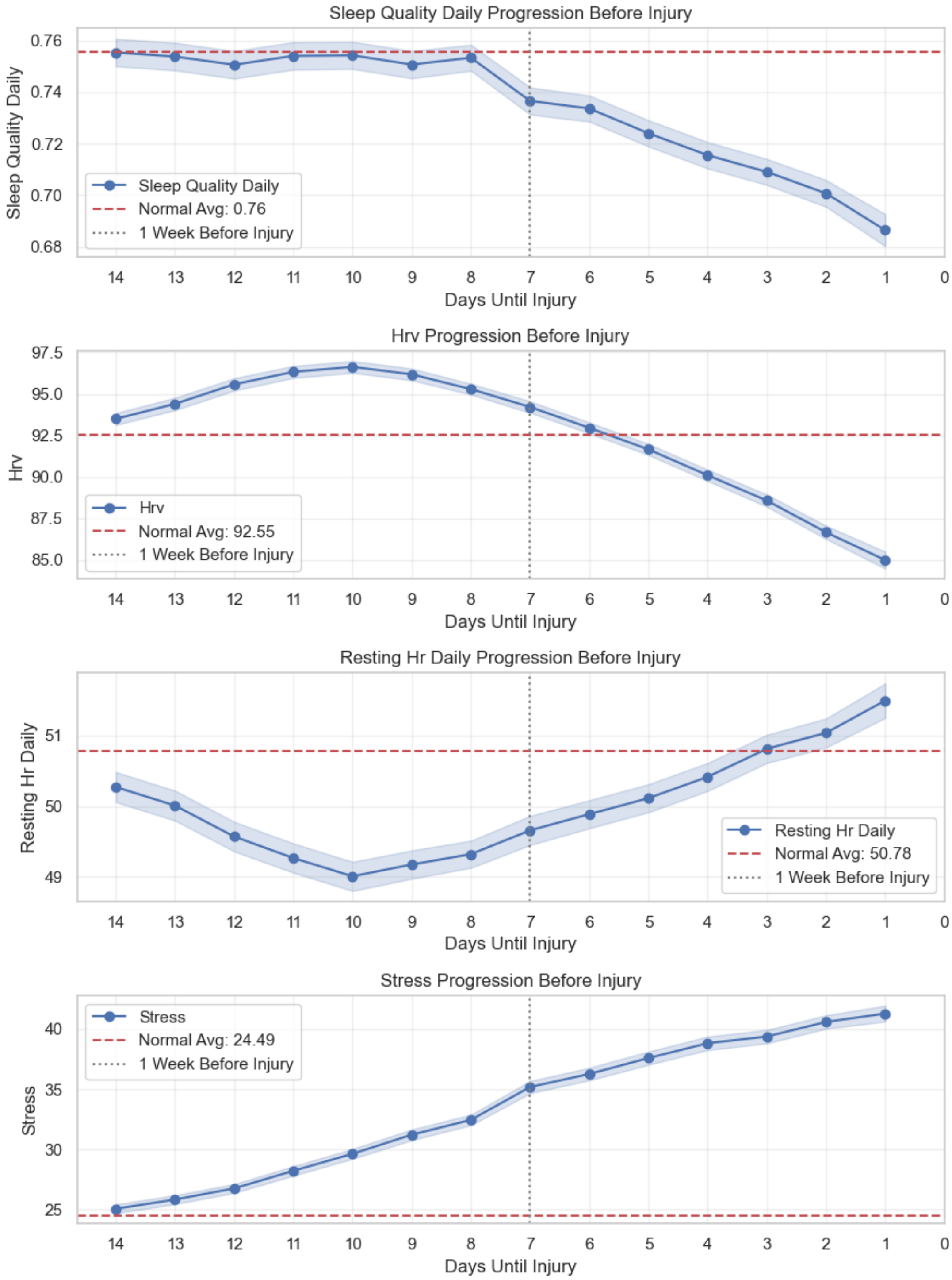

Figure 5.5: Time Series Analysis of Recovery Metrics Before Injury

The progression patterns reveal distinct physiological changes approximately 7 days before injury occurrence (marked by the vertical dotted line), suggesting this may represent a critical transition point in athlete recovery status. Most notably, HRV demonstrates a biphasic pattern characterized by supranormal values (days 14-8) followed by a progressive decline toward injury. This pattern aligns with findings from Plews et al. [67], who describe how both increases and decreases in HRV can indicate maladaptive responses in elite athletes, suggesting that interpreting HRV changes requires understanding the individual's baseline and current training context.

The RHR data shows an inverse relationship to HRV, with values initially below normal average (days 14-10), followed by a gradual increase beginning around day 7, and eventually exceeding the normal average approximately 3 days before injury. This reciprocal relationship between HRV and RHR is consistent with the autonomic nervous system regulation described by Buchheit [118], who demonstrated that parasympathetic withdrawal typically manifests as both decreased HRV and increased RHR during periods of physiological strain.

Sleep quality demonstrates a relatively stable pattern until day 7, after which a steady decline is observed. This degradation in sleep quality coincides with the rise in stress levels, which show a continuous increase throughout the two-week period, with acceleration in the final week. The concurrent deterioration of sleep quality and escalation of stress is consistent with research by Hausswirth et al. [136], who found that functional overreaching in endurance athletes was associated with both increased perceived stress and decreased sleep quality.

The collective pattern suggests a potential sequence of physiological events preceding injury: initial autonomic dysregulation (reflected in elevated HRV) followed by progressive deterioration in recovery capacity (declining HRV, increasing resting heart rate), accompanied by worsening sleep quality and rising stress levels. This supports the hypothesis that monitoring these interconnected physiological markers may provide an early warning system for injury risk, potentially allowing for intervention during the critical transition period around day 7.

5.2.2.3 Individual Variation in Warning Patterns

While Figure 5.5 illustrated the average progression of recovery metrics before injury, Figure 5.6 reveals the substantial individual variation that exists within these patterns. Each thin line represents an individual athlete's physiological response in the two weeks preceding injury, with the thicker lines showing the population averages for each metric.

A critical challenge when using synthetic data for machine learning applications is avoiding overly ”clean” patterns that fail to represent real-world complexity. To be valuable for training predictive models, simulated data must capture appropriate individual variation; otherwise, the classification task becomes artificially simple and may not generalize well to actual athlete monitoring scenarios [141]. This visualization demonstrates how the synthetic dataset successfully incorporates significant heterogeneity in physiological responses, creating a more realistic foundation for predictive modelling.

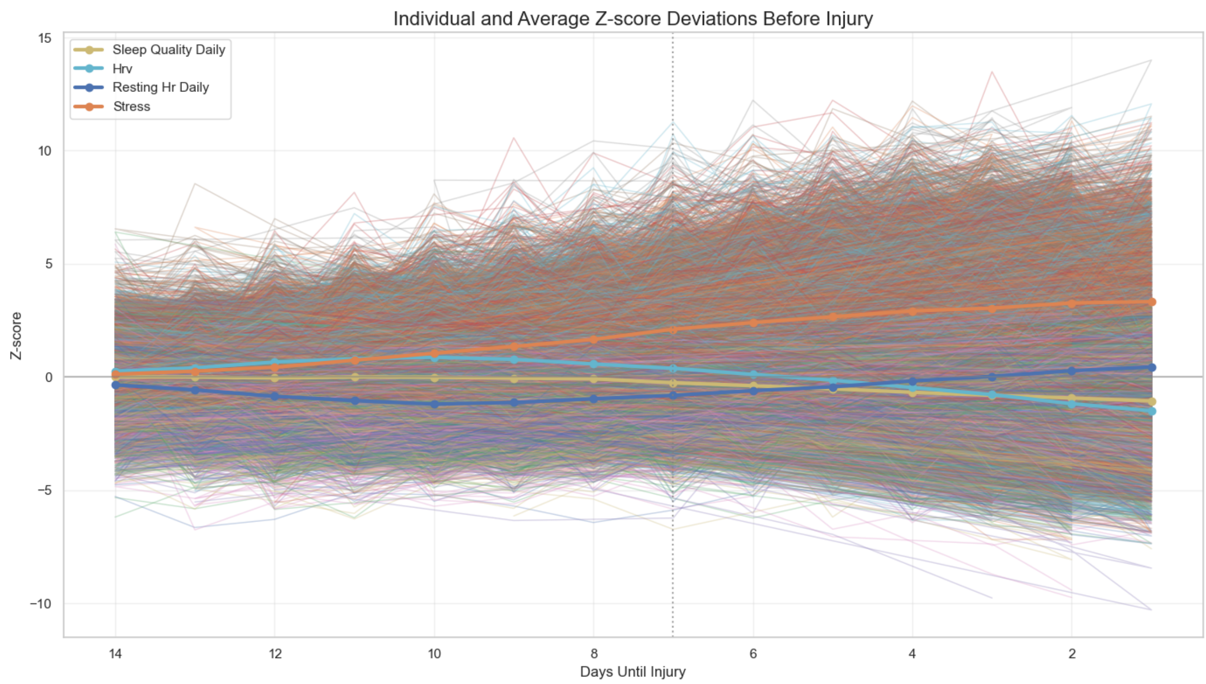

Figure 5.6: Individual and Average Z-score Deviations Before Injury

The visualization demonstrates considerable heterogeneity in how athletes manifest pre-injury physiological changes. Some individuals exhibit extreme deviations (z-scores reaching beyond ±10), while others show more moderate alterations. This variation is particularly pronounced after day 7 (the critical transition point identified earlier), where the spread of individual trajectories widens substantially across all metrics.

For heart rate variability (HRV), while the average trend shows the biphasic pattern discussed previously, individual athletes display diverse responses. Some maintain elevated HRV until just days before injury, while others show early and steep declines.

Sleep quality similarly shows varied responses, with some athletes exhibiting dramatic deterioration in the final days before injury, while others maintain relatively stable patterns.

Resting heart rate demonstrates an inverse relationship to HRV at the population level, but individual trajectories reveal that this relationship is not uniformly expressed across all athletes. Some show the classic pattern of rising heart rate concurrent with falling HRV, while others display more complex interactions between these metrics.

Stress levels show the most consistent directional trend across individuals (generally increasing), yet the magnitude of increase varies substantially. This supports the hypothesis that while stress elevation may be a universal component of the pre-injury state, the threshold at which this elevation becomes problematic likely differs between athletes.

5.2.3 Validation of Training Load Management

This subsection evaluates the second research objective of this thesis:

To ensure that the simulated training programs adhere to established sports science principles for load management, thereby creating a realistic environment where injury occurrence is predominantly influenced by recovery factors

This goal should ensure that a realistic environment is created that reflects competitive age-group triathletes often relying on coaches that design their training following scientific principles in order to minimize injury risk and maximize performance gains. The consequence of this careful design choice in the synthetic data generation should lead to injuries mainly originating from factors outside of training and hence outside of a coaches control.

To verify that this goal was met, three critical aspects are analyzed:

- **Periodization Structure**
- **Sport-specific Load Distribution**
- **Acute-to-Chronic Workload Ratios**

5.2.3.1 Periodization Analysis

The periodization analysis examines how training loads progress week after week, making sure that a controlled progression takes place to create the necessary stimuli for performance gains timed according to race proximity, while avoiding drastic increases that could impose risk.

The two plots 5.7 and 5.8 visualize these progressions. Figure 5.7 depicts the weekly planned load progression in a training plan generated through the data generation framework, it also visualizes the current training phase as well as race weeks. Figure 5.8 visualizes the weekly percentage change in planned TSS of the same training plan, with the dashed red line representing the 10% risk threshold [142].

The year starts with the base phase lasting approximately until week 13. This phase is characterized by training volumes progressively increasing from around 650 weekly TSS to 900 weekly TSS. Recovery weeks with decreased load are strategically placed approximately every third week to facilitate adaptation. The percentual load change approaches or slightly exceeds the 10% threshold in week 3 (10.23%) and in week 7 (10.18%), demonstrating a cautious approach to load progression during this foundational period.

This is followed by a build phase from weeks 12-22, characterized by sustained high training loads consistently around 900 TSS. The fact that the 900 TSS threshold is never surpassed likely indicates the corresponding athlete's weekly capacity limit, a common constraint for amateur

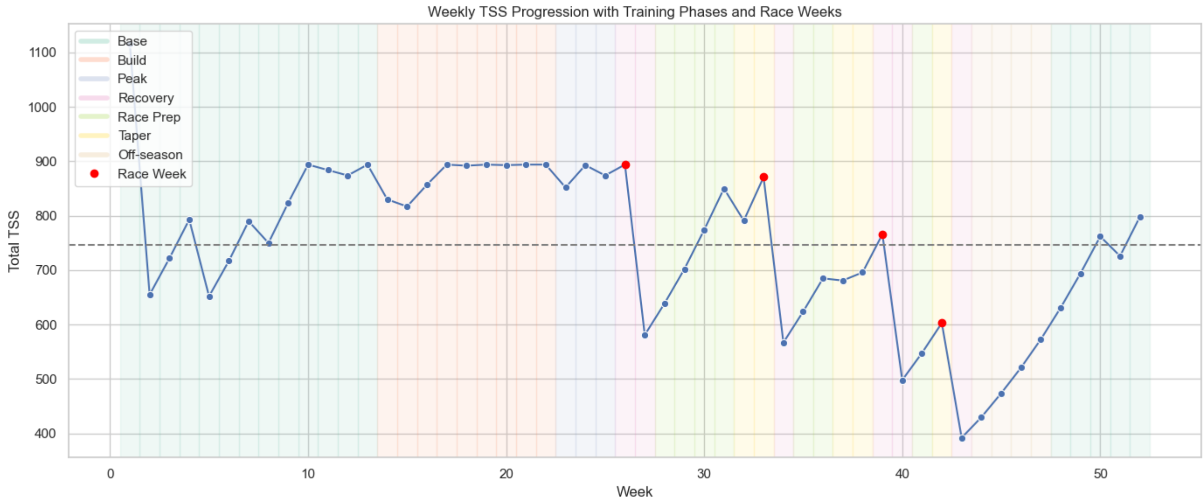

Figure 5.7: Weekly TSS Progression with Training Phases

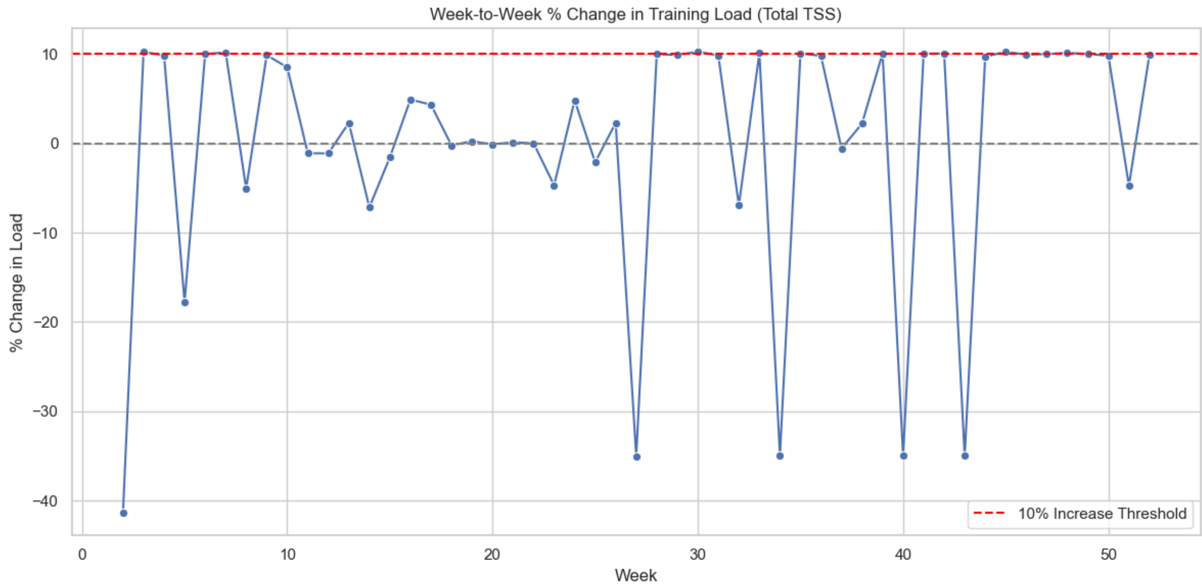

Figure 5.8: Week-to-Week % Change in Training Load

athletes balancing competitive ambitions with career, family, and other obligations. During this phase, the 10% increase threshold is never reached, suggesting a focus on load maintenance rather than progression once capacity is established.

