Universiteit Leiden

# A Data-Driven Analysis of Personalized Performance Predictors in Competitive Endurance Runners

Bilal Nour (s3874095)

Supervisors:
Dr. A. Knobbe & Guus Toussaint

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 18, 2026

**Abstract**

In competitive endurance running, optimizing an athlete's training plan is critical for their success. With the introduction of commercial smartwatches that are capable of tracking heart rate and GPS location, it is possible to identify patterns in an athlete's preparation that produced a given performance. However, due to athletes being capable of competing only a set number of times each year and the relatively recent widespread adoption of watch data, there are significant constraints on the models that can be used due to data sparsity. These constraints are further tightened as coaches and athletes alike demand models that are highly interpretable to have any utility in logically adjusting training. Therefore, this study uses highly interpretable models to evaluate and identify the relationship between the collected wearable data and final performance of an athlete (measured in IAAF score). For global modeling, linear Mixed-Effects and Lasso models were deployed to discover the presence of any linear relationships within the data. This was followed by a Subgroup Discovery that was used to find non-linear localized pockets of data that were considered interesting. The final results presented a clear conclusion that global linear models largely failed to explain any variance within the data, indicating that the relationship between training load and race performance cannot be effectively modeled linearly or pooled across different athletes. Conversely, Subgroup Discovery successfully identified a number of significant thresholds for the athletes. Ultimately, this research found that translating sparse wearable data into practical coaching insights requires personalized, non-linear subgroup discovery rather than global modeling.

# Contents

# 1 Introduction

Endurance running is defined as any continuous running effort where the aerobic energy system is the primary pathway for energy production. These events typically encompass distances of 1500 m and greater, taking place primarily on roads, trails, or 400 m tracks. To minimize the time required to traverse a given distance, elite athletes rely on highly developed physiological attributes that allow them to maintain elevated speeds. Among these attributes, three are widely agreed to have the greatest influence [6]: maximal oxygen uptake ($VO_2$ max), running economy, and lactate threshold. $VO_2$ max represents the absolute aerobic capacity of an athlete and is the maximum rate that oxygen can be transported to working muscles [2]. Running economy quantifies the efficiency of an athlete at a sub maximal pace and is closely related to biomechanical efficiency of an athlete. Finally, lactate threshold is defined as the inflection point where the build-up of lactate exceeds the body's clearance capacity [10].

Beyond these internal physiological attributes, an athlete's performance is also heavily influenced by external factors such as weather [8], stress, and sleep. Analyzing these factors in combination can help identify aspects of an athlete's preparation that may be either detrimental or supportive to their overall performance.

Generally, athletes would use lab tests to calculate their physiological metrics; however, this can be time-consuming and costly, as measurements must be taken repeatedly in short intervals. Although frequent lab testing is unrealistic for the majority of athletes, commercial wearable trackers and sensors are now widely used to measure training volume and heart rate. These sensors collect data on sleep quality and other recovery metrics.

A significant challenge in sports data science is that athletes exhibit extreme variability in their physiological responses to training and external stimuli. This is especially the case for highly competitive athletes who are already significant outliers. Therefore, predictive modeling cannot rely solely on generalized population data and should instead be highly individualized.

Approaching this from a machine learning perspective, predicting race outcomes from daily factors is inherently difficult due to the time series nature present in the data. General predictive models would fail due to their inability to capture the temporal nature of load, namely that load closer to race day has vastly different biological impact (e.g., acute fatigue) than loads accumulated months prior (e.g., chronic base-building).

Besides temporal lag, athletic performance suffers from extreme sparsity. Elite track runners may compete in only 20 maximum effort competitions each year, with this being even fewer for marathon and road runners as the distance causes extreme strain on the athletes' bodies. This directly limits the kind of models that can be used to understand performance, excluding complex models like Recurrent Neural Networks (RNNs). Although RNNs are suitable for time-series data, they are both black box and extremely data hungry. This "black-box" nature makes it difficult for a coach or an athlete to interpret the results and adjust their training accordingly, as the output of the model is extremely non-linear. Instead, coaches are looking for clear, actionable insights, such as specific volume thresholds or the weather sensitivity of their athlete. This could help coaches avoid overtraining athletes or find optimal tapering durations for competitions.

To address these challenges, this study conducts an individualized, hierarchical analysis of four elite

male Dutch endurance runners. The cohort consists of a range of specialists from the marathon down to the 1500 m, all possessing performances that place them within the top 800 globally. Normalizing their results using the World Athletics (IAAF) scoring tables allows us to create a single target variable that can be used to research global relationships using global continuous models (Mixed-Effects and Lasso Regression) as well as non-linear local techniques such as subgroup discovery.

The primary goal of this paper is to extract mathematically validated, interpretable insights for athletes. Specifically, this study is guided by the following research questions:

- Does historical wearable data combined with environmental data carry statistically significant predictive power for elite race performance?
- Can a Mixed-Effects regression framework effectively model physiological load and preserve individual baselines for athletes with sparse performance data?
- Which specific temporal windows and engineered features are the most critical drivers of race-day success?
- What localized, non-linear physiological thresholds exist within the data, and how can they be translated into actionable training constraints for coaches?

# 2 Background

## 2.1 Sports Physiology and Training Load

Physiological improvement occurs when an athlete undergoes a training stimulus, which over an extended time frame forces biological adaptations. The most challenging aspect of an athlete's preparation is identifying the optimal training load that maximizes adaptation while minimizing the risk of injury. Simply increasing the volume or intensity of training does not yield linear improvement; instead, adaptation occurs up to a threshold, beyond which the athlete's risk of overtraining and injury increases drastically. Training load is generally categorized into two groups: external and internal loads. External load refers to the mechanical work completed by the athlete and is measured using metrics such as distance, pace, and overall session duration. Conversely, internal load is the relative physiological and psychological stress experienced by the athlete [9]. One of the clearest internal load metrics is heart rate; it is directly linked to exercise intensity, as working muscles require greater volumes of oxygenated blood to sustain higher energy outputs [3].

Beyond training, environmental conditions on the day of performance can strongly influence race outcomes. High temperatures during a race have repeatedly been shown to degrade endurance performance, as heat can induce severe thermoregulatory strain. This reduces the oxygen delivery to working muscles and causes premature cardiovascular drift [8]. This drift directly causes a reduction in the athlete's capacity to perform and is an important factor that should be considered when modeling performance.

Another environmental factor included in this study is elevation; it is a cornerstone of modern elite athlete training, as athletes use it to increase their $VO_2$ max. Although racing at altitude as an endurance runner can hinder performance due to a lower partial pressure of oxygen (hypoxia),

training, or better yet, living at altitude has been shown to be a powerful physiological stimulus, which improves the athlete's maximum oxygen uptake ($VO_2$ max) and overall aerobic capacity [12]. Therefore it critical to identify the individual utility of altitude training and its impact as a race day condition.

## 2.2 Quantifying Training Load

A significant challenge in sports analytics is the quantification of training load that consists of highly variable intensities. Endurance training uses interval training, characterized by high-intensity reps at a specific pace related to the athlete's goal, followed by a recovery. The high intensities are held for only a fraction of the athlete's total training session and are not accurately accounted for in metrics such as average heart rate or maximum heart rate. Thus, a feature that effectively captures the physiological toll of a session must be utilized to ensure that short but high-intensity training is weighted fairly in comparison to long but low-intensity sessions. To address this, Banister (1976) introduced the Training Impulse (TRIMP) model [5]. TRIMP multiplies the duration of an effort by a heart rate-derived intensity coefficient, applying an exponential weighting factor to account for the disproportionate physiological stress that occurs as an athlete approaches their maximum heart rate.

For each recorded training session, the total internal load is computed using the following mathematical formulation:

$$\text{TRIMP} = D \cdot \Delta HR \cdot e^{b \cdot \Delta HR} \tag{1}$$

where:

- $D$ represents the total duration of the training session in minutes.
- $\Delta HR$ represents the fractional heart rate reserve (the normalized cardiovascular intensity), calculated as:

$$\Delta HR = \frac{HR_{ex} - HR_{rest}}{HR_{max} - HR_{rest}} \tag{2}$$

$HR_{ex}$ is the average heart rate during the exercise, $HR_{rest}$ is the athlete's resting heart rate, and $HR_{max}$ is the athlete's absolute maximum heart rate. The exponent coefficient $b$ controls the slope of the lactate curve. In alignment with established physiological literature for male athletes, $b$ is set to the standard constant of 1.92.

## 2.3 IAAF score

Endurance athletes rarely compete in a single distance, as training multiple disciplines builds different aspects of fitness in preparation for their specialization. For example, a 1500 m specialist would do a shorter 800 m race in order to build speed and lactate tolerance for their 1500 m competitions. Additionally, quantifying performance using race times is inherently non-transferable, as a ten-second improvement in the 1500 m is a far more challenging feat than a similar improvement in the 10,000 m.

This demands a scoring system that is comparable across disciplines while accounting for their difficulty. Originally developed by Dr. Bojidar Spiriev in 1982 [16], the World Athletics Federation created the IAAF Scoring Tables. The tables provide a standardized metric that maps absolute race times to a universal point system. This allows for an objective statistical comparison of performances across track and field events. Because these tables are designed exclusively for within-gender analysis, cross-gender performance comparisons are not possible. Thus, the cohort consists exclusively of male athletes, with all reported IAAF scores being calculated using the men's scoring tables.

The scoring framework is non-linear, as it is designed to reflect the biological reality of marginal gains becoming exponentially more difficult to achieve at the elite boundaries of human performance.

As the WA framework is created using the performance distribution across a set period of time, it is critical that these distributions be updated in order to account for the developments in the sport. These shifts can occur due to the introduction of improved equipment, such as super shoes, or improvements in training methodologies. The World Athletics Federation updates its tables every three or so years. To ensure a fair but modern benchmark, we selected the 2022 scoring table [16], as it captures the inflation of performances after the introduction of super shoes and encapsulates the period where most of the performances occurred.

For track events, the points calculation is driven by the following mathematical formula:

$$Points = a \cdot (b - T)^c$$

where:

- $T$ is the athlete's raw performance time in seconds.
- $b$ represents the theoretical baseline limit (a performance threshold yielding zero points).
- $a$ and $c$ are event-specific parameters derived from the global statistical distribution of world-class performances.

The exponent $c$ ensures the curve remains progressive, disproportionately rewarding times that approach or exceed current world records. It should be noted that although the scoring table follows an exponential trend, an equation is not provided by World Athletics. Instead, a scoring table with a threshold scoring system is used that approximately follows the power law outlined above.

## 2.4 Wearable Sensors

Wearable sports sensors have only recently been popularized for general use. While early iterations existed in the 2000s, the mid-2010s saw an explosion in adoption due to the introduction of smartwatches [17]. Thus, using consumer-grade smartwatch data to predict race performance is a relatively understudied subject. The large feature spaces (daily physiological metrics) but relatively sparse data points makes it difficult to find meaningful relationships. The viability of consumer-grade wearables has been proven by studies such as Broughton (2023), who used Fitbit watch data to predict resting heart rate, showcasing that wearable sensors are biologically sufficient to predict physiological variability across an extended period [4]. However, unlike this study, our

primary target variable suffers from severe scarcity, due to athletes competing only a few times each year.

### 2.5 Machine Learning in Sports Analytics

Machine learning has been increasingly applied to elite sports, where championships are won by the smallest of margins. Athletes desire a better understanding of their individual drivers of performance and how they can use this knowledge to maximize their results. Because preparation is a continuous, long-term process, modeling performance inherently requires sequential or time-series approaches.