Weeks 23 to 25 mark a peak phase, with a deliberate reduction in TSS in week 25 preceding a race in week 26. The race week is followed by a recovery phase lasting until week 27. In this plot, the most common phase during the week was taken for the color coding, suggesting the TSS drop in week 25 could be attributed to a portion of the week being a taper period in preparation for competition.

Three further races take place in relatively close succession, which impacts the traditional periodization structure that preceded the first race. Instead, a compressed cycle of race prep → taper → race → recovery is repeated, demonstrating how the periodization model adapts to accommodate multiple competitive events in sequence.

The highest training loads are observed during the build and peak phases, reaching approximately 900 TSS, while significant reductions occur during recovery and taper phases, sometimes dropping below 400 TSS. The 10% increase threshold is occasionally approached or marginally exceeded, particularly after recovery weeks. However, these increases are never excessive, with the maximum of 10.256% occurring in week 30 during a race prep phase. The significant drops in load visible in Figure 5.8 align precisely with recovery and taper periods, further reinforcing the structured nature of the training plan.

The analysis of this synthetic training plan demonstrates adherence to established sports science principles for load management. The controlled progression during base phases, strategic implementation of recovery weeks, maintenance of appropriate load ceilings, and race-specific tapering all reflect evidence-based periodization strategies. These features create a realistic training environment where weekly TSS progressions generally stay within or very close to recognized safety thresholds while still providing sufficient training stimulus.

5.2.3.2 Sport Distribution

Cross-training in triathlon allows for greater training volumes as discipline-specific muscles experience varied stress patterns. For instance, running-specific muscles are less affected during cycling, but the aerobic gains made during cycling, positively affect running performance as well [143]. The distribution among the three disciplines is driven by multiple factors, including individual strengths, race-day requirements, and injury risk considerations—particularly important being the fact that running is the most biomechanically impactful discipline of the three [144]. Figure 5.9 visualizes the distribution among bike, run, and swim training loads in the synthetic dataset.

Figure 5.9 visualizes the distribution among bike, run, and swim in the synthetic dataset.

Bike Cycling makes up the majority of the training load with a mean of 53.94%. The distribution is relatively tight, with most values falling between 50% and 58%. This aligns with typical triathlon training approaches that emphasize cycling volume due to its longer duration on race day and lower injury risk profile. The consistency in cycling allocation across the dataset reflects the foundational role it plays in triathlon training.

Run Running averages 25.12% of the total TSS and shows a wider spread, ranging from about 11% to almost 40%. This variability likely reflects differences in individual biomech-

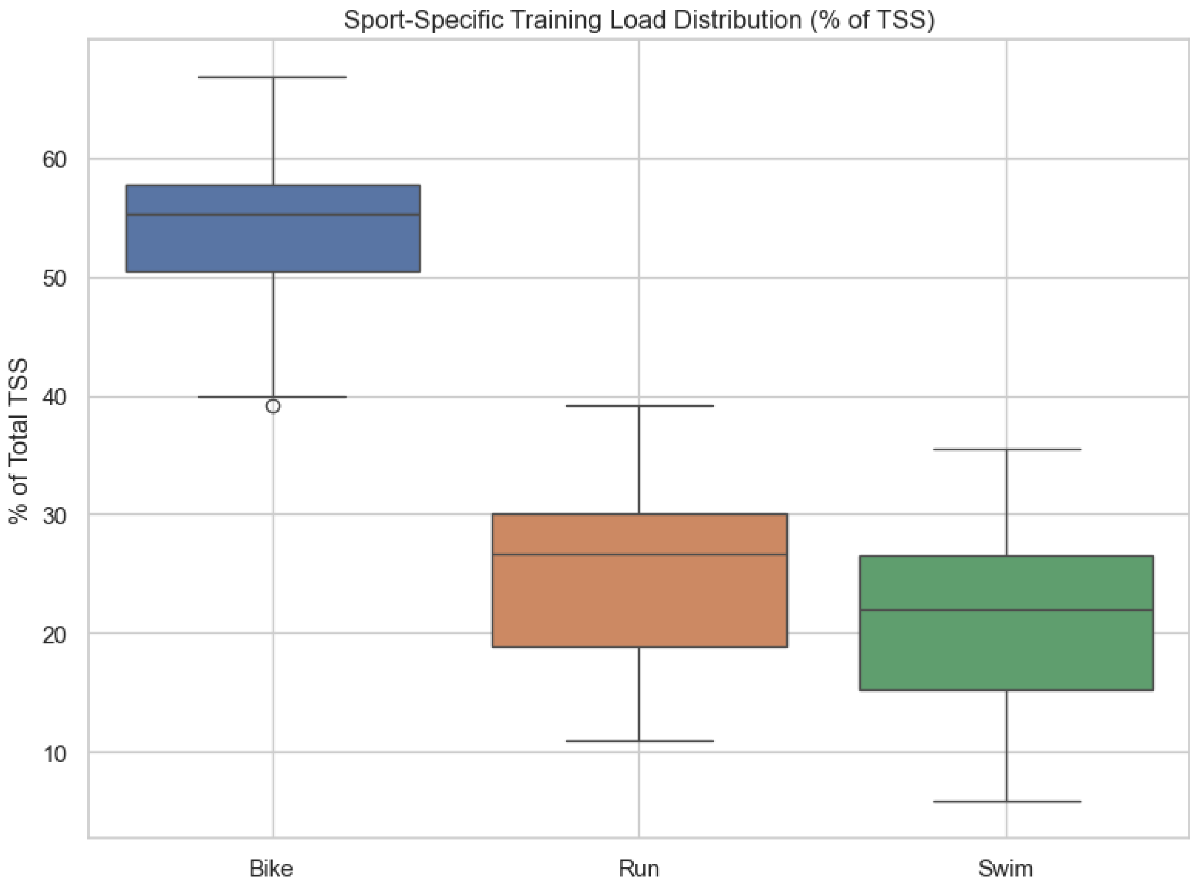

Figure 5.9: Sport-Specific Training Load Distribution

anical tolerance, running economy, and training background. Athletes with stronger swimming backgrounds may reallocate more focus to running at the expense of swimming volume. The moderation in running volume compared to cycling is consistent with evidence-based triathlon training that balances performance gains against injury risk [144].

Swim Swimming on average accounts for 20.95% of total TSS, displaying the broadest distribution from 5.83% to 35.5%. This substantial variation can be explained through differences in swimming backgrounds, technical proficiency, and access limitations to swimming facilities. The relatively lower overall proportion aligns with its shorter duration in most triathlon events.

The discipline distribution patterns observed in the synthetic dataset closely mirror recommended practices in triathlon training literature, which typically advocate for a cycling-dominant approach with carefully managed running volumes and technically-focused swimming [86, 87].

5.2.3.3 Adherence to Load Management Principles

Workload management is a critical aspect of injury prevention in endurance sports training. The Acute Workload Ratio (ACWR) provides valuable insights into how current training load (acute) compares to established fitness (chronic), helping to identify periods of potential injury risk [123].

Figure 5.10 displays the ACWR progression throughout the training plan already analyzed (5.2.3.1), while Figure 5.11 shows the distribution of ACWR values across all simulated athletes in the dataset.

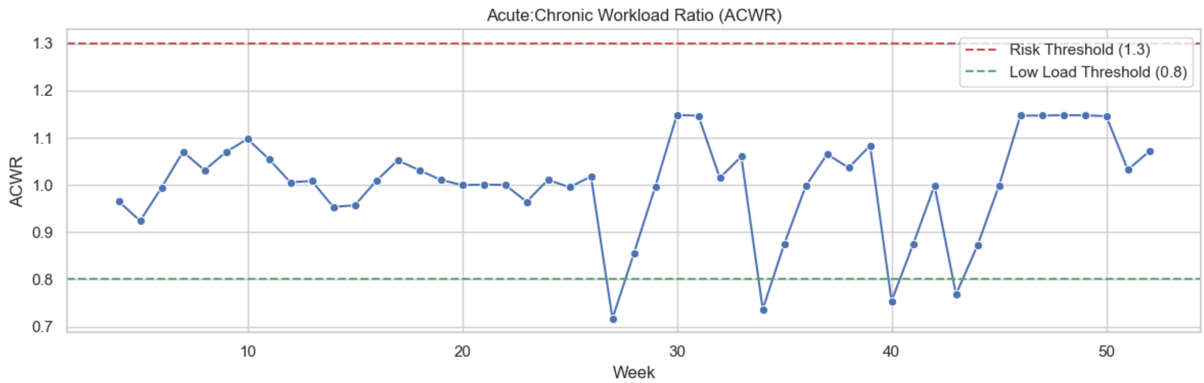

Figure 5.10: Acute Workload Ratio (ACWR) in a specific Training Plan

The weekly ACWR plot in Figure 5.10 demonstrates careful management of training loads throughout the season. The ratio generally fluctuates between 0.8 and 1.15, indicating controlled progressions that rarely exceed established risk thresholds [123]. The upper risk threshold of 1.3 (red dashed line) is never breached, while the lower threshold of 0.8 (green dashed line) is occasionally crossed during planned recovery and taper periods (weeks 27, 32, 39, and 44), which aligns with the race schedule observed in the periodization analysis 5.2.3.1.

These strategic dips below 0.8 represent intentional reductions in acute load to facilitate recovery, not random training errors. Notably, the ACWR returns to the optimal range of 0.8-1.3 [123] immediately following these recovery periods, demonstrating methodical load management. The highest recorded ACWR values (approximately 1.15) occur during build phases when training loads are intentionally increased to stimulate adaptation, yet remain well below the high-risk threshold of 1.3.

Figure 5.11 reveals that across all simulated athletes, ACWR values follow a normal distribution centered around 1.0, with the vast majority falling within the widely accepted ”safe zone” of 0.8 to 1.3 (indicated by the green vertical lines) [123]. The red vertical line at 1.5 represents the high-risk zone threshold [123], which contains only a negligible portion of the distribution. This demonstrates that the synthetic training plans predominantly operate within established

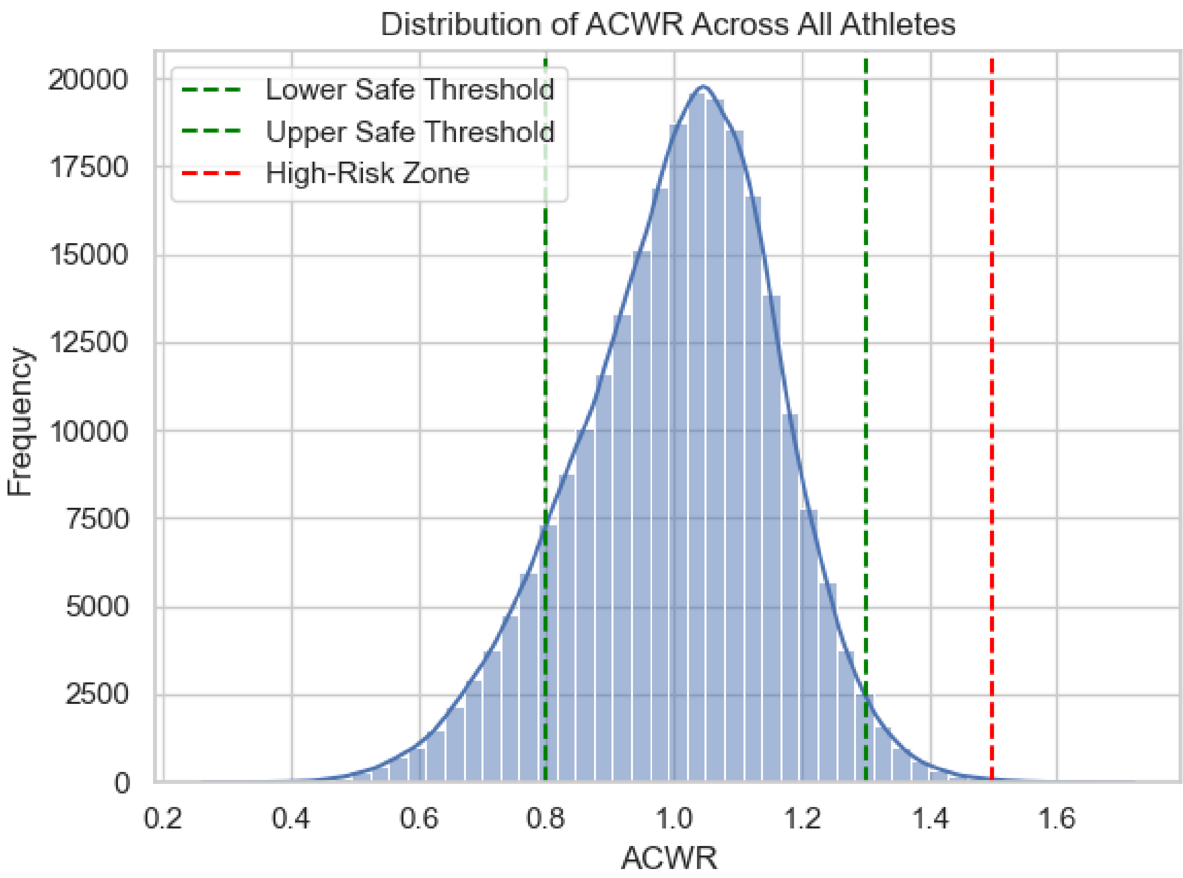

Figure 5.11: Distribution of ACWR Across All Athletes

safe parameters for workload progression.

The ACWR distribution validates that the simulated training plans successfully implement evidence-based workload management principles. By designing training that respects these established thresholds, the synthetic dataset creates a promising environment where excessive training loads are rarely the primary cause of injury.

The synthetic data also models the natural fluctuations in ACWR that would be expected in real-world training scenarios, avoiding an artificially perfect pattern that would misrepresent actual training conditions. This realistic variability, while remaining within safe parameters, creates an appropriate context for investigating how recovery metrics interact with workload management to influence injury risk in triathletes.

5.3 Machine Learning Experiments

This section evaluates the third research objective of this thesis:

To implement and evaluate multiple machine learning algorithms for injury prediction using *the synthetic dataset, with particular emphasis on identifying early warning signs that precede overuse injuries*

Building upon the exploratory data analysis presented earlier, which revealed distinct physiological patterns preceding injury events, this section examines how effectively machine learning models can leverage these patterns for predictive purposes. The experiments are designed to address two critical questions: first, whether automated approaches can reliably detect the subtle pre-injury warning signs identified in 5.2.2.1, and second, how the choice of data splitting strategy affects model performance in real-world application scenarios.

The section begins with a brief recap of the experimental setup introduced in Section 4.2, incorporating design decisions informed by the insights gained during EDA—particularly the identification of a critical 7-day window before injury onset as discussed in 5.2.2.2. Subsequently, the performance of the implemented algorithms is evaluated across both athlete-based and time-based splitting strategies. Finally, the analysis examines which features contribute most significantly to prediction accuracy, validating whether the models align with the physiological understanding established through exploratory analysis.

5.3.1 Experimental Setup

The experimental framework builds upon the methodology established in Section 4.2, incorporating specific refinements based on EDA findings. Three algorithms were selected for evaluation: XGBoost, Random Forest, and Lasso Regression. These models were chosen following findings of related work, and represent different approaches to injury prediction—gradient boosting, ensemble learning, and regularized linear modelling.

5.3.1.1 Prediction Window

Following the temporal analysis presented in Subsection 5.2.2.2, which identified clear physiological deterioration approximately 7 days before injury, a 7-day prediction window was adopted. This window represents the critical period where intervention could potentially prevent injury while maintaining practical relevance for real-world applications.

5.3.1.2 Data Splitting Strategies

Two splitting approaches were implemented to evaluate different aspects of model generalization:

- **Athlete-based split**: Uses 80% of athletes (800 athletes) to train the model and the remaining 20% (200 athletes) reserved for testing. This strategy assesses the model's ability to generalize to entirely new individuals, simulating real-world deployment scenarios where the model encounters athletes with no prior training data.

- **Time-based split**: Training data comprised all athlete records from January through October, while testing was conducted exclusively on November and December data across all athletes. By training on earlier data and predicting future events, this scenario simulates real-world deployment where models must predict future injuries based on historical data

5.3.1.3 Performance Metrics

Model performance was evaluated using:

- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: Measures the model's ability to distinguish between injured and non-injured instances across all classification thresholds. It provides a general overview of discriminative performance.

- **Precision-Recall Metrics**: Focus specifically on the minority class (injuries), providing a more sensitive assessment when positive events are rare. Precision (positive predictive value) indicates the proportion of predicted injuries that were true injuries, while recall (sensitivity) measures the proportion of actual injuries that were correctly identified. The area under the precision-recall curve (PR-AUC) summarizes this relationship across thresholds.

- **F1 Score**: Represents the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. Particularly valuable when missing an injury (false negative) and falsely predicting an injury (false positive) are both critical but differently weighted risks.

- **Feature Importance Analysis**: Evaluates which input features most strongly influence model predictions. This assessment is important for validating that model behavior is physiologically plausible (*e.g.,* captures the patterns identified during EDA), thus enhancing trustworthiness for deployment in athlete monitoring systems.

5.3.2 Model Performance Evaluation

5.3.2.1 Overall Performance Comparison

Table 5.1 presents the performance metrics for all three models across both data splitting strategies. The results demonstrate that all models achieve strong discriminative ability for injury prediction, with AUC scores ranging from 0.849 to 0.858 and average precision scores between 0.700 and 0.726.

Table 5.1: Model Performance Across Splitting Strategies

Model	Athlete-Based Split		Time-Based Split	
	AUC	Average Precision	AUC	Average Precision
LASSO	0.849	0.710	0.855	0.725
Random Forest	0.853	0.700	0.855	0.718
XGBoost	0.852	0.709	0.858	0.726

XGBoost emerges as the best performing model, when considering both splitting strategies together (AUC = 0.858, AP = 0.726). However, the performance differences are relatively small, with AUC values varying by less than 1% across both splitting strategies. This convergence of performance across fundamentally different algorithms—from linear (LASSO) to ensemble (Random Forest) to gradient boosting (XGBoost)—suggests that the synthetic dataset contains clear and consistent injury patterns that are readily detectable regardless of the modelling approach.

This observation has important methodological implications. It indicates that the predictive signal in the data is sufficiently strong and well-structured that even simpler models like LASSO can capture the essential relationships. This finding supports the quality of the synthetic data generation process.

The relatively high average precision scores (0.700-0.726) are particularly encouraging given the class imbalance inherent in injury prediction tasks. These results suggest that the models can effectively identify injury risks while maintaining reasonable false positive rates, a crucial requirement for practical deployment in athlete monitoring systems.

5.3.2.2 Athlete-Based Splitting Results

Figures 5.12, 5.13, and 5.14 display ROC and Precision-Recall curves for LASSO, random forest, and XGBoost respectively, under the athlete-based splitting strategy. This evaluation approach assesses the models' ability to generalize to entirely unseen athletes, mimicking a real-world deployment scenario where the system encounters individuals not present in the training data.

The results are remarkably consistent: Random Forest stands out slightly with an AUC of 0.853, while XGBoost follows at 0.852 and LASSO at 0.849. In terms of precision-recall performance, LASSO actually outperforms the other models with an average precision of 0.710, compared to XGBoost's 0.709 and Random Forest's 0.700.

What is particularly interesting about these ROC curves is their early steep rise, indicating that all models achieve high true positive rates at very low false positive rates. This suggests they are quite good at identifying the most obvious injury risks before they start making mistakes.

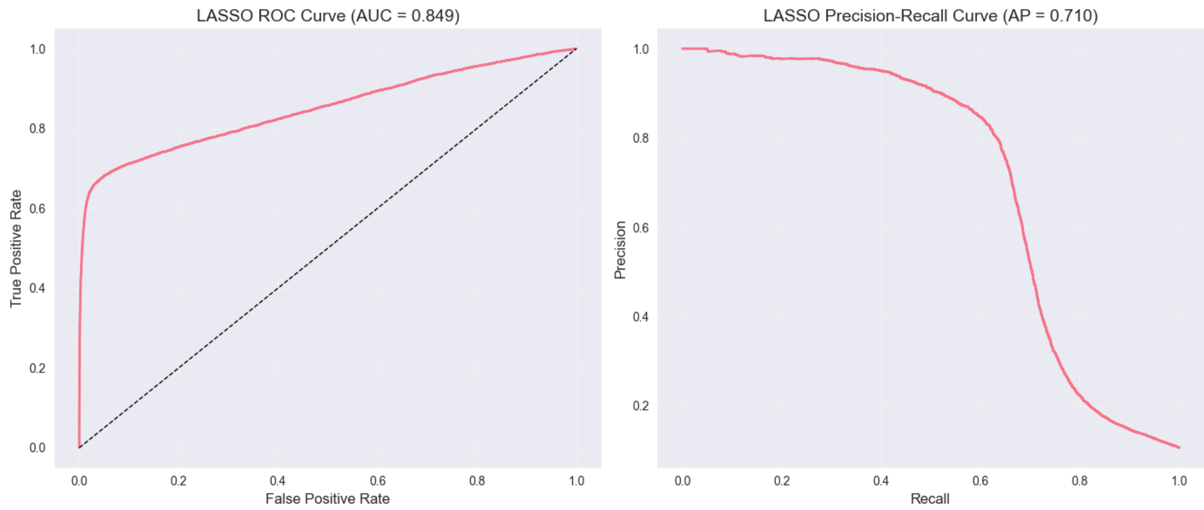

Figure 5.12: LASSO ROC and Precision-Recall Curves

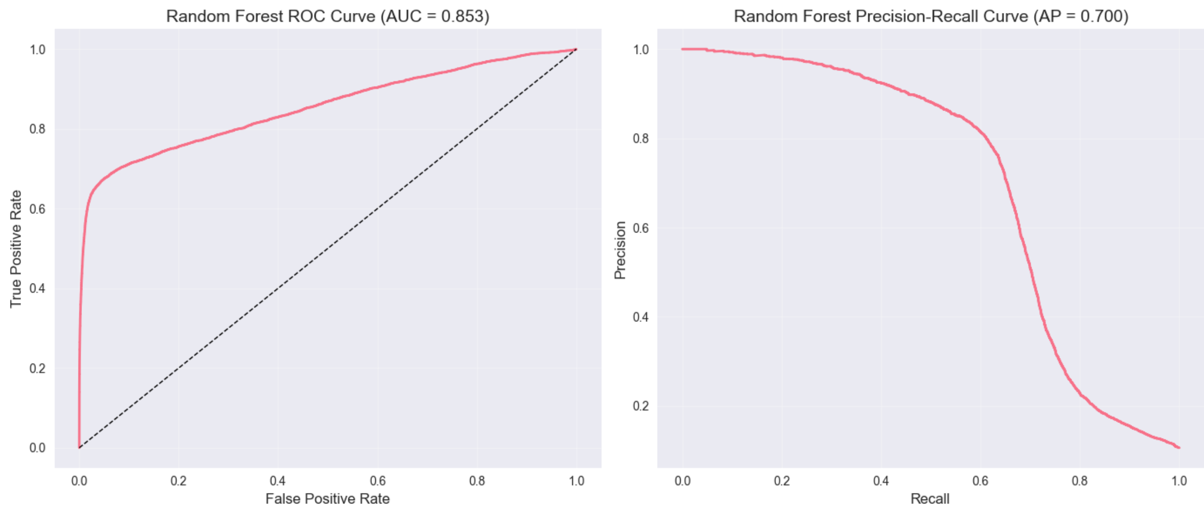

Figure 5.13: Random Forest ROC and Precision-Recall Curves

The precision-recall curves tell a complementary story: all models maintain extremely high precision (around 0.98-1.0) when being very selective, only losing precision as they try to catch more subtle cases (around 0.6 Recall).

5.3.2.3 Time-Based Splitting Results

The time-based splitting strategy mirrors how these systems would learn from the past to predict the future. Figures 5.15, 5.16, and 5.17 reveal that this approach consistently outperforms the athlete-based split. XGBoost performs best with an AUC of 0.858, while LASSO and Random Forest tie at 0.855. The precision-recall metrics show similar improvements, XGBoost achieves an average precision of 0.726, LASSO follows closely at 0.725, and Random Forest manages 0.718.

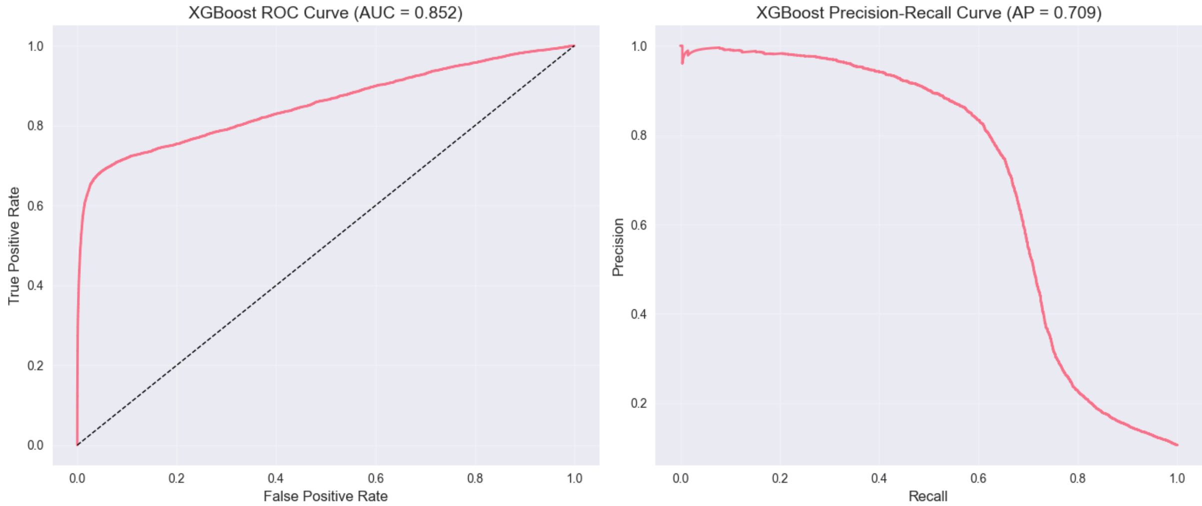

Figure 5.14: XGBoost ROC and Precision-Recall Curves

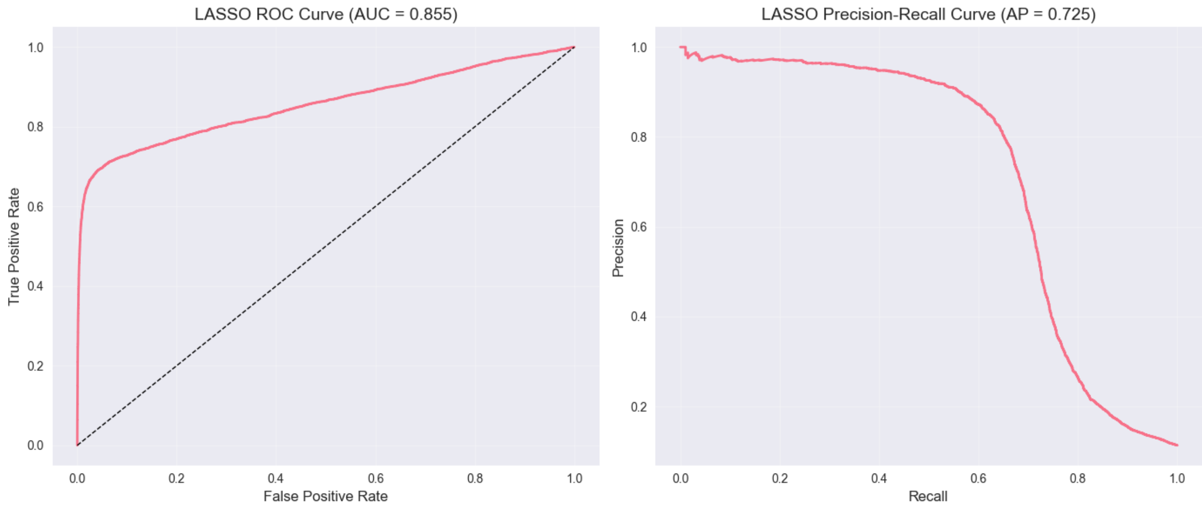

Figure 5.15: LASSO ROC and Precision-Recall Curves

These improvements suggest that individual athlete characteristics significantly influence injury patterns. The models perform better when they can learn from an athlete's own historical data rather than trying to generalize from other athletes' patterns. This reflects the EDA observations (subsection 5.2.2.3) that each athlete has unique physiological responses and injury risk factors that are best captured when the model has access to their personal history. This is particularly evident in the precision-recall curves, where all models show better performance in the middle recall ranges (0.4-0.7), indicating they are getting better at catching those moderately subtle injury patterns.