Modern data science frequently deploys complex architectures such as Recurrent Neural Networks (RNNs) or Gradient Boosted Trees to handle these temporal dependencies. These models present a fundamental tradeoff between predictive complexity and interpretability. While these models can produce highly accurate predictions, they lack the necessary interpretability to be utilized by domain experts without post hoc explainability techniques [4]. Additionally, as athletes compete infrequently, these data-hungry, opaque models become entirely unviable. To bridge this gap between sparse data and actionable insights, researchers have instead used simple, interpretable models rather than deep, black-box architectures [1, 11].

Pooling can be employed when an individual athlete's data is too sparse to support such a model in isolation, allowing the sample size to be artificially increased while individual differences within the cohort are still preserved. Avalos et al. [1] illustrate this combination directly by successfully applying a Linear Mixed-Effects Model (LMM) to a cohort of elite swimmers. By pooling athlete data within a single statistical framework, this approach uncovers athlete-specific relationships between training load and performance that would otherwise remain undetectable given the limited number of observations.

Where sufficient individual data exists, pooling becomes unnecessary. Knobbe [11] addresses precisely this scenario, identifying athlete-specific drivers of performance through Least Absolute Shrinkage and Selection Operator (Lasso) regression and Subgroup Discovery (SD), the latter used to uncover non-linear performance thresholds.

The pooling approach and individualized analyses demonstrate that complex temporal relationships do not require deep sequential layers [1, 11]. Instead, they can be elegantly captured using discrete, time-lagged rolling window aggregations, directly informing the acute and chronic TRIMP windows utilized in this thesis.

## 3 Methodology

This chapter provides the foundational blueprint of the study and how raw data is transformed into interpretable insights of performance. The coming section first explain key detail about the feature engineering pipeline such as mathematical transformation and temporal aggregation. Then it formally defines the global continuous models (Linear Mixed-Effects and Lasso) and local models (Subgroup Discovery). Finally it will introduce the specific performance metrics and how the models will be validated using Cross validation and Swap randomization.

## 3.1 Feature Engineering

In order to use raw continuous data it is critical to first transform the data into mathematical structure that is compatible with predictive modeling. As raw time series data cannot be directly inputted into a global regression model.

Therefore this section introduces a generalized feature engineering pipeline that digests the longitudinal data and outputs features that are compatible with our models. The pipeline consist of four primary stages: creating features that are aligned correctly with the target variable, extracting raw mechanical and environmental metrics, categorizing training into 3 distinct physiological zones, and applying rolling temporal aggregations. The final output of this process is a multidimensional feature matrix that is optimized for both linear regression and non-linear subgroup discovery.

### 3.1.1 Temporal Alignment

To have a valid comparison across multiple endurance disciplines, the standardized World Athletics (IAAF) performance score was defined as the target variable. A fundamental challenge in predicting a time series target variable is preventing temporal data leakage. If a model has access to the physiological data that generated the performance, its predictive validity would be compromised. Thus, to ensure the validity of the model, a temporal shift was introduced across the feature space. This would mean that for a given race occurring on Day $t$, all historical mechanical load and aggregated features were computed utilizing data that was shifted to Day $t-1$. This boundary would ensure that any mechanical and cardiovascular exertion of the race is excluded.

This shift will not be present in race-day environmental conditions, as they serve as a direct acute physiological factors rather than training adaptations. As the conditions of a competition directly impact the final result and provide no direct information of the performance, environmental variables (such as perceived temperature, wind speed, and humidity) were sampled simply as Day $t$.

### 3.1.2 Raw data

To create the final engineered feature, it is critical to first understand the raw data extracted from the commercial wearable trackers and historical weather data. The raw data has two distinct temporal resolutions, the first being session-level metrics which are determined by when the athlete records a training session, and the second being daily-level metrics which are recorded once every 24 hours. In addition to this distinction, raw data can be separated into three distinct domains.

**Internal Load:** These variables capture the cardiovascular strain exerted by the athlete.

- **Average Heart Rate:** The mean cardiovascular intensity sustained during a specific training session (bpm).
- **Maximum Heart Rate:** The absolute peak cardiovascular exertion reached during the session (bpm).
- **Resting Heart Rate:** A daily-level metric representing the athlete's baseline cardiovascular state at complete rest (bpm).

**External Load:** These variables quantify the physical exertion of an athlete. While distance was

extracted, duration serves as the primary anchor for training volume, as it is more indicative of load irrespective of topography or athlete fitness.

- **Duration:** The total active time of the training session (seconds).
- **Distance:** The total spatial distance covered (meters).
- **Elevation Gain:** The total vertical ascent achieved during the activity (meters).

**Environmental Load:** These daily-level metrics represent the weather conditions that act as external influnce to performance.

- **Perceived Temperature (Feels Like):** The ambient air temperature adjusted for wind chill and relative humidity (°C).
- **Humidity:** The relative moisture saturation of the air (%).
- **Wind Speed:** The velocity of the surrounding air currents (km/h).

Additionally, discipline is also included in the dataset as an athlete final IAAF score is directly related to their specialization. Therefore this relationship must be represented in the dataset. It should be noted that this is only the case for subgroup discovery as it can handle nominal data unlike the regression models.

Athletes also frequently train aerobic fitness using other modalities, such as cycling and swimming; therefore, a distinction was made between running and cross-training, as both the physical load and aerobic gains can differ depending on the training type. Instead of defining a single feature exclusively for cross-training, a supporting feature was engineered that accumulated total training volume by adding cross-training on top of the athlete's normal running. This composite feature accurately models cross-training in its true context: as an absolute physiological supplement to baseline running volume rather than an independent training focus.

### 3.1.3 Zones

When using raw wearable data to calculate internal training loads or zones, the data must first be filtered to ensure that no hardware artifacts are used in their calculations. As wearable optical sensors are susceptible to misreadings due to insufficient contact between the sensors and skin or for a number of other factors, they can produce physiologically impossible spikes or artificial resting lows.