The stronger performance in time-based splitting is encouraging for practical deployment. It suggests that as more historical data is accumulated, predictions should become increasingly accurate. However, the still-strong performance in athlete-based splitting means the models are

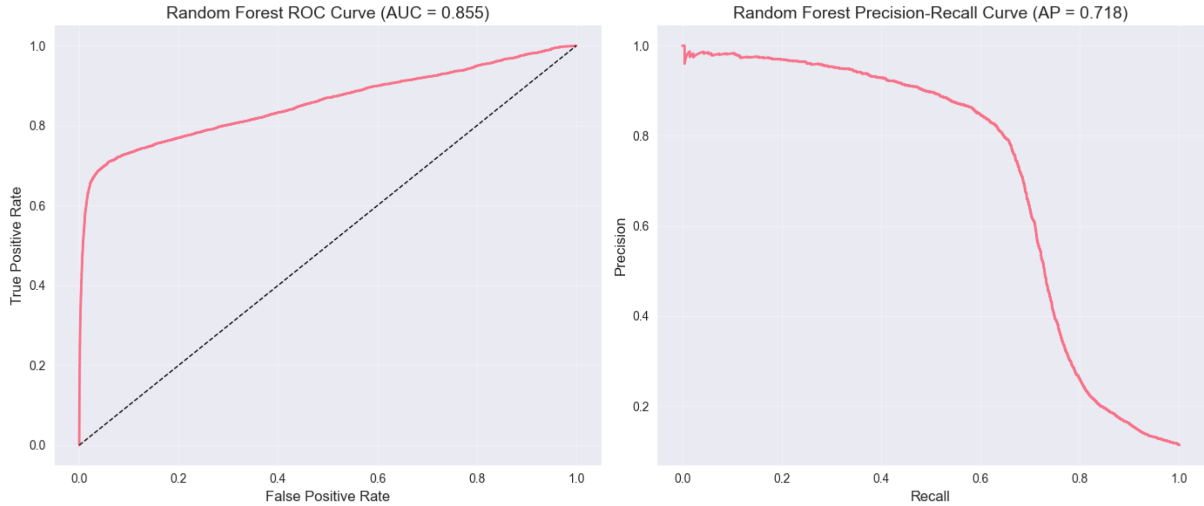

Figure 5.16: Random Forest ROC and Precision-Recall Curves

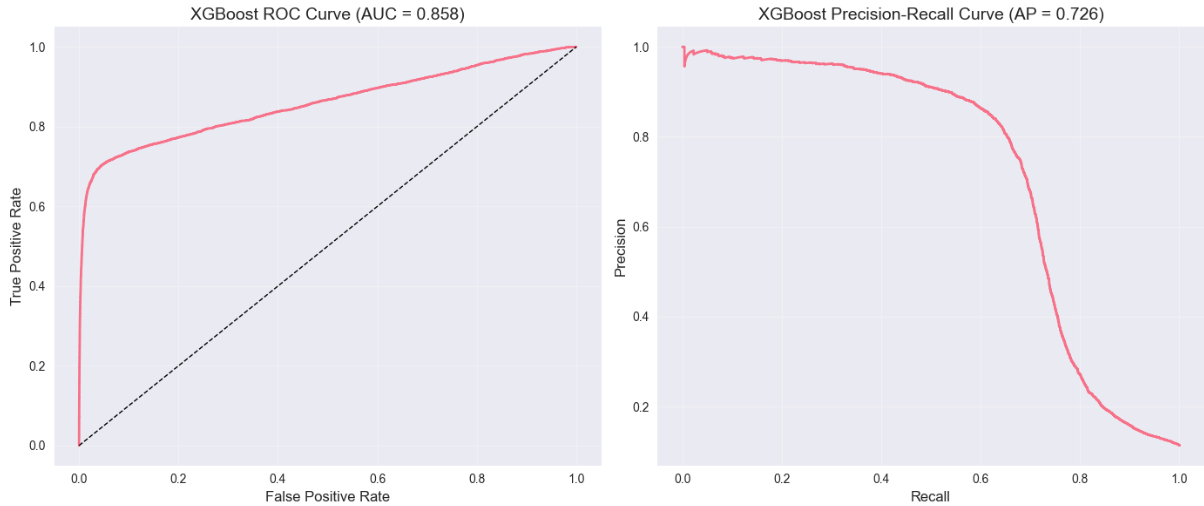

Figure 5.17: XGBoost ROC and Precision-Recall Curves

not just memorizing temporal trends, they are learning generalizable patterns that can be applied to new athletes as well.

5.3.2.4 Splitting Strategy Comparison

The comparative analysis between athlete-based and time-based splitting strategies reveals intriguing patterns about how these models learn to predict injuries. Figure 5.18 visualizes the performance differences across both strategies for all three models.

5.3.2.5 Performance Gap

The time-based split consistently outperforms the athlete-based split across all models, with AUC improvements ranging from 0.2% (Random Forest) to 0.6% (LASSO and XGBoost).

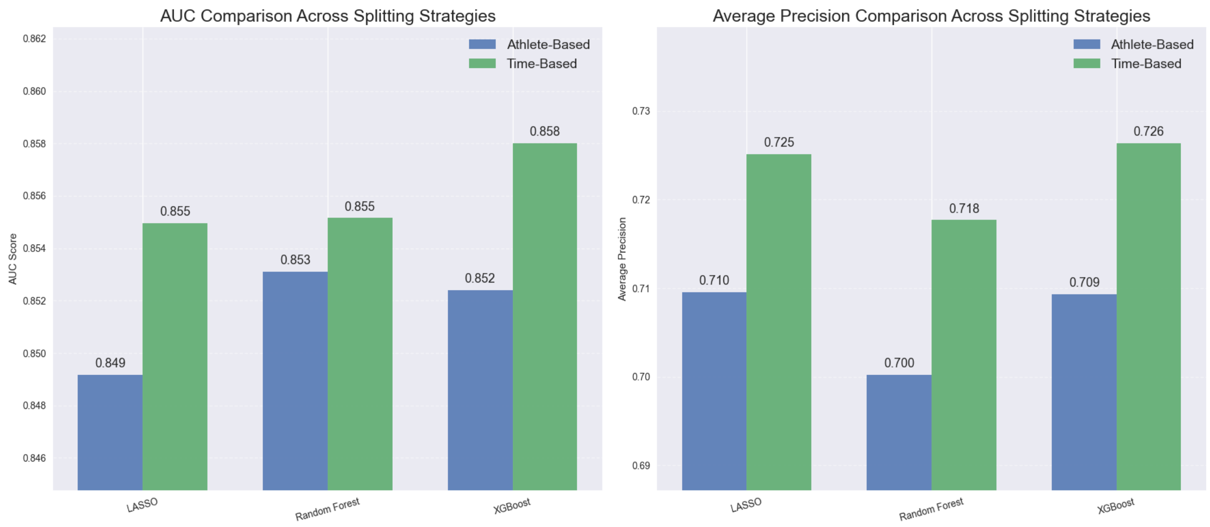

Figure 5.18: Performance Comparison Across Splitting Strategies

While these differences might seem modest, they reveal something fundamental about injury prediction: individual athlete differences play a significant role in injury patterns. When the model can learn from the same athletes' historical data (as in time-based splitting), it performs better than when it must generalize to completely new athletes.

The average precision metrics show even more pronounced improvements in the time-based approach: LASSO sees a 2.1% increase (from 0.710 to 0.725), Random Forest improves by 2.5% (from 0.700 to 0.718), and XGBoost gains 2.4% (from 0.709 to 0.726). These larger improvements in precision suggest that the time-based approach is particularly better at reducing false positives—a crucial consideration for practical deployment, where false alarms can lead to unnecessary interventions and increased athlete anxiety.

5.3.2.6 Deployment Strategy

The superior performance of time-based splitting suggests that in real-world applications, models should be continuously updated with recent data rather than attempting to build static, comprehensive athlete profiles. This observation leads to several practical implications:

- **Rolling deployment model:** Systems should be designed to regularly retrain on recent historical data, potentially using a sliding window approach where older data is gradually phased out as new patterns emerge.
- **Cold start handling:** For new athletes with limited historical data, the time-based model can still make reasonable predictions based on general temporal patterns, while the system progressively collects athlete-specific information.

5.3.3 Feature Importance Analysis

The comparative analysis of feature importance across LASSO, Random Forest, and XGBoost models reveals distinct yet complementary patterns in identifying key predictors of injury risk. Notably, each model captures different aspects of the physiological progression toward injury, reflecting their inherent algorithmic structures.

5.3.3.1 LASSO

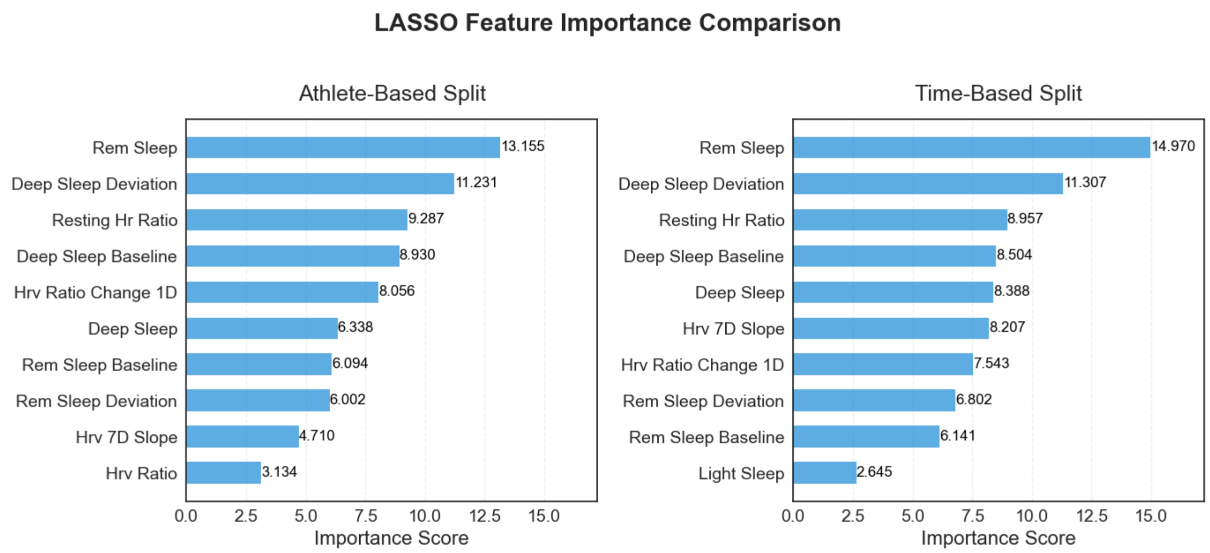

Figure 5.19: LASSO Top 10 Features across Time-Based and Athlete-Based Splits

The LASSO model's feature selection (Figure 5.19) aligns strongly with the correlation analysis presented in Figure 5.4. Sleep-related metrics, which demonstrated the strongest negative correlations with injury (deep sleep: $r = -0.19$, REM sleep: $r = -0.12$), emerge as dominant predictors. REM sleep exhibits the highest importance scores across both data splits (athlete-based: 13.155, time-based: 14.970), followed by deep sleep deviation (athlete-based: 11.231, time-based: 11.307).

This consistency validates the findings from Section 5.2.2.1 that recovery metrics, particularly sleep quality, are critical indicators of injury risk. The model's emphasis on baseline and deviation metrics for both REM and deep sleep suggests that LASSO effectively captures the dual importance of absolute values and individual variability. The prominence of resting heart rate ratio (athlete-based: 9.287, time-based: 8.957) further corresponds to the positive correlation identified between elevated resting heart rate and injury occurrence ($r \approx 0.06$).

The inclusion of HRV metrics, including HRV ratio change and 7-day slope, reflects the biphasic pattern observed in Figure 5.5, where HRV initially elevated before declining sharply in the days preceding injury. This suggests that LASSO successfully captures both immediate disruptions and longer-term trends in autonomic function.

5.3.3.2 Random Forest

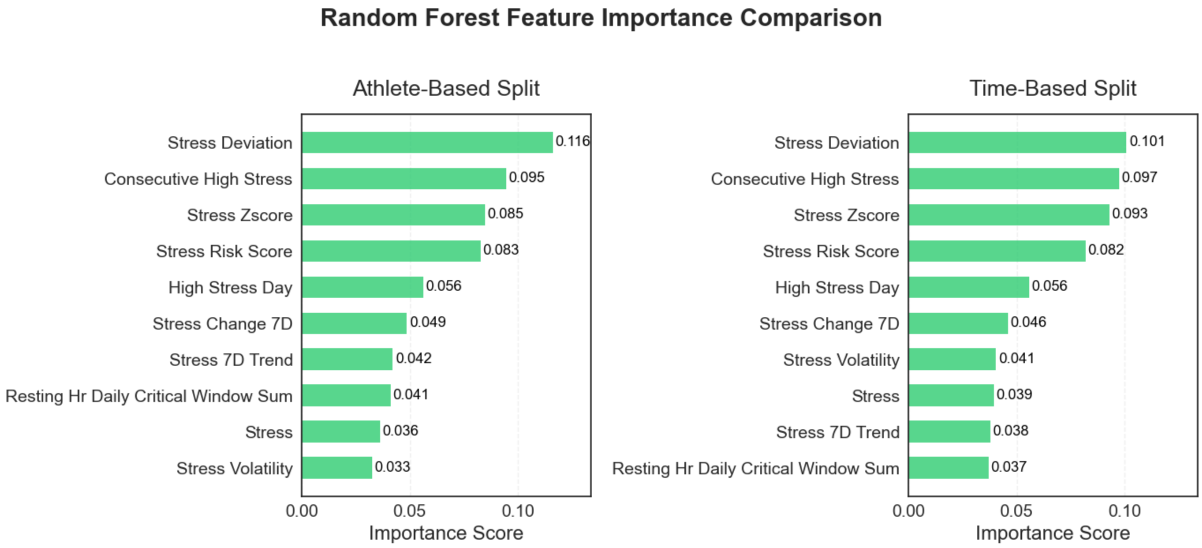

Figure 5.20: Random Forest Top 10 Features across Time-Based and Athlete-Based Splits

In strong contrast to LASSO, Random Forest (Figure 5.20) elevates stress-related features to primary importance, despite the correlation analysis showing only a moderate positive correlation between stress and injury ($r \approx 0.04$). Stress deviation leads in both splits (athlete-based: 0.116, time-based: 0.101), followed by consecutive high stress and stress z-score.

This apparent discrepancy highlights Random Forest's strength in capturing non-linear relationships and interaction effects not evident in the point-biserial correlation analysis. The model's emphasis on cumulative stress metrics aligns with the temporal pattern identified in Figure 5.5, where stress showed continuous escalation throughout the two-week pre-injury period, beginning as early as 14 days before injury onset.

The dominance of stress-related features across both splits suggests that Random Forest identified stress as a fundamental driver of injury risk, potentially acting as an upstream factor influencing other physiological metrics. This interpretation is supported by the clustering of multiple stress metrics (deviation, consecutive days, z-score, and risk score) among the top features, indicating that Random Forest captures both acute and chronic stress manifestations.

5.3.3.3 XGBoost

XGBoost's feature selection (Figure 5.21) demonstrates the most dynamic pattern, particularly in its treatment of acute stress indicators. The high stress day feature shows dramatically dif-

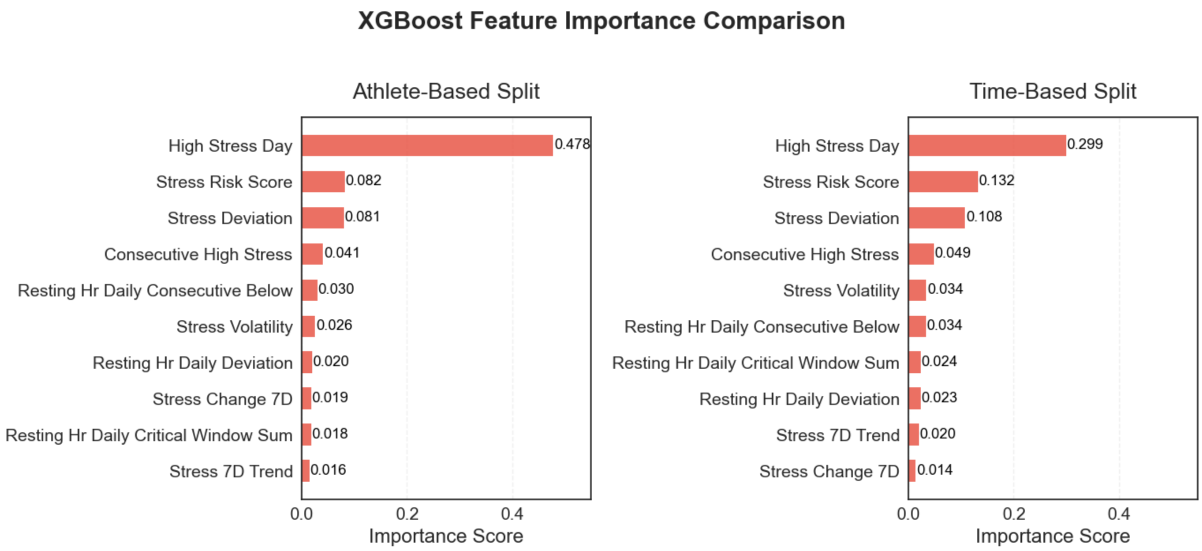

Figure 5.21: XGBoost Top 10 Features across Time-Based and Athlete-Based Splits

ferent importance between splits (athlete-based: 0.478, time-based: 0.299), suggesting that the manifestation of stress-related injury risk may be highly individualized.

This variability complements the findings from Section 5.2.2.3, where the analysis revealed substantial heterogeneity in pre-injury patterns across athletes. XGBoost's sensitivity to both acute and chronic stress indicators may reflect its ability to capture the critical transition period approximately 7 days before injury, as identified in Figure 5.5.

The model's feature prioritization suggests a hierarchical relationship: acute stress (high stress day) as the primary indicator, followed by composite scores (stress risk score) and deviations from baseline. This hierarchy may reflect XGBoost's ability to identify non-linear thresholds, where acute stress spikes become particularly dangerous when combined with underlying physiological vulnerability.

5.3.3.4 Cross-Model Insights

The divergent feature importance rankings across models provide complementary perspectives on injury risk factors:

1. **LASSO** emphasizes fundamental physiological states (sleep, HRV, resting heart rate) that correlate directly with injury risk, consistent with the linear relationships identified in the correlation analysis.

2. **Random Forest** identifies stress as a primary driver, capturing the continuous escalation pattern observed in the temporal analysis and potentially acting as a mediating factor for other physiological disruptions.

3. **XGBoost** focuses on acute manifestations and threshold effects, particularly in stress-related metrics, reflecting the individualized and non-linear nature of injury risk progression.

The consistency in feature selection between athlete-based and time-based splits for LASSO and Random Forest suggests robust generalization, while XGBoost's variability highlights the importance of considering individual differences in injury risk modelling. This aligns with the substantial individual variation observed in Section 5.2.2.3, where athletes demonstrated diverse physiological trajectories preceding injury.

A critical observation across all three models is the consistent absence of training load metrics (such as actual_TSS) among the top predictors. This finding provides compelling evidence that the second research objective—creating a realistic environment where injury occurrence is predominantly influenced by recovery factors rather than training load—was successfully achieved. The models' collective prioritization of recovery-related parameters (sleep quality, HRV, stress levels) over load metrics reflects the minimal correlation between actual_TSS and injury occurrence identified in Section 5.2.2.1, further validating the synthetic dataset's design principles.

5.4 Discussion

5.4.1 Achievement of Research Objectives

5.4.1.1 Physiological Plausibility of Synthetic Data

The first research objective aimed to develop a simulation framework capable of generating physiologically plausible athlete data with realistic pre-injury patterns. The extensive evaluation in Section 5.2 demonstrates that this objective was successfully achieved through several key findings:

The synthetic data exhibited physiological metrics that aligned closely with established clinical ranges for competitive age-group triathletes. Heart rate variability values (mean: 92.3ms, SD: 10.17ms) matched those reported in endurance athletes by Plews et al. [67], while resting heart rate distributions showed the expected bradycardic adaptation with 92.46% of values below 60 bpm. The trimodal sleep quality distribution accurately captured distinct recovery states, with proportions consistent with the defined athlete lifestyle profiles.

More significantly, the framework generated realistic interdependencies between physiological metrics. The correlation analysis (Section 5.2.1.2) revealed physiologically sound relationships such as the negative correlation between HRV and RHR (r = -0.23) and the strong negative correlation between sleep quality and stress (r = -0.72). These patterns demonstrate that the simulation avoided the common pitfall of treating physiological variables as independent parameters.

The dynamic response patterns to training loads (Figure 5.3) further validated the framework's biological plausibility. High-load training sessions produced expected acute physiological responses, including HRV suppression of approximately 3% and RHR elevation exceeding 1.5%, and were followed by appropriate recovery trajectories.

5.4.1.2 Sports Science Principles in Training Programs

The second research objective focused on ensuring simulated training programs adhered to established sports science principles, thereby creating an environment where injury occurrence would be predominantly influenced by recovery factors. This objective was comprehensively achieved as evidenced by multiple aspects of the evaluation:

The periodization analysis (Section 5.2.3.1) demonstrated systematic load progressions with appropriate base, build, peak, and recovery phases. Particularly noteworthy was the control of load increases, which rarely exceeded the 10% safety threshold established by Nielsen et al. [142]. The highest observed increase of 10.256% during a race preparation phase showed that the framework maintained realistic load progressions while avoiding excessive jumps.

Sport-specific load distribution achieved typical triathlon training allocations, with cycling comprising 53.94% of training stress, running 25.12%, and swimming 20.95%. This distribution reflects evidence-based practices that prioritize cycling volume while managing running impact to reduce injury risk, as documented by Hreljac [144].

The adherence to load management principles was definitively shown through the ACWR analysis. The distribution of ACWR values (Figure 5.11) centered around 1.0, with the vast majority falling within Gabbett's safe zone of 0.8-1.3 [123]. This careful load management created the intended effect: training load showed minimal correlation with injury occurrence ($r \approx 0.01$), confirming that injuries in the synthetic dataset primarily emerged from recovery-related factors rather than training errors.

5.4.1.3 Effectiveness of Machine Learning Approach

The third research objective targeted the implementation and evaluation of machine learning algorithms for injury prediction, with emphasis on identifying early warning signs. This objective was successfully met through comprehensive experimental validation:

All three implemented algorithms (XGBoost, Random Forest, and LASSO) achieved strong discriminative performance with AUC scores ranging from 0.849 to 0.858 and average precision scores between 0.700 and 0.726. The consistency across fundamentally different algorithmic approaches suggests that the synthetic dataset contains clear and detectable injury patterns.

The models successfully identified the critical pre-injury warning signs discovered during exploratory analysis. Feature importance analysis demonstrated that recovery metrics consistently ranked as top predictors across all models, aligning with the correlation analysis findings from Section 5.2.2.1. LASSO emphasized sleep-related metrics, Random Forest prioritized stress patterns, and XGBoost captured acute stress manifestations – collectively providing a comprehensive view of injury risk factors.

The superior performance of time-based splitting (AUC improvement of 0.2-0.6%) over athlete-based splitting revealed valuable insights about deployment strategies. This finding suggests that continuous model updating with recent data would be more effective than static athlete profiling, providing practical guidance for real-world implementation.

5.4.1.4 Synthetic Data as a Methodological Bridge

The fourth research objective was to demonstrate how synthetic data generation could overcome data privacy regulations and missing labels. The previous evaluation provides evidence of the potential of this approach, but with important considerations:

The framework successfully generated a dataset of 1,000 athletes, each with 365 days of data, creating 365,000 records that would be extremely difficult to collect in reality due to privacy constraints and compliance costs. The synthetic nature of the data eliminates privacy concerns while maintaining the statistical properties required for meaningful analysis.

The approach overcame the challenge of missing labels by generating complete ground truth labels for injury events, including the precise timing and physiological context. This complete labelling enabled both detailed exploratory analysis and supervised learning approaches, which would be difficult with real data due to incomplete injury reporting and subjective labelling issues.

It must be emphasized that the models developed in this thesis have not been evaluated with real athlete data, so their real-world performance remains theoretical. However, the findings provide promising indications of how such models might perform in practical applications. It is particularly encouraging that the time-based splitting consistently outperformed athlete-based splitting across all algorithms. This suggests a natural deployment pathway where models can be initially developed using synthetic data and then refined through continuous retraining as real historical data becomes available from onboarding users.

This incremental adaptation approach fits well with privacy-conscious deployment strategies, as it minimizes the initial real-world data requirements while building a foundation for incremental improvement. Each new user's data contributes to model refinement without requiring a large up-front dataset that could raise privacy concerns. The synthetic data thus serves not only as a substitute but also as a scaffolding that supports a gradual transition to real-world implementation, potentially accelerating the development of effective injury prediction systems while maintaining strict privacy standards.

5.4.2 Practical Implications

5.4.2.1 Implementation Considerations

The results of this thesis offer several important considerations for the implementation of injury prediction systems in real-world athlete monitoring contexts. First, the superior performance of the time-based splitting approach suggests that implementation architectures should be designed with temporal adaptation as a core feature. Systems should incorporate automated retraining pipelines that regularly update models with recent historical data, possibly using a sliding window approach where older data is gradually phased out.

Integration with existing wearable technology ecosystems presents both opportunities and challenges. Unlike the specialised laboratory equipment used in the studies by Petrovsky et al. [42], the focus on metrics commonly captured by commercial devices (HRV, RHR, sleep quality) makes widespread implementation feasible. However, data standardisation across different wearable platforms remains a challenge. Implementation solutions should include preprocessing modules that harmonise data from different sources.

Privacy considerations need to be carefully addressed, particularly given the sensitivity of physiological data. The synthetic data approach presented in this thesis provides a pathway for initial model development without privacy concerns, but real-world implementations would need to establish robust data governance protocols. This could include federated learning approaches where models are updated locally on athletes' devices before only anonymised parameter updates are shared, preserving individual privacy while enabling collective improvement.

5.4.2.2 Intervention Design

The injury prediction models developed in this thesis make proactive intervention strategies possible, that could significantly reduce injury occurrence in triathletes. Following the temporal patterns identified in Section 5.2.2.2, interventions should be designed to trigger approximately 7 days before potential injury onset.

This prediction system is particularly well-suited for integration with coaching platforms where athletes' training plans are already stored and managed. When an athlete's incoming physiological data begins to match the pre-injury patterns identified by the ML models, the system could automatically trigger warnings with specific, actionable recommendations. These interventions would work directly within the existing training calendar, rather than requiring separate monitoring tools or manual adjustments.

For example, if declining sleep quality and elevated stress levels are detected, similarly to patterns preceding historical injuries, the system could suggest modifications to upcoming scheduled sessions. These might include:

- Dynamic intensity adjustments: Automatically reducing the intensity targets of upcoming high-intensity workouts while preserving the overall structure of the session
- Strategic session swapping: Recommending the rescheduling of high-impact sessions (like threshold runs) with lower-impact alternatives (such as technique-focused swims)
- Recovery insertion: Proactively converting planned training days to active recovery sessions when risk exceeds certain thresholds
- Sleep prioritization: Suggesting earlier completion of evening workouts to protect sleep duration when sleep metrics show concerning trends

5.4.2.3 Pathway to Real-World Deployment

As mentioned, one of the objectives and the findings made in this thesis are designed to be a methodological strategy, to overcome data limitations, and are hence designed to integrate real-world data as it becomes available after deployment.

This transition requires a strategic, phased approach. The initial deployment should leverage the models trained on synthetic data as a foundation, with clear communication to users about the preliminary nature of the predictions. This transparency builds trust and sets appropriate expectations regarding model accuracy. The refinement pathway should follow a staged progression:

1. **Data Collection Phase:** Upon deployment, the system should prioritize high-quality data collection across the key physiological metrics identified in the feature importance analysis. This should include explicit consent processes and transparent data governance to maintain user trust.
2. **Calibration Phase:** As individual user data accumulates, personalised baselines should be established for each metric, addressing the individual variation observed in Section 5.2.2.3. This calibration phase might last 4-6 weeks, providing sufficient data to capture normal physiological fluctuations.
3. **Model Adaptation Phase:** Following the time-based splitting approach that demon-

strated superior performance, the models should be progressively retrained as real user data becomes available. Initially, this might involve model fine-tuning rather than complete retraining, gradually shifting from synthetic to real data influence.

4. **Feedback Integration:** User feedback on prediction accuracy and intervention effectiveness should be systematically collected and incorporated into model refinement. This closed-loop approach allows for continuous improvement while building a valuable labelled dataset of real injury occurrences.

5.4.3 Comparison with Existing Literature

This thesis extends and complements existing literature on injury prediction in several important ways. The synthetic data approach provides a novel solution to the data scarcity challenge highlighted by Hohl et al. [9] and Lange et al. [23]. While Rossi et al. [7] and Naglah et al. [44] relied on limited datasets from professional football teams, and Lövdal et al. [8] used data from elite runners, the synthetic framework enabled the creation of a comprehensive dataset representing diverse athlete profiles and training scenarios.

The emphasis on recovery metrics as primary predictors is in line with emerging research trends, but differs from traditional approaches that focus predominantly on training load. The finding that sleep quality, heart rate variability, and stress levels were more predictive than training stress scores validates Halson's [41] framework, which emphasized the importance of internal load monitoring over external metrics. Similarly, the critical role of sleep quality identified is consistent with Ayala et al.'s [45] finding that sleep quality was a significant predictor of hamstring injuries in football players.

The temporal analysis of pre-injury patterns extends the methodology of several previous studies. The identification of a critical transition point approximately 7 days before injury provides greater specificity than the broad "within one week" window used by Naglah et al. [44]. In addition, the detailed progression patterns for different physiological metrics provide more nuanced insights than the day/week comparisons presented by Lövdal et al. [8]. This approach builds on Rothschild et al.'s [43] Markov unfolding technique by explicitly modelling the trajectory of physiological deterioration.

The comparative analysis of model performance makes a methodological contribution to the field of sports analytics. Unlike Rothschild et al. [43], who found that LASSO models outperformed more complex algorithms, this thesis demonstrated comparable performance between LASSO, Random Forest, and XGBoost. This suggests that the structural patterns in injury development can be captured by diverse modelling approaches, providing flexibility for real-world implementation based on computational constraints or interpretability requirements.