**Data Artifact Filtration:** Therefore, in order to ensure the integrity of our engineered features, a strict biological boundary was applied. This boundary was enforced by excluding any heart rate falling below 30 bpm or exceeding 230 bpm. In addition to this, a stable resting heart rate (RHR) was calculated by extracting the 5th percentile of their daily resting measurements across a year. As RHR is critical to future calculations of Zones and Load, it was critical to create a stable value that does not fluctuate drastically due to athlete compliance or other short-term factors.

However, unlike RHR, Maximum Heart Rate ($HR_{max}$) is a relatively stable measurement that slowly degrades at 0.7 bpm regardless of fitness level [18]. In order to ensure that the $HR_{max}$ is a valid peak rather than just a sensor spike, a gradient-based thresholding technique was used on the percentile distribution of the athlete's heart rate data. As such, a valid ($HR_{max}$) was the point at

which the gradient between percentiles spiked significantly above the average gradient. Then, with the assumption that the captured ($HR_{max}$) is valid, it is used as an anchor point. It is applied with the biological decay formula (0.7 bpm $\times$ years elapsed) forward and backward in time from the date of the anchor point [18].

**Fractional Heart Rate Reserve (HRR):** Using filtered $HR_{max}$ and RHR values allows for the normalization of each recorded training session using Fractional Heart Rate Reserve ($\Delta HR$). This metric represents the athlete's true exertion relative to total cardiovascular capacity and acts as a critical variable for both the calculation of Bannister's TRIMP and zones:

$$\Delta HR = \frac{HR_{ex} - HR_{rest}}{HR_{max} - HR_{rest}}$$

where $HR_{ex}$ is the average heart rate during the given exercise.

**Physiological Zoning:** To effectively quantify the physiological stress of a session, traditional endurance planning has three intensity zones. These zones are defined by their physiological markers that are present at a given intensity and are closely linked to heart rate. They are formally defined as aerobic threshold, lactate threshold 1 (LT1), and lactate threshold 2 (LT2) [15].

Due to the athlete lacking a lactate step test, a generalized intensity thresholds was instead applied based on the established literature for mapping $\Delta$HR (Heart Rate Reserve) to the aforementioned thresholds for highly trained athletes [7], the following boundaries were established:$\Delta HR$:

- **Zone 1 (Aerobic Base):** Exertion strictly $\leq 70\%$ of HRR.
- **Zone 2 (Threshold/Tempo):** Exertion between 70% and 80% of HRR.
- **Zone 3 (Anaerobic Capacity):** Exertion between 80% and 90%.

In categorizing the total duration of a session into three distinct buckets, the methodology successfully captures the distribution of intensity, thus producing highly interpretable features that can be used in subsequent models to differentiate changes in intensity and volume.

#### 3.1.4 Aggregation

In order for a model to effectively grasp the temporal significance of a training session, it is critical to create features that effectively represent different windows of time. This is done by aggregating daily metrics using multi-day rolling temporal windows. These windows represent acute fatigue, intermediate adaptation, and chronic baseline fitness.

**Zero-Padding for Rest Days:** Endurance training naturally contains rest days; as such, an athlete's historical training data will have dates with zero recorded sessions. For the mathematical continuity of the rolling windows, the timeline was strictly zero-padded. Thus, any day lacking a recorded session was assumed to be a rest day, making any physiological load or volume set to zero prior to aggregation.

**Mathematical Primitives:** In order to capture different aspects of load within a given window, daily metrics were transformed using the three distinct primitives outlined below:

- **Sum ($\Sigma$):** Used to quantify cumulative volume, such as total duration or the total TRIMP accumulated over the window.
- **Mean ($\mu$):** Used to capture the average physiological or environmental strain.
- **Max (max):** Used to anchor peak exertion markers, such as the $HR_{max}$ of a given training session.

**Acute:Chronic Workload Ratio (ACWR)** An additional feature, ACWR, was added after the aggregation process; this feature gives insight into the athlete's relationship between their baseline fitness and acute fatigue. It is a well-supported feature that can be used to monitor injury risk [13] and performance readiness. It is mathematically defined as:

$$ACWR = \frac{\text{Average Acute Load}}{\text{Average Chronic Load}}$$

The acute load was derived using the cross TRIMP 7-day window, while the chronic load was defined as the cross TRIMP 28-day window. This feature's utility stems from its interpretability, especially in the case of Subgroup Discovery, as it helps identify specific thresholds where an athlete was optimally primed for competition.

## 3.2 Models

When investigating the utility of any model, it is critical to first create a simpler model that acts as a comparison before building more complex models. This ensures that a highly complex model does not underperform in comparison to simpler models, as if the performance does not increase with the complexity of the model, it would invalidate the utility of said model.

For that reason, this study uses a hierarchical framework of progressively complex models. The pipeline begins with a naive baseline estimator and progresses to global regression models (Linear Mixed-Effects and Lasso Regression) and then onto a local nonlinear technique (Exploratory Subgroup Discovery) that identifies specific biological thresholds.

### 3.2.1 The Naive Baseline

The Naive baseline acts as a control model that all future models must exceed in performance. This allows for the effective evaluation of the predictive power of wearable sensors and environmental conditions. It should therefore be noted that the naive baseline model should be sufficiently simple; thus, it will be defined as a trailing mean model. This model operates under the assumption that an athlete's future performance is solely determined by their immediate past performances.

The trailing mean model works by using temporal windows that slide across the athlete's performances. Let $\hat{Y}_t$ represent the predicted standardized performance (expressed as a World Athletics IAAF scoring point total) for an event at time $t$. The generalized formula is formally defined as the mean of the previous three performances:

$$\hat{Y}_t = \frac{1}{3}\sum_{k=1}^{3} Y_{t-k}$$

The decision to use a three-race window ($n = 3$) was mainly to smooth out fluctuations in athlete fitness, as looking at just the current race would be too sensitive to noise; however, selecting a larger window leads to significant data loss, as the selected window size is equal to the amount of data discarded for the buffer.