Finally, this thesis addresses the specific gap identified in literature regarding triathlon-specific injury prediction models. While Kienstra et al. [40] explored injury mechanisms in triathletes theoretically, and Chen [46] proposed advanced clustering methods for athlete risk profiling, neither developed practical, deployable models for early warning detection. By focusing specifically on the multidisciplinary nature of triathlon training and the unique recovery challenges it presents, this thesis provides a tailored solution for a sporting population that has received limited attention in predictive analytics literature.

5.5 Limitations of Synthetic Data Approach

5.5.1 Physiological Simplifications

Despite the demonstrated physiological plausibility of the synthetic dataset, several important biological mechanisms were necessarily simplified or omitted from the simulation framework. These simplifications represent potential limitations in the model's ability to fully capture the complexity of real athlete physiology.

Biomechanical factors that play a critical role in many triathlon injuries were not explicitly modelled in the simulation. While training loads were distributed across swimming, cycling and running to reflect different impact profiles, the specific biomechanical stresses associated with each discipline - such as running gait abnormalities, cycling fit issues or swimming technique errors - were not included. These factors can significantly influence injury susceptibility in ways that may not be fully captured by the recovery metrics prioritised in current models.

The menstrual cycle and its effects on female athletes were not included in the physiological simulation. Research by Bruinvels et al. [145] has shown that hormonal fluctuations during the menstrual cycle can significantly affect recovery capacity, injury risk and performance in female athletes. The omission of these cyclical patterns may limit the applicability of the findings to female triathletes, who make up a significant proportion of the sport's participants.

The timing of training in relation to circadian rhythms was not considered in the model. The simulation treated training sessions as discrete stress events, without accounting for potential differences in physiological response based on time of day. Research suggests that training at different times may influence hormonal responses, recovery dynamics and adaptation patterns [146]. This simplification may obscure important temporal relationships between training schedule and injury risk that would be present in real athlete data.

Although the model included sleep duration and stage distribution, it did not include sleep phase timing, sleep onset latency, or the relationship between sleep timing and circadian alignment. These factors may significantly influence the quality of recovery beyond what is captured by duration and stage percentages alone.

Nutritional factors that significantly influence recovery and adaptation [147] were modelled indirectly through the athletes' lifestyle profiles rather than as explicit variables. Elements such as energy availability, macronutrient timing and micronutrient status can significantly alter an athlete's ability to recover from training stress. The absence of these nutritional parameters may lead to an incomplete representation of the recovery process, potentially missing important interactions between nutritional status and injury susceptibility.

5.5.2 External Validity

A fundamental limitation of this thesis is the lack of validation against real athlete data. Extensive efforts have been made to ensure physiological plausibility through literature-based parameter ranges and relationships, and in some cases assumptions, but the resulting patterns remain theoretical constructs until validated with real athlete populations.

The synthetic population may not fully represent the diversity of real triathlete populations. Although the framework accounted for variations in training background, lifestyle factors and physiological parameters, these were based on literature-derived distributions rather than empirical sampling of actual triathlete cohorts. This approximation may miss important sub-populations or extreme cases that exist in the real world but were not accounted for in the simulation parameters.

The injury models generated are based on clear physiological deterioration patterns, which may be more pronounced in the synthetic data than in noisy real-world measurements. Wearable devices in real-world use scenarios often suffer from missing data, measurement errors and inconsistent user compliance, which could obscure the predictive patterns identified in this research. The clean, continuous data streams generated by the simulation framework may therefore overestimate the potential predictive performance achievable in practice.

The timeframes for physiological responses were derived from the literature, but may not accurately reflect real-world progression patterns. The critical 7-day window identified prior to injury onset requires empirical validation, as actual injury development may follow shorter or longer trajectories depending on individual factors not fully captured in the simulation.

5.5.3 Model Transfer Challenges

The application of models trained on synthetic data to real-world athlete monitoring presents several significant challenges that must be addressed for successful deployment. These transfer challenges represent important limitations that could affect initial model performance in real-world applications.

The domain shift between synthetic and real data distributions is perhaps the most challenging. Despite efforts to ensure physiological plausibility, systematic differences between simulated data patterns and actual wearable device measurements could lead to degraded model performance when applied to real athletes. These differences could be due to device-specific measurement characteristics, environmental factors not accounted for in the simulation, or subtle physiological relationships that have been oversimplified.

Calibration requirements for individual athletes may be greater than expected. The synthetic data approach assumed that individual variability could be addressed through baseline monitoring and normalisation, but real-world athletes may exhibit more complex patterns that require longer adaptation periods or more sophisticated personalisation approaches than current models account for.

Feature engineering differences are likely to emerge between synthetic and real data sets. The clean, regular structure of the synthetic data enabled straightforward computation of features such as rolling averages, z-scores and deviation metrics. Real wearable data often contain irregularities, missing values and measurement artefacts that require more robust pre-processing pipelines, potentially altering the information content of the derived features.

Temporal consistency may differ significantly between synthetic and real-world measurements. The simulation framework assumed consistent daily measurements of all physiological parameters, whereas real-world athlete monitoring typically suffers from variable compliance, device charging issues and intermittent use patterns. These inconsistencies could undermine the effectiveness of the time-series features that proved most valuable in the synthetic analysis.

Inter-device variability is another significant transfer challenge. The simulation created idealised physiological metrics without taking into account measurement differences between different wearable device manufacturers. In practice, HRV or sleep staging algorithms may differ significantly between brands, potentially requiring device-specific model adaptations to maintain predictive performance.

These transfer challenges underscore the importance of the phased deployment strategy outlined in Section 5.4.2.3, which acknowledges the preliminary nature of synthetic-based models and establishes a structured pathway for adaptation as real-world data become available. While these limitations are significant, they do not diminish the value of the synthetic approach as a methodological bridge to overcome initial data limitations in injury prediction research.

6 Conclusion

This concluding chapter summarizes the main findings and contributions of this thesis, reflects on key considerations that emerged during the research process, and outlines promising directions for future work.

6.1 Summary

This thesis introduces a novel synthetic data generation framework to address data limitations imposed by privacy restrictions common in health and sports science applications. The framework makes a significant methodological contribution by enabling the use of predictive models that are otherwise constrained by the scarcity of large-scale, labelled datasets. The key contributions are as follows:

First, a simulation framework was developed to generate physiologically plausible athlete data, including injury labels preceded by realistic warning patterns. The synthetic dataset successfully reproduced key physiological relationships and temporal progression patterns consistent with established sports science literature on injury development.

Second, the simulation framework generated annual training data based on synthetic training schedules that were designed according to established principles of load management and periodization. This approach effectively simulated the goal environment in which injuries are primarily due to recovery factors rather than training errors.

Third, multiple machine learning models (LASSO, Random Forest, and XGBoost) were evaluated using different data partitioning strategies. All models showed comparable success in identifying pre-injury patterns, with time-based partitioning consistently outperforming athlete-based approaches. This finding provides valuable insight into real-world deployment strategies.

Finally, a pathway for transitioning from synthetic to real-world data applications was proposed. The framework was specifically designed to progressively incorporate incoming athlete data, allowing models to evolve from synthetic foundations to increasingly personalised and empirically validated systems as real-world data becomes available.

6.2 Considerations

This thesis journey provided significant insights that have influenced both the research methodology and understanding of data science challenges in healthcare and sports domains.

The research process began with a seemingly straightforward plan: partner with an established triathlon coaching company to gain access to athlete data. When this opportunity became unfeasible due to privacy regulations, despite having advanced to a stage where collaboration appeared certain, the project initially faced considerable uncertainty. However, this setback ultimately yielded valuable methodological insights.

Through discussions with the company, it became evident that they confronted challenges beyond privacy concerns—they lacked reliable injury labels in their dataset, a critical component for supervised learning approaches. This revelation necessitated a broader examination of the prevalence of this problem in sports and health sciences. This investigation revealed a common constraint: the most valuable data for prediction tasks is frequently difficult to access due to privacy regulations, ethical considerations, and labelling challenges.

Being required to reconsider the research approach led to the exploration of synthetic data generation as an alternative methodology. What began as a contingency plan evolved into the central methodological contribution of this thesis, demonstrating the importance of adaptability in research design.

Working with synthetic data enhanced appreciation for domain expertise in data science applications. Creating physiologically plausible simulations required integrating knowledge from multiple disciplines, including exercise physiology, sports science, and triathlon coaching. This interdisciplinary nature emphasized that effective data scientists in specialized domains must develop sufficient expertise to recognize when models align with, or diverge from, established knowledge frameworks.

Perhaps the most valuable observation was the recognition that idealized datasets rarely exist in practice. Applied contexts often require working with imperfect, incomplete, or proxy data. Developing methods to address these limitations, whether through synthetic data generation, transfer learning, or other approaches, represents an important direction in applied machine learning, particularly in domains with stringent privacy considerations.

This research journey transformed what initially appeared as a methodological constraint into an opportunity for innovation, potentially yielding a contribution with greater impact than the original research design. This experience reinforced the value of persistence and methodological flexibility when confronting data access challenges that are increasingly prevalent in contemporary research environments.

6.3 Future Work

This thesis opens several promising avenues for future research.

First, the data generation framework could be further refined by addressing the limitations outlined in Section 5.5, such as the incorporation of menstrual cycle effects in female athletes and the integration of circadian rhythm influences. Including these additional physiological factors could enhance the realism and diversity of the synthetic datasets.

Second, given access to a limited amount of real-world data, an important direction would be to explore adversarial approaches for refining the synthetic data generation. For example, generative adversarial networks (GANs) could be used, where a discriminator network is trained to discriminate between real and synthetic data, thus challenging and improving the ability of the generator to produce physiologically realistic patterns.

Finally, the ultimate research avenue is to investigate deployment strategies and develop continuous integration (CI) and continuous deployment (CD) pipelines for incorporating incoming real-world data. Such systems would allow for progressive validation and adaptation of models, allowing them to evolve from synthetic training to empirically based, personalised applications. Successful operationalisation of this process would demonstrate the practicality of overcoming initial data limitations through synthetic-first modelling approaches.

Beyond sports science, the synthetic data methodology proposed in this thesis could be applied to other health-related domains where data scarcity due to privacy concerns is a critical barrier. Areas such as rare disease research, mental health monitoring, and early intervention programs could benefit from domain-knowledge-driven synthetic data generation to bootstrap predictive models, overcoming the initial ”chicken-and-egg” challenge of requiring large datasets for system development.

In summary, while this thesis has established a foundation for synthetic data-based injury prediction in triathletes, it represents just the beginning of a promising research direction. The methodological approach developed here offers a template for addressing similar challenges across sports science and healthcare domains where privacy concerns and labelling difficulties have traditionally hindered the development of predictive systems.

Bibliography

[1] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ser. ACL '01. USA: Association for Computational Linguistics, 2001, p. 26–33. [Online]. Available: https://doi.org/10.3115/1073012.1073017

[2] GeeksforGeeks, "Data cleansing - introduction," https://www.geeksforgeeks.org/data-cleansing-introduction/, 2025, accessed: 2025-02-03.

[3] B. McShane, "Data preprocessing methods to improve model performance: Discretization transforms," 2024. [Online]. Available: https://medium.com/@broganmcshane.code/data-preprocessing-methods-to-improve-model-performance-discretization-transforms-ab0d9154d2ed

[4] C. Taylor, "Data normalization," 2020, accessed: 2025-02-04. [Online]. Available: https://cyberhoot.com/cybrary/data-normalization/

[5] C. A. Andersen, B. Clarsen, T. V. Johansen, and L. Engebretsen, "High prevalence of overuse injury among iron-distance triathletes," *British Journal of Sports Medicine*, vol. 47, no. 12, pp. 857–861, 2013.

[6] T. Soligard, M. Schwellnus, J.-M. Alonso, R. Bahr, B. Clarsen, H. P. Dijkstra, T. Gabbett, M. Gleeson, M. Hägglund, M. R. Hutchinson *et al.*, "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*, vol. 50, no. 17, pp. 1030–1041, 2016.

[7] A. Rossi, L. Pappalardo, P. Cintia, F. M. Iaia, J. Fernández, and D. Medina, "Effective injury forecasting in soccer with gps training data and machine learning," *PloS one*, vol. 13, no. 7, p. e0201264, 2018.

[8] S. S. Lövdal, R. J. Den Hartigh, and G. Azzopardi, "Injury prediction in competitive runners with machine learning," *International journal of sports physiology and performance*, vol. 16, no. 10, pp. 1522–1531, 2021.

[9] B. Hohl, H. F. Satizábal, and A. Perez-Uribe, “Unveiling the potential of synthetic data in sports science: A comparative study of generative methods,” in *International Conference on Artificial Neural Networks*. Springer, 2024, pp. 162–175.

[10] G. M. Migliaccio, J. Padulo, and L. Russo, “The impact of wearable technologies on marginal gains in sports performance: An integrative overview on advances in sports, exercise, and health,” *Applied Sciences*, vol. 14, no. 15, p. 6649, 2024.

[11] G. Aroganam, N. Manivannan, and D. Harrison, “Review on wearable technology sensors used in consumer sport applications,” *Sensors*, vol. 19, no. 9, p. 1983, 2019.

[12] E. J. Post, “Sagnac effect,” *Reviews of Modern Physics*, vol. 39, no. 2, p. 475, 1967.

[13] V. Patel, A. Orchanian-Cheff, R. Wu *et al.*, “Evaluating the validity and utility of wearable technology for continuously monitoring patients in a hospital setting: systematic review,” *JMIR mHealth and uHealth*, vol. 9, no. 8, p. e17411, 2021.

[14] D. Fuller, E. Colwell, J. Low, K. Orychock, M. A. Tobin, B. Simango, R. Buote, D. Van Heerden, H. Luan, K. Cullen *et al.*, “Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review,” *JMIR mHealth and uHealth*, vol. 8, no. 9, p. e18694, 2020.

[15] B. Haslam, A. Gordhandas, C. Ricciardi, G. Verghese, and T. Heldt, “Distilling clinically interpretable information from data collected on next-generation wearable sensors,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 1729–1732.

[16] M. G. Siegel, M. J. Rossi, and J. H. Lubowitz, “Artificial intelligence and machine learning may resolve health care information overload,” 2024.

[17] J. Russell Stuart and P. Norvig, *Artificial intelligence: a modern approach*. Prentice Hall Upper Saddle River, 2009.

[18] Y. Yan, *Machine Learning Fundamentals*. John Wiley Sons, Ltd, 2022, ch. 2, pp. 19–46. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119817512.ch2

[19] F. J. Reis, R. K. Alaiti, C. S. Vallio, and L. Hespanhol, “Artificial intelligence and machine-learning approaches in sports: Concepts, applications, challenges, and future perspectives,” *Brazilian Journal of Physical Therapy*, p. 101083, 2024.

[20] O. Chapelle, B. Scholkopf, and A. Zien, “Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews],” *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[21] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.

[22] X. Zhu, *Semi-supervised learning with graphs*. Carnegie Mellon University, 2005.

[23] L. Lange, N. Wenzlitschke, and E. Rahm, “Generating synthetic health sensor data for privacy-preserving wearable stress detection,” *Sensors*, vol. 24, no. 10, p. 3052, 2024.

[24] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller, “Synthetic data–what, why and how?” *arXiv preprint arXiv:2205.03257*, 2022.