### 3.2.2 Linear Mixed-Effects Regression

The Linear Mixed-Effects regression is used to simply predict the athlete's performance based solely on their training load. This model is used as it allows for athletes with extremely sparse data (¡ 20) to be included in the study, potentially finding some interesting relationships between load for said individuals. Thus, by pooling the entire athlete cohort into a single dataset, the LMM can identify general population-level responses to training loads (fixed effects) and preserve the unique biological baseline of each athlete (random intercepts).

As LMM optimization is highly sensitive to multicollinearity and high-dimensionality datasets, it is intentionally restricted to a dataset comprising Bannister TRIMP loads that contains features for acute and chronic loads. Before modeling, Group-Mean Centering is applied to the TRIMP metrics to separate within-athlete variance (acute physiological deviation) from between-athlete variance (genetic capacity). Features are then Z-score scaled (mean=0, std=1) so the optimizer treats all features equally regardless of their original units.

The applied architectural formula is defined as:

$$Y_{ij} = (\beta_0 + u_{0j}) + \beta_1 X_{1,ij} + \cdots + \beta_k X_{k,ij} + \epsilon_{ij}$$

Where:

- $Y_{ij}$ is the IAAF performance score of athlete $j$ in race $i$.
- $\beta_0$ is the global fixed intercept.
- $u_{0j}$ is the random intercept for athlete $j$, capturing their unique, inherent physiological baseline.
- $\beta_k$ are the fixed-effect coefficients for the $k$-th physiological feature (e.g., centered acute TRIMP).
- $\epsilon_{ij}$ is the residual error, assumed to be normally distributed as $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

Further optimization of temporal lag for a training load was executed using a grid search, in which an acute and a chronic window were selected from a set of window sizes. This study used window sizes of [1, 7, 14] for the acute window and [28, 90] for the chronic window.

### 3.2.3 Regularized Linear Regression

This study uses Lasso to evaluate the global linear relationship between load and race-day performance. Building upon the LMM model, Lasso provides an increase in complexity by mathematically discovering the specific features that are most indicative of performance and then using them to inform the final predictions. This is in direct contrast with the previous LMM model that has a defined feature space of just training load. As such, the inputs for the two models drastically differ, with Lasso having the capacity to safely handle over 100 features that are highly correlated with overlapping temporal windows. In addition to this, it can handle the ratio imbalance that is present due to the number of features significantly exceeding the number of actual race observations per athlete, which would typically cause any standard OLS model to overfit.

Least Absolute Shrinkage and Selection Operator (Lasso) applies an $L_1$ regularization penalty to the standard optimization objective. This then dynamically shrinks the coefficients of redundant or highly correlated features to exactly zero. Given an athlete's historical race performances $y_i$ and the high-dimensional matrix of engineered features $x_{ij}$, the objective function is defined as:

$$\hat{\beta}^{Lasso} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right\}$$

where $\alpha$ dictates the degree to which the regularization penalty is enforced, and $p$ represents the total number of features in the dataset. $\alpha$ is not just a fixed value, but instead must be learned from the data during the training phase. Therefore, this study uses a nested cross-validation procedure to calculate $\alpha$ and assess the overall stability of the selected features.

### 3.2.4 Subgroup Discovery

Global regression models naturally assume the presence of a continuous and linear relationship between load and performance. However, this is often not the case in human physiology, as there are multiple instances where biology exhibits a nonlinear "step-function" response. For example, racing in high-temperature conditions is plausible up until a specific threshold where severe overheating occurs and performance plummets drastically. This is just one of many examples in human biology that showcases a stable fluctuation in performance until a threshold.

Thus, to identify these localized nonlinear inflection points, subgroup discovery is used, as it can handle the same high-dimensional dataset as Lasso. It does this by searching the hypothesis space to find specific logical rules that isolate pockets of race performance that are considered to be interesting based on an evaluation metric.

The algorithm constructs rules by using a combination of simple logical rules that are then applied to the feature space (e.g., $X_1 \leq c_1 \wedge X_2 > c_2$). These outputs are inherently highly interpretable, making them easy for domain experts to understand and implement changes based on the findings. However, due to the nature of subgroup discovery, it is susceptible to false subgroups that are just noise in the dataset rather than a valid result. Therefore, both support constraints and swap randomization are applied to ensure the validity of the results.

## 3.3 Validation and Evaluation Framework

Before accepting the results of a model as significant, it is first critical to validate them. The validation methods differ for global and local models; for the former, we can test the results on out-of-sample data using standard $k$-fold cross-validation. For the latter, we use swap randomization, which determines a noise threshold in which values below this threshold can be considered statistical artifacts.

Although $k$-fold cross-validation is a standard in machine learning, it is not applicable to this study due to the scarcity of race data. This is because partitioning an already limited dataset leads to highly unstable parameter estimations and stunts the performance of the model. To resolve this without sacrificing the validation process, this study uses Leave-One-Out Cross Validation (LOOCV).

### 3.3.1 LOOCV

This technique maximizes the available training data by iteratively training the model on $N - 1$ observations and testing it on a single held-out instance. This process is repeated $N$ times until every race has been predicted strictly out-of-sample. The final evaluation metric, the Coefficient of Determination ($R^2$), is then calculated globally by aggregating all $N$ independent predictions ($\hat{y}_i$) against the true performance values ($y_i$). This approach was utilized to evaluate both the Linear Mixed-Effects Model (LMM) and the Lasso regression.

However, one vulnerability of LOOCV is that it can introduce temporal data leakage, as it can theoretically use future data to predict the past. This is due to LOOCV shuffling the training data during the validation process. However, this is mitigated through feature engineering, which shifts all race observations (Day $t - 1$); this ensures that each race observation is independent, therefore preserving the integrity of the validation process.

### 3.3.2 Validating Lasso Regression

The previously described LOOCV procedure is only sufficient in cases where a model's parameters can be estimated directly from the training fold. For a Lasso model, this is simply not the case, as it has an additional hyperparameter $\alpha$ that must be selected from within the data. If $\alpha$ was selected using the entire dataset and then tested, it would produce an overly optimistic out-of-sample performance. Thus, to prevent this bias, nested cross-validation was utilized for the Lasso model. It works by finding the optimal $\alpha$ from a separate cross-validation procedure on the training set data that excludes the test observation. This ensures that the $\alpha$ found is sufficiently generalizable to out-of-sample data points.