[25] GeeksforGeeks, “Synthetic Data Generation,” 2024, accessed: 2025-02-23. [Online]. Available: https://www.geeksforgeeks.org/synthetic-data-generation/#architecture-and-techniques-of-synthetic-data-generation

[26] K. S. Seiler and G. Ø. Kjerland, “Quantifying training intensity distribution in elite endurance athletes: is there evidence for an “optimal” distribution?” *Scandinavian journal of medicine & science in sports*, vol. 16, no. 1, pp. 49–56, 2006.

[27] Moez Ali, “Synthetic Data Generation: A Hands-On Guide in Python,” 2024, accessed: 2025-02-23. [Online]. Available: https://www.datacamp.com/tutorial/synthetic-data-generation

[28] A. Seth, J. L. Hicks, T. K. Uchida, A. Habib, C. L. Dembia, J. J. Dunne, C. F. Ong, M. S. DeMers, A. Rajagopal, M. Millard *et al.*, “Opensim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement,” *PLoS computational biology*, vol. 14, no. 7, p. e1006223, 2018.

[29] M. Bulat, N. K. Can, Y. Z. Arslan, and W. Herzog, “Musculoskeletal simulation tools for understanding mechanisms of lower-limb sports injuries,” *Current Sports Medicine Reports*, vol. 18, no. 6, pp. 210–216, 2019.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.

[31] D. P. Kingma, M. Welling *et al.*, “Auto-encoding variational bayes,” 2013.

[32] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. ” O’Reilly Media, Inc.”, 2022.

[33] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *International journal of computer science*, vol. 1, no. 2, pp. 111–117, 2006.

[34] D. Patel, "Data discretization," *Medium*, 2022, accessed: 2025-02-03.

[35] GeeksforGeeks, "What is data normalization?" 2024, accessed: 2025-02-04. [Online]. Available: https://www.geeksforgeeks.org/what-is-data-normalization/

[36] M. Kellmann, M. Bertollo, L. Bosquet, M. Brink, A. J. Coutts, R. Duffield, D. Erlacher, S. L. Halson, A. Hecksteden, J. Heidari *et al.*, "Recovery and performance in sport: consensus statement," *International journal of sports physiology and performance*, vol. 13, no. 2, pp. 240–245, 2018.

[37] M. Kellmann, "Preventing overtraining in athletes in high-intensity sports and stress/recovery monitoring," *Scandinavian journal of medicine & science in sports*, vol. 20, pp. 95–102, 2010.

[38] O.-P. Nuuttila, A. Nummela, E. Korhonen, K. Häkkinen, and H. Kyröläinen, "Individualized endurance training based on recovery and training status in recreational runners," *Medicine and science in sports and exercise*, vol. 54, no. 10, p. 1690, 2022.

[39] F. Crellin, "Hockey-dutch women's coach sticks with heart-rate monitors to build a powerhouse," *Reuters*, August 2 2024. [Online]. Available: https://www.reuters.com/sports/olympics/hockey-dutch-womens-coach-sticks-with-heart-rate-monitors-build-powerhouse-2024-08-02

[40] C. M. Kienstra, T. R. Asken, J. D. Garcia, V. Lara, and T. M. Best, "Triathlon injuries: Transitioning from prevalence to prediction and prevention," *Current sports medicine reports*, vol. 16, no. 6, pp. 397–403, 2017.

[41] S. L. Halson, "Monitoring training load to understand fatigue in athletes," *Sports medicine*, vol. 44, no. Suppl 2, pp. 139–147, 2014.

[42] D. V. Petrovsky, V. I. Pustovoyt, K. S. Nikolsky, K. A. Malsagova, A. T. Kopylov, A. A. Stepanov, V. R. Rudnev, E. I. Balakin, and A. L. Kaysheva, "Tracking health, performance and recovery in athletes using machine learning," *Sports*, vol. 10, no. 10, p. 160, 2022.

[43] J. A. Rothschild, T. Stewart, A. E. Kilding, and D. J. Plews, "Predicting daily recovery during long-term endurance training using machine learning analysis," *European Journal of Applied Physiology*, pp. 1–12, 2024.

[44] A. Naglah, F. Khalifa, A. Mahmoud, M. Ghazal, P. Jones, T. Murray, A. S. Elmaghraby, and A. El-Baz, "Athlete-customized injury prediction using training load statistical re-

cords and machine learning," in *2018 IEEE international symposium on signal processing and information technology (ISSPIT)*. IEEE, 2018, pp. 459–464.

[45] F. Ayala, A. López-Valenciano, J. A. G. Martín, M. D. S. Croix, F. J. Vera-Garcia, M. del Pilar Garcia-Vaquero, I. Ruiz-Pérez, and G. D. Myer, "A preventive model for hamstring injuries in professional soccer: Learning algorithms," *International journal of sports medicine*, vol. 40, no. 05, pp. 344–353, 2019.

[46] Q. Chen, "Identification of potential injury risk factors and prediction model construction of athletes using data mining algorithm," *Journal of Electrical Systems*, vol. 20, no. 6s, pp. 1048–2058, 2024.

[47] N. Rommers, R. Rössler, E. Verhagen, F. Vandecasteele, S. Verstockt, R. Vaeyens, M. Lenoir, E. D'Hondt, and E. Witvrouw, "A machine learning approach to assess injury risk in elite youth football players," *Medicine and science in sports and exercise*, vol. 52, no. 8, pp. 1745–1751, 2020.

[48] J. L. Oliver, F. Ayala, M. B. D. S. Croix, R. S. Lloyd, G. D. Myer, and P. J. Read, "Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players," *Journal of science and medicine in sport*, vol. 23, no. 11, pp. 1044–1048, 2020.

[49] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[50] M. Ehrhart, B. Resch, C. Havas, and D. Niederseer, "A conditional gan for generating time series data for stress detection in wearable physiological sensor data," *Sensors*, vol. 22, no. 16, p. 5969, 2022.

[51] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[52] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[53] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[54] K. L. Hamstra-Wright, K. C. Huxel Bliven, and C. Napier, "Training load capacity, cumulative risk, and bone stress injuries: a narrative review of a holistic approach," *Frontiers in sports and active living*, vol. 3, p. 665683, 2021.

[55] J. H. Falk Neto, E. C. Parent, V. Vleck, and M. D. Kennedy, "The training characteristics of recreational-level triathletes: Influence on fatigue and health," *Sports*, vol. 9, no. 7, p. 94, 2021.

[56] M. G. C. Assis, J. G. G. Barbosa Junior, A. Seffrin, V. Ribeiro dos Anjos Souza, L. Vivan, M. A. Matos Rodrigues, C. A. B. de Lira, R. L. Vancini, B. Knechtle, K. Weiss *et al.*, "Maximal oxygen uptake, muscular oxidative capacity, and ventilatory threshold in amateur triathletes: Eight-month training follow-up," *Open Access Journal of Sports Medicine*, pp. 9–17, 2024.

[57] A. D'Souza, A. Bucchi, A. B. Johnsen, S. J. R. Logantha, O. Monfredi, J. Yanni, S. Prehar, G. Hart, E. Cartwright, U. Wisloff *et al.*, "Exercise training reduces resting heart rate via downregulation of the funny channel hcn4," *Nature communications*, vol. 5, no. 1, p. 3775, 2014.

[58] C. Castro-Diehl, A. V. Diez Roux, S. Redline, T. Seeman, P. McKinley, R. Sloan, and S. Shea, "Sleep duration and quality in relation to autonomic nervous system measures: the multi-ethnic study of atherosclerosis (mesa)," *Sleep*, vol. 39, no. 11, pp. 1927–1940, 2016.

[59] H. A. Young and D. Benton, "Heart-rate variability: a biomarker to study the influence of nutrition on physiological and psychological health?" *Behavioural pharmacology*, vol. 29, no. 2 and 3, pp. 140–151, 2018.

[60] E. Ralevski, I. Petrakis, and M. Altemus, "Heart rate variability in alcohol use: a review," *Pharmacology Biochemistry and Behavior*, vol. 176, pp. 83–92, 2019.

[61] D. J. Bentley, J. Newell, and D. Bishop, "Incremental exercise test design and analysis: implications for performance diagnostics in endurance athletes," *Sports medicine*, vol. 37, pp. 575–586, 2007.

[62] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," *Journal of the american college of cardiology*, vol. 37, no. 1, pp. 153–156, 2001.

[63] F. Billaut and D. Bishop, "Muscle fatigue in males and females during multiple-sprint exercise," *Sports medicine*, vol. 39, pp. 257–278, 2009.

[64] E. F. Coyle, "Integration of the physiological factors determining endurance performance ability," *Exercise and sport sciences reviews*, vol. 23, no. 1, pp. 25–64, 1995.

[65] K. Mj, "The effects of training on heart rate: a longitudinal study," *Ann med exp biol fenn*, vol. 35, pp. 307–315, 1957.

[66] S. Lacke, "Women in triathlon: Where are we now?" 2020, accessed: 7-Apr-2025. [Online]. Available: https://www.triathlete.com/culture/news/women-in-triathlon-where-are-we-now?

[67] D. J. Plews, P. B. Laursen, J. Stanley, A. E. Kilding, and M. Buchheit, "Training adaptation and heart rate variability in elite endurance athletes: opening the door to effective monitoring," *Sports medicine*, vol. 43, pp. 773–781, 2013.

[68] C. R. Bellenger, J. T. Fuller, R. L. Thomson, K. Davison, E. Y. Robertson, and J. D. Buckley, "Monitoring athletic training status through autonomic heart rate regulation: a systematic review and meta-analysis," *Sports medicine*, vol. 46, pp. 1461–1486, 2016.

[69] A. Hamilton, "Lactate threshold: 2-2 t-20," N/A, accessed: 7-Apr-2025. [Online]. Available: https://www.sportsperformancebulletin.com/training/base-endurance-training/lactate-threshold-2-2-t-20

[70] P. J. Puccinelli, C. A. de Lira, R. L. Vancini, P. T. Nikolaidis, B. Knechtle, T. Rosemann, and M. S. Andrade, "The performance, physiology and morphology of female and male olympic-distance triathletes," in *Healthcare*, vol. 10, no. 5. MDPI, 2022, p. 797.

[71] A. Johnson, "Power comparison: Podium finishers of the past four years," N/A, accessed: 1-May-2025. [Online]. Available: https://www.trainingpeaks.com/blog/power-comparison-podium-finishers-of-the-past-four-years/

[72] D. J. Taylor, "Transitioning from participating triathlete to competitive triathlete," 2021, accessed: 1-May-2025. [Online]. Available: https://www.triathlete.com/training/transitioning-from-participating-triathlete-to-competitive-triathlete/

[73] S. L. Halson, "Sleep in elite athletes and nutritional interventions to enhance sleep," *Sports medicine*, vol. 44, no. Suppl 1, pp. 13–23, 2014.

[74] A. Hunter, "The development of the training stress score," N.A., accessed: 2025-04-02. [Online]. Available: https://www.trainingpeaks.com/blog/the-development-of-the-training-stress-score/

[75] G. Support, "What is the exercise load feature on my garmin® device?" 2024, accessed: 3-Apr-2025. [Online]. Available: https://support.garmin.com/en-US/?faq=YhApFCgKQN64763OBiHGf5

[76] Ironman, "Plan your 2025 race calendar in europe south africa: Spring and summer race schedule confirmed," 2024, accessed: 7-Apr-2025. [Online]. Available: https://www.ironman.com/news/plan-your-2025-race-calendar-europe-south-africa-spring-and-summer-race-schedule-confirmed

[77] Z. Wang, Y. □. Wang, W. Gao, and Y. Zhong, "Effects of tapering on performance in endurance athletes: A systematic review and meta-analysis," *PloS one*, vol. 18, no. 5, p. e0282838, 2023.

[78] T. L. Stöggl and B. Sperlich, "The training intensity distribution among well-trained and elite endurance athletes," *Frontiers in physiology*, vol. 6, p. 295, 2015.

[79] Garmin, "Garmin run coach," 2024, accessed: 3-Apr-2025. [Online]. Available: https://www.garmin.com/en-IE/garmin-technology/garmin-coach/garmin-run-coach/

[80] J. Friel, "K.i.s.s periodization," 2010, accessed: 3-Apr-2025. [Online]. Available: https://joefrieltraining.com/kiss-periodization/

[81] G. Whyte, "How to increase training volume?" 2024, accessed: 3-Apr-2025. [Online]. Available: https://www.220triathlon.com/training/training-plans/how-to-increase-training-volume

[82] A. Boike, "Progressive overload," 2024, accessed: 3-Apr-2025. [Online]. Available: https://www.performancerungym.com/running-tips/progressive-overload

[83] B.-A. Burick, "So…what is a deload week and do you need one?" 2024, accessed: 7-Apr-2025. [Online]. Available: https://barbend.com/deload-week

[84] M. Lebrun, "The dos and don'ts of deload weeks," 2024, accessed: 7-Apr-2025. [Online]. Available: https://www.runnersworld.com/training/a62149398/deload-weeks/

[85] Triathlete.com, "What's the right way to approach a recovery week?" 2018, accessed: 7-Apr-2025. [Online]. Available: https://www.triathlete.com/training/recovery/dear-coach-how-should-i-approach-a-recovery-week/

[86] C. Quinn, "Tri-al error 12: What is the typical distribution of training swim/bike/run for triathletes?" 2023, accessed: 4-Apr-2025. [Online]. Available: https://medium.com/@thrillchester/tri-al-error-12-what-is-the-typical-distribution-of-training-swim-bike-run-for-triathletes-827b49e377da

[87] N. Odell, "Balancing swim-bike-run in triathlon training," N/A, accessed: 4-Apr-2025. [Online]. Available: https://www.trainingpeaks.com/blog/balancing-swim-bike-run-in-triathlon-training/

[88] . T. Team, "80/20 training: What it is and how it works," 2019, accessed: 4-Apr-2025. [Online]. Available: https://www.220triathlon.com/training/80-20-training-what-it-is-and-how-it-works

[89] M. Fitzgerald, "How to use the 80/20 training rule to nail your triathlon," N/A, accessed: 4-Apr-2025. [Online]. Available: https://www.trainingpeaks.com/blog/using-the-80-20-rule-to-balance-triathlon-training-intensity/

[90] A. Freeman, "80/20 triathlon training: The what, why, and how," 2024, accessed: 4-Apr-2025. [Online]. Available: https://www.triathlete.com/training/80-20-triathlon-training-the-what-why-and-how/

[91] C. Farrell and D. R. Turgeon, "Normal versus chronic adaptations to aerobic exercise," 2021.

[92] J. Leavitt, "How much deep, light, and rem sleep do you need?" 2024, accessed: 7-Apr-2025. [Online]. Available: https://www.healthline.com/health/how-much-deep-sleep-do-you-need

[93] WHOOP, "Understanding sleep cycles and the stages of sleep," 2019, accessed: 7-Apr-2025. [Online]. Available: https://www.whoop.com/us/en/thelocker/stages-of-sleep-cycles/

[94] Garmin, "How garmin watches track your sleep, calculate sleep score," 2025, accessed: 7-Apr-2025. [Online]. Available: https://www.garmin.com/en-US/blog/fitness/how-garmin-watches-track-your-sleep-calculate-sleep-score/

[95] L. A. Cunha, J. A. Costa, E. A. Marques, J. Brito, M. Lastella, and P. Figueiredo, "The impact of sleep interventions on athletic performance: a systematic review," *Sports Medicine-Open*, vol. 9, no. 1, p. 58, 2023.

[96] G. Technology, "Understanding hrv status on your garmin device," N/A, accessed: 4-Apr-2025. [Online]. Available: https://www.garmin.com/en-US/garmin-technology/health-science/hrv-status/

[97] Polar, "Heart rate zones," 2025, accessed: 7-Apr-2025. [Online]. Available: https://www.polar.com/en/guide/heart-rate-zones

[98] T. Hogue, "Coros heart rate zones: The ultimate guide," 2024, accessed: 7-Apr-2025. [Online]. Available: https://coros.com/stories/coros-metrics/c/coros-heart-rate-zones-the-ultimate-guide

[99] WHOOP, "A simple guide to training with heart rate zones," 2025, accessed: 7-Apr-2025. [Online]. Available: https://www.whoop.com/us/en/thelocker/max-heart-rate-training-zones/

[100] E. F. Coyle and J. Gonzalez-Alonso, "Cardiovascular drift during prolonged exercise: new perspectives," *Exercise and sport sciences reviews*, vol. 29, no. 2, pp. 88–92, 2001.

[101] R. Nelesen, Y. Dar, K. Thomas, and J. E. Dimsdale, "The relationship between fatigue and cardiac functioning," *Archives of internal medicine*, vol. 168, no. 9, pp. 943–949, 2008.

[102] J. L. Storniolo, R. Esposti, and P. Cavallari, "Heart rate kinetics and sympatho-vagal balance accompanying a maximal sprint test," *Frontiers in Psychology*, vol. 10, p. 2950, 2020.

[103] B. Roels, L. Schmitt, S. Libicz, D. Bentley, J. Richalet, and G. Millet, "Specificity of v˙ o2max and the ventilatory threshold in free swimming and cycle ergometry: comparison between triathletes and swimmers," *British journal of sports medicine*, vol. 39, no. 12, pp. 965–968, 2005.

[104] R. Bartels, E. Prodel, M. C. Laterza, J. R. P. de Lima, and T. Peçanha, "Heart rate recovery fast-to-slow phase transition: Influence of physical fitness and exercise intensity," *Annals of Noninvasive Electrocardiology*, vol. 23, no. 3, p. e12521, 2018.

[105] D. Río-Rodríguez, E. Iglesias-Soler, and M. Fernandez del Olmo, "Set configuration in resistance exercise: muscle fatigue and cardiovascular effects," *PLoS One*, vol. 11, no. 3, p. e0151163, 2016.

[106] A. Coggan, "Cycling power zones explained," N/A, accessed: 7-Apr-2025. [Online]. Available: https://www.trainingpeaks.com/blog/power-training-levels/

[107] A. Sargeant, "Human power output and muscle fatigue," *International journal of sports medicine*, vol. 15, no. 03, pp. 116–121, 1994.

[108] C. R. Abbiss and P. B. Laursen, "Models to explain fatigue during prolonged endurance cycling," *Sports medicine*, vol. 35, pp. 865–898, 2005.

[109] A. Nimmerichter, R. G. Eston, N. Bachl, and C. Williams, "Longitudinal monitoring of power output and heart rate profiles in elite cyclists," *Journal of sports sciences*, vol. 29, no. 8, pp. 831–839, 2011.

[110] S. Seiler and E. Tønnessen, "Intervals, thresholds, and long slow distance: the role of intensity and duration in endurance training." *Sportscience*, vol. 13, 2009.

[111] J. C. Martin, D. L. Milliken, J. E. Cobb, K. L. McFadden, and A. R. Coggan, "Validation of a mathematical model for road cycling power," *Journal of applied biomechanics*, vol. 14, no. 3, pp. 276–291, 1998.

[112] J. Friel, "Joe friel's quick guide to setting zones," N/A, accessed: 7-Apr-2025. [Online]. Available: https://www.trainingpeaks.com/learn/articles/joe-friel-s-quick-guide-to-setting-zones/

[113] Garmin International, *Garmin Connect Activity API Documentation*, confidential internal documentation. Not publicly available.

[114] N. B. Murray, T. J. Gabbett, A. D. Townshend, and P. Blanch, "Calculating acute: chronic workload ratios using exponentially weighted moving averages provides a more sensitive indicator of injury likelihood than rolling averages," *British journal of sports medicine*, vol. 51, no. 9, pp. 749–754, 2017.

[115] TrainingPeaks, "What is the performance management chart?" N/A, accessed: 7-Apr-2025. [Online]. Available: https://www.trainingpeaks.com/learn/articles/what-is-the-performance-management-chart/

[116] Garmin, "What is the acute and chronic load feature of my garmin device?" N/A, accessed: 7-Apr-2025. [Online]. Available: https://support.garmin.com/en-US/?faq=C6iHdy0SS05RkoSVbFz066

[117] D. J. Plews, P. B. Laursen, A. E. Kilding, and M. Buchheit, "Heart rate variability in elite triathletes, is variation in variability the key to effective training? a case comparison," *European journal of applied physiology*, vol. 112, no. 11, pp. 3729–3741, 2012.

[118] M. Buchheit, "Monitoring training status with hr measures: do all roads lead to rome?" *Frontiers in physiology*, vol. 5, p. 73, 2014.

[119] M. D. Milewski, D. L. Skaggs, G. A. Bishop, J. L. Pace, D. A. Ibrahim, T. A. Wren, and A. Barzdukas, "Chronic lack of sleep is associated with increased sports injuries in adolescent athletes," *Journal of Pediatric Orthopaedics*, vol. 34, no. 2, pp. 129–133, 2014.

[120] H. H. Fullagar, S. Skorski, R. Duffield, D. Hammes, A. J. Coutts, and T. Meyer, "Sleep and athletic performance: the effects of sleep loss on exercise performance, and physiological and cognitive responses to exercise," *Sports medicine*, vol. 45, no. 2, pp. 161–186, 2015.

[121] A. Ivarsson, U. Johnson, M. B. Andersen, U. Tranaeus, A. Stenling, and M. Lindwall, "Psychosocial factors and sport injuries: meta-analyses for prediction and prevention," *Sports medicine*, vol. 47, pp. 353–365, 2017.

[122] A. Esmaeili, W. G. Hopkins, A. M. Stewart, G. P. Elias, B. H. Lazarus, and R. J. Aughey, "The individual and combined effects of multiple factors on the risk of soft tissue non-contact injuries in elite team sport athletes," *Frontiers in physiology*, vol. 9, p. 1280, 2018.

[123] T. J. Gabbett, "The training—injury prevention paradox: should athletes be training smarter and harder?" *British journal of sports medicine*, vol. 50, no. 5, pp. 273–280, 2016.

[124] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[125] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[126] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[127] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[128] B. J. Maron and A. Pelliccia, "The heart of trained athletes: cardiac remodeling and the risks of sports, including sudden death," *Circulation*, vol. 114, no. 15, pp. 1633–1644, 2006.

[129] Garmin, "2024 Garmin Connect data report," https://www.garmin.com/en-US/blog/general/2024-garmin-connect-data-report/, 2024, accessed: 2025-04-26.

[130] F. Togo and M. Takahashi, "Heart rate variability in occupational health—a systematic review," *Industrial health*, vol. 47, no. 6, pp. 589–602, 2009.

[131] J. D. Minkel, S. Banks, O. Htaik, M. C. Moreta, C. W. Jones, E. L. McGlinchey, N. S. Simpson, and D. F. Dinges, "Sleep deprivation and stressors: evidence for elevated negative affect in response to mild stressors when sleep deprived." *Emotion*, vol. 12, no. 5, p. 1015, 2012.

[132] S. Seiler, O. Haugen, E. Kuffel *et al.*, "Autonomic recovery after exercise in trained athletes: intensity and duration effects," *Medicine and science in sports and exercise*, vol. 39, no. 8, p. 1366, 2007.

[133] J. Stanley, J. M. Peake, and M. Buchheit, "Cardiac parasympathetic reactivation following exercise: implications for training prescription," *Sports medicine*, vol. 43, pp. 1259–1277, 2013.

[134] J. Achten and A. E. Jeukendrup, "Heart rate monitoring: applications and limitations," *Sports medicine*, vol. 33, pp. 517–538, 2003.

[135] L. Bosquet, S. Merkari, D. Arvisais, and A. E. Aubert, "Is heart rate a convenient tool to monitor over-reaching? a systematic review of the literature," *British journal of sports medicine*, vol. 42, no. 9, pp. 709–714, 2008.

[136] C. Hausswirth, J. Louis, A. Aubry, G. Bonnet, R. Duffield, and Y. Le Meur, "Evidence of disturbed sleep and increased illness in overreached endurance athletes," *Medicine and science in sports and exercise*, 2014.

[137] M. Lastella, G. E. Vincent, R. Duffield, G. D. Roach, S. L. Halson, L. J. Heales, and C. Sargent, "Can sleep be used as an indicator of overreaching and overtraining in athletes?" *Frontiers in Physiology*, vol. 9, p. 436, 2018.

[138] E. A. Copenhaver and A. B. Diamond, "The value of sleep on athletic performance, injury, and recovery in the young athlete," *Pediatric annals*, vol. 46, no. 3, pp. e106–e111, 2017.

[139] P. Hellard, M. Avalos, F. Guimaraes, J.-F. Toussaint, and D. B. Pyne, "Training-related risk of common illnesses in elite swimmers over a 4-yr period." *Medicine and science in sports and exercise*, vol. 47, no. 4, pp. 698–707, 2015.

[140] J. B. Mann, K. R. Bryant, B. Johnstone, P. A. Ivey, and S. P. Sayers, "Effect of physical and academic stress on illness and injury in division 1 college football players," *The Journal of Strength & Conditioning Research*, vol. 30, no. 1, pp. 20–25, 2016.

[141] G. S. Collins, J. A. de Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L.-M. Yu, K. G. Moons *et al.*, "External validation of multivariable prediction models: a systematic review of methodological conduct and reporting," *BMC medical research methodology*, vol. 14, pp. 1–11, 2014.

[142] R. Ø. Nielsen, E. T. Parner, E. A. Nohr, H. Sørensen, M. Lind, and S. Rasmussen, "Excessive progression in weekly running distance and risk of running-related injuries: an association which varies according to type of injury," *journal of orthopaedic & sports physical therapy*, vol. 44, no. 10, pp. 739–747, 2014.

[143] G. P. Millet, R. B. Candau, B. Barbier, T. Busso, J.-D. Rouillon, and J. C. Chatard, "Modelling the transfers of training effects on performance in elite triathletes," *International journal of sports medicine*, vol. 23, no. 01, pp. 55–63, 2002.

[144] A. Hreljac, "Impact and overuse injuries in runners," *Medicine & Science in Sports & Exercise*, vol. 36, no. 5, pp. 845–849, 2004.

[145] G. Bruinvels, R. Burden, N. Brown, T. Richards, and C. Pedlar, “The prevalence and impact of heavy menstrual bleeding (menorrhagia) in elite and non-elite athletes,” *PLoS One*, vol. 11, no. 2, p. e0149881, 2016.

[146] H. Chtourou and N. Souissi, “The effect of training at a specific time of day: a review,” *The Journal of Strength & Conditioning Research*, vol. 26, no. 7, pp. 1984–2005, 2012.

[147] G. L. Close, C. Sale, K. Baar, and S. Bermon, “Nutrition for the prevention and treatment of injuries in track and field athletes,” *International journal of sport nutrition and exercise metabolism*, vol. 29, no. 2, pp. 189–197, 2019.

Declaration of Aids and Tools

Aid / Tool	**Intended use**
Claude	Rewording sections; Support with interpreting outputs of exploratory data analysis; Support with research work
ChatGPT	Rewording sections; Support with interpreting outputs of exploratory data analysis; Support with research work
DeepL Write	Rewording throughout the work
GitHub Copilot	Assistance with coding and script development
Overleaf LaTeX Editor	Document preparation
Python in Visual Studio Code	Data simulation and analysis

A Appendix

A.1 GitHub Repository

The complete code implementation for this thesis is available in the following GitHub repository:

https://github.com/redsleo02/Bachelorarbeit

A.1.1 Repository Structure

The repository is organized into the following key directories:

- `/synthetic_data_generation/`: Contains the implementation of the synthetic data simulation framework.
- `/notebooks/`: Contains Jupyter notebooks documenting the analytical process:
  - `1_eda.ipynb`: Initial analysis of synthetic dataset characteristics
  - `2_preprocessing_feature_engineering.ipynb`: Data preparation and transformation procedures
  - `3_ML_analysis.ipynb`: Implementation and evaluation of machine learning models
- `/explored_unused_solutions/`: Contains alternative approaches explored during research, that can be used to gain access to real data given more time:
  - Web interface prototype to collect data through the Garmin Connect API
  - Implementation for Strava API data access and authentication

For detailed documentation, please refer to the README.md file in the repository.

A.2 Access to Synthetic Datasets

The complete synthetic datasets generated during this research exceed GitHub's file size limitations and therefore are not included directly in the repository. These datasets can be accessed through the following alternative methods:

- **Zenodo Repository**: All synthetic datasets have been archived with persistent DOIs on Zenodo at:
  https://doi.org/10.5281/zenodo.15401061
- **Dataset Regeneration**: The complete datasets can be regenerated by running the main simulation script from the repository root:

```
python synthetic_data_generation/main.py
```

A.2.1 Dataset Description

The synthetic datasets include:

Dataset	Size	Description
athletes.csv	465.2 kB	Contains simulated profiles for 1,000 athletes including physiological parameters and lifestyle factors
activity_data.csv	115.4 MB	Daily training loads, intensities, and durations across a simulated annual cycle
daily_data.csv	71.2 MB	Daily recovery markers including HRV, sleep quality, astress, and injury labels

A.3 Declaration of Authorship

The following text is to be attached to the thesis.

“I hereby declare,

- that I have written this thesis independently,
- that I have written the thesis using only the aids specified in the index;
- that all parts of the thesis produced with the help of aids have been declared;
- that I have handled both input and output responsibly when using AI. I confirm that I have therefore only read in public data or data released with consent and that I have checked, declared and comprehensibly referenced all results and/or other forms of AI assistance in the required form and that I am aware that I am responsible if incorrect content, violations of data protection law, copyright law or scientific misconduct (e.g. plagiarism) have also occurred unintentionally;
- that I have mentioned all sources used and cited them correctly according to established academic citation rules;
- that I have acquired all immaterial rights to any materials I may have used, such as images or graphics, or that these materials were created by me;
- that the topic, the thesis or parts of it have not already been the object of any work or examination of another course, unless this has been expressly agreed with the faculty member in advance and is stated as such in the thesis;
- that I am aware of the legal provisions regarding the publication and dissemination of parts or the entire thesis and that I comply with them accordingly;
- that I am aware that my thesis can be electronically checked for plagiarism and for third-party authorship of human or technical origin and that I hereby grant the University of St.Gallen the copyright according to the Examination Regulations as far as it is necessary for the administrative actions;
- that I am aware that the University will prosecute a violation of this Declaration of Authorship and that disciplinary as well as criminal consequences may result, which may lead to expulsion from the University or to the withdrawal of my title.”

By submitting this thesis, I confirm through my conclusive action that I am submitting the Declaration of Authorship, that I have read and understood it, and that it is true.

Date and signature

……………………………………………