Applying LOOCV to Lasso might determine the accuracy of our model for unseen races; however, it fails to actually identify the specific features that produced said performance, as for every test-train split, the exact $\alpha$ and features can vary. So, in order to identify and quantify the impact of the features, it is critical to add an additional interpretation phase.

Thus, once the out-of-sample performance is established, a secondary independent Lasso model is fitted to the entire dataset. This model performs no predictions; instead, it is only used to identify the exact coefficients for each of the surviving features. The purpose of the initial step was purely

to find a fair estimate of the model's out-of-sample performance; now that a fair estimate was discovered, a new model was trained.

It should be explicitly stated that the features presented in this study are not directly evaluated on the held-out data. The link between the two phases is an assumption that, as the interpretation model applies the exact same procedure as the training model and the training dataset almost completely overlaps with the full dataset, it would therefore be reasonable to expect the two models to select similar features. However, it should be noted that this cannot be guaranteed, and the interpretation model's true predictive accuracy remains unverified.

As stated before, fitting $N$ separate models during LOOCV leads to independent feature selections. Due to the significant imbalance between data points and features, there exists a real possibility that a selected feature will vary across folds. A feature appearing in the interpretation model is therefore not sufficient evidence for it being a physiological driver rather than an artifact of noise of the particular sample used to fit that model. To address this, the frequency of a feature appearing in a fold is tracked during LOOCV. This provides a precise metric for the robustness of a feature and follows the general logic of stability selection [14].

It should be noted that the current validation process is likely to produce no results for athletes with insufficient data, as a direct consequence of dividing an already limited training dataset to find an alpha during the cross-validation process. To account for this, athletes with insufficient data were omitted from this additional inner $k$-fold split. Instead, $\alpha$ was selected by using the full training dataset excluding the single held-out data point. However, even under these conditions, it is still likely that regularization shrinks all coefficients to zero for athletes with insufficient data.

### 3.3.3 Swap Randomization

As mentioned before, subgroup discovery evaluates a massive hypothesis space in order to find interesting subgroups; however, this makes the output highly susceptible to the false discovery rate. As random noise in the data might be wrongly attributed to be a subgroup, a method proposed by Arno Knobbe [11] called Swap Randomization is utilized.

It works by randomly sampling the target variable across 100 iterations and then calculating the maximum quality score that can be achieved by pure mathematical chance. After the threshold is calculated, only subgroups that possess a quality score that exceeds this randomized upper bound with a significance level of $\alpha = 0.05$ are considered to be valid. This ensures that any subgroup discovered represents an actual biological driver and not just noise.

### 3.3.4 Evaluation Metrics

**Global Metric ($R^2$):** All global models are evaluated using the Coefficient of Determination ($R^2$), which calculates the proportion of total variance in the target variable (IAAF score) that can be explained by the independent features (e.g., training loads and environmental factors). $R^2$ is defined as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2} \tag{3}$$

where $y_i$ is the actual race performance, $\hat{y}_i$ is the model's predicted performance, and $\overline{y}$ is the historical mean performance of the cohort.

**Local Metric (Z-Score):** Although $R^2$ is useful when determining the performance of global models, it is unsuitable for subgroup discovery. As the primary objective of subgroup discovery is to find specific pockets where performance is unusually high or low, the Z-score is instead used as the primary quality measure.

The Z-score measures the deviation of a subgroup distribution from the athlete's global baseline mean, evaluating if said deviation can be attributed to pure chance. The formal mathematical definition is outlined as:

$$Z = \frac{\overline{Y}_{sg} - \overline{Y}}{SE} = \frac{\overline{Y}_{sg} - \overline{Y}}{\sigma/\sqrt{n_{sg}}} \tag{4}$$

where $\overline{Y}_{sg}$ is the mean IAAF score of the races inside the subgroup, $\overline{Y}$ is the athlete's global mean across all races, $\sigma$ is the standard deviation of all race scores, and $n_{sg}$ is the number of races covered by the subgroup rule.

# 4 Experimental Setup

This section details the practical execution of the methodology outlined above. It introduces the athletes in further detail and the standard used for data collection, then discusses the model hyperparameters and implementations. The goal of this section is to clearly explain the experiment so that it is possible to replicate our methodology and reach similar results.

## 4.1 Data Collection

The data collected from each athlete consisted of their Garmin watch data and race outcomes. To ensure that the athletes used in the study were competitive, strict inclusion criteria were used. These were defined as follows: all athletes must have at least one race performance exceeding 1000 points on the World Athletics scoring table. This ensures that any athlete included in the study is within the top 800 athletes globally; within the cohort, there is an individual with a ranking in the top 200 for their respective discipline, showcasing that the athletes considered are elite on the global stage.

At the time of data collection, the mean age of the cohort was $25.3 \pm 4.5$ years. The cohort consists of male athletes who have a wide range of specializations, from long-distance (marathon to 10,000 m) to middle-distance (1500 m to 5,000 m) running. As the athletes all compete at such a high level, tracking of sessions was consistent. Therefore, any days that lacked a recorded session were assumed to be rest days, under the assumption that any maximum-effort or structured training session would be recorded.

After removing races that the athletes did not finish or that lacked sufficient watch data, the final dataset of valid performances per subject is detailed as:

- **Athlete 01:** 49 valid performances

- **Athlete 02:** 30 valid performances
- **Athlete 03:** 22 valid performances
- **Athlete 04:** 55 valid performances

The weather data was collected using the Open-Meteo free API route. Using the coordinates provided by the watch data, the data was retrieved from the API, recording the environmental features outlined in the methodology. Race observations were acquired using Atletiek.nu, which is the Dutch website that hosts a database of athletic performances.

**Hardware and Data Structure** The specific hardware of each athlete differed; however, all wearable-derived load metrics were captured using consumer-grade Garmin devices. This ensured consistent formatting across all athletes, which allows for the same data processing pipeline across all individuals. The specific device output uses Garmin's standardized Flexible and Interoperable Data Transfer (.FIT) protocol and Session Metadata, which are listed in further detail below:

1. **High-Fidelity data (.FIT Files):** These binary files contain the second-by-second measurements that are required to accurately compute internal load metrics, such as Bannister's TRIMP and zones.
2. **Session Metadata (JSON):** In order to save time and maximize accuracy, session-level JSON summaries were instead used to extract absolute duration, coordinates (latitude/longitude), and session-level heart rate metrics.

## 4.2 Implementation and Hyperparameter Configurations

All data preprocessing, feature engineering, and modeling were written in Python. To ensure the reproducibility of the methodology, this section details the specific libraries and hyperparameters utilized for each of the models.

### 4.2.1 Linear Mixed-Effects Regression

The global mixed-effects regression architecture was implemented using the `mixedlm` module from the Python `statsmodels` library. As mentioned in the methodology, directly inputting raw physiological metrics into a pooled model could lead to gradient descent optimization failure. Therefore, Group-Mean Centering and Z-score standardization were personally implemented and acquired from the Python `scikit-learn` library, respectively.

In addition to this, the `statsmodels` implementation was configured to use the Conjugate Gradient (`cg`) optimizer during the fitting process, as standard optimization algorithms suffer from convergence instability when calculating the random intercept for athletes with sparse race observations.

### 4.2.2 Regularized Linear Regression

The Lasso regression model was implemented using the `LassoCV` estimator from the `scikit-learn` library, following the nested cross-validation procedure detailed in Section 3.3.2. The maximum iterations and convergence tolerance were increased due to the model struggling to converge for

specific athletes. The maximum iterations parameter was set to 10,000, and the convergence tolerance was set to $1 \times 10^{-3}$.

#### 4.2.3 Subgroup Discovery

Subgroup discovery was implemented using `pysubdisc`, a Python wrapper for the Cortana Java-based SubDisc software. The algorithm was applied to all 4 athletes' datasets. It should be noted that the values were not scaled, as it is both unnecessary for the functionality of subgroup discovery and makes the output difficult to interpret.

The target variable was defined as a continuous numeric target, with the specific hyperparameters being defined as:

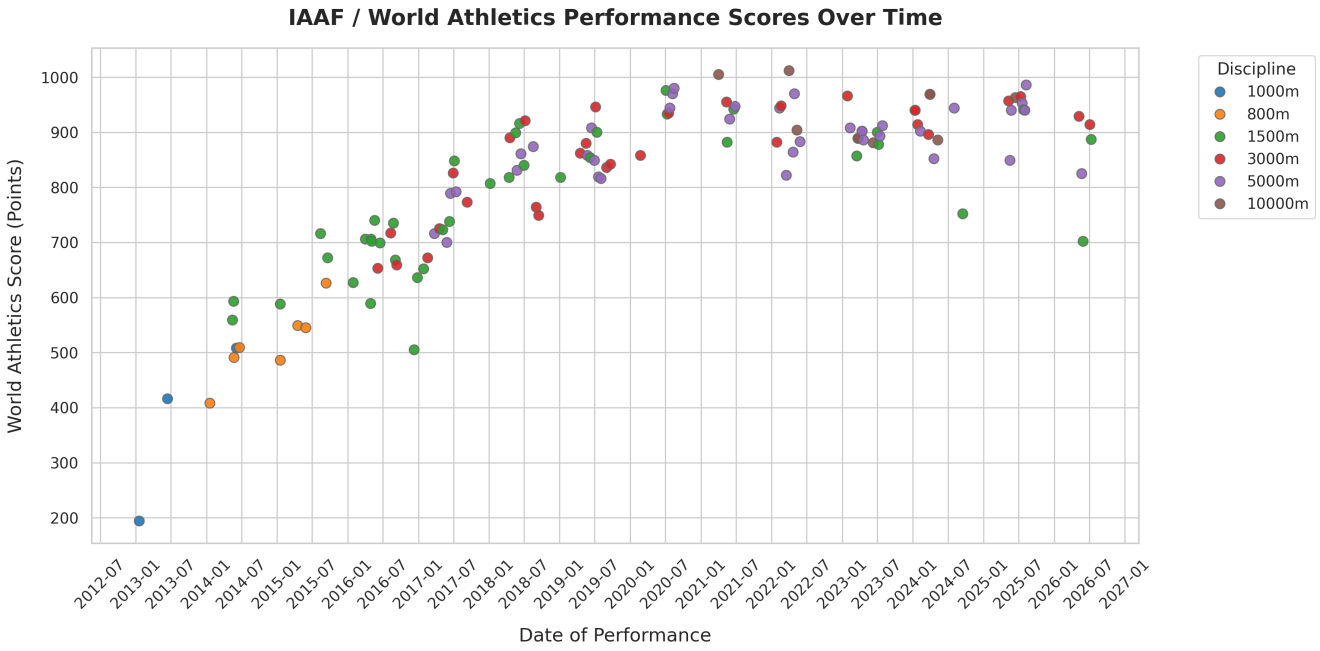
- **Quality Measure:** "Z_SCORE" to identify subgroups that deviated significantly from the athlete's baseline mean.
- **Numeric Strategy:** NUMERIC_BEST.
- **Search Depth:** Only depths 1 and 2 were considered, as increasing the depth directly reduces the interpretability of the output. The depth controls how many conditions can be combined into a single rule.
- **Support Constraint:** $\max(5,\ 0.10 \times N)$, ensuring any reported rule covers at least 10% of the athlete's races and preventing rules built on just one or two outlier events.
- **Statistical Validation:** Swap randomization with 100 permutations at $\alpha = 0.05$. Both positive and negative deviations were searched; however, only positively deviating subgroups are reported in the results.

# 5 Results

This chapter presents the results of each of the models and a comparison of their performance. It should be noted that not all athlete results will be shown for the sake of brevity. Instead, only athletes with interesting and significant results will be selected for further investigation. The two athletes with the most data, Athlete 01 ($n = 49$) and Athlete 04 ($n = 55$), will likely be the main focus due to having the most data points.

## 5.1 Dataset Variance and Naive Baseline

Before further investigation using complex models, it is critical to first understand the inherent variance of an athlete's performance across their career. As shown in the figure below, an athlete's career does not progress in a strictly linear fashion; instead, it grows rapidly during initial stages before plateauing as they mature and asymptote towards their physiological limit.




As shown in the diagram, as the athlete progressed through their career, there is an inherent variance in both the discipline the athlete competes in and the score of those performances.

Table 1: Performance of the Naive Baseline Estimator

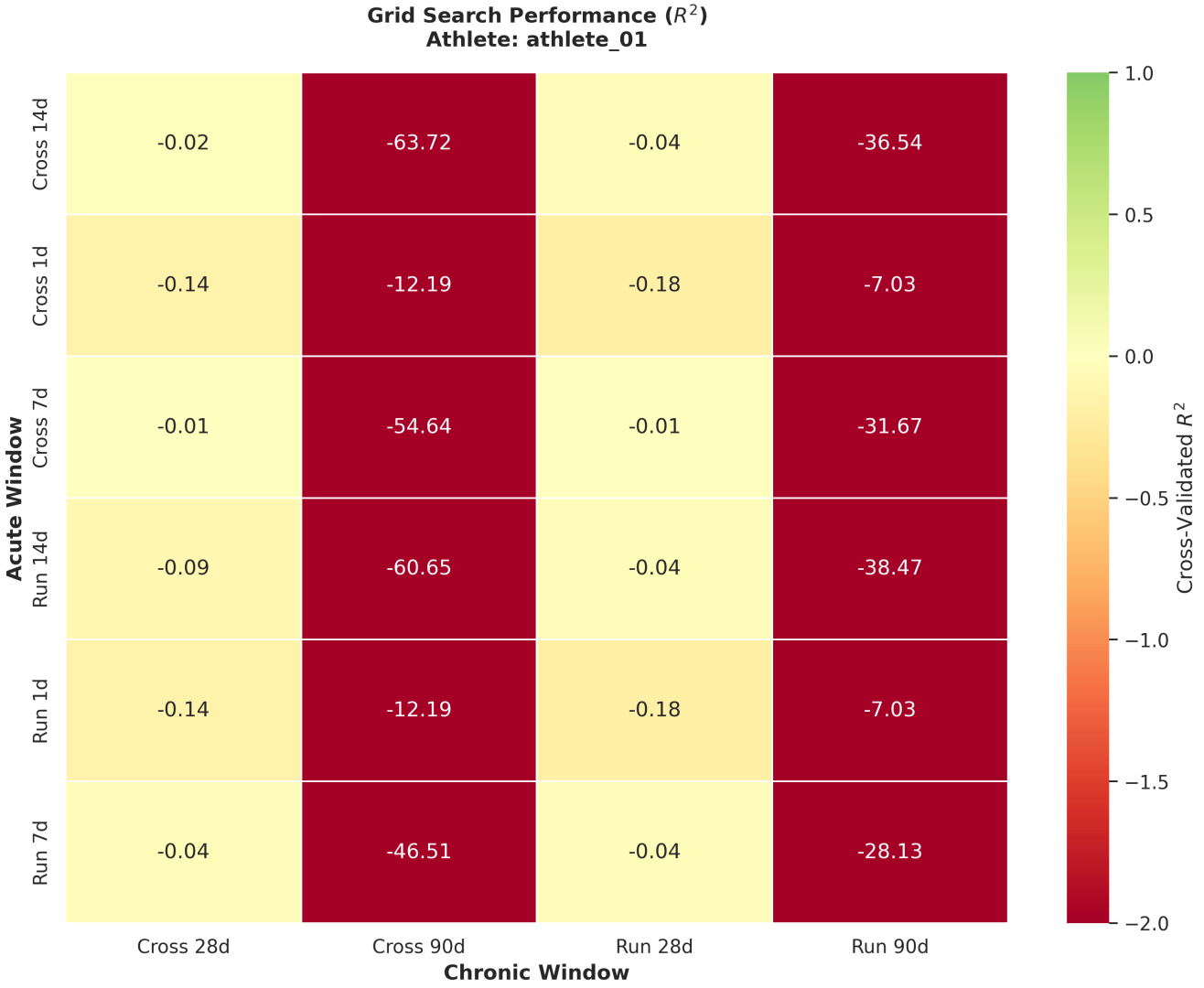
| **Subject** | **Baseline** $R^2$ |
|---|---|
| Athlete 01 | -1.3467 |
| Athlete 02 | -0.2838 |
| Athlete 03 | -0.3728 |
| Athlete 04 | -7.1270 |

The severely negative $R^2$ values—most notably for Athlete 04 ($R^2 = -7.127$) and Athlete 01 ($R^2 = -1.347$)—indicate that using an athlete's past three races to predict a performance is significantly worse than simply guessing the global mean. A possible explantion for the models extremely weak performance for Athlete 04 is that the models athletes performance swings drasticly within a short time frame, which is clearly illustrated by their career timeline. On the other hand, the two athletes with the greatest performance both lack sufficient data for the variance to impact the naive baseline's prediction.

## 5.2 Linear Mixed-Effects

The main objective of using the LMM was to determine if historical internal training loads could accurately predict performance better than the naive moving average, with the secondary objective of identifying the optimal acute and chronic windows for each athlete that lead to the greatest performance.

The results of LMM grid search differ greatly for the global population and the individual athletes. When modeling the athletes as a single pooled dataset, the optimal windows were discovered to be 14 days for the acute window and 28 days for the chronic window. This global configuration had a macro-averaged $R^2$ of -0.0955. On the other hand, predicting individual athletes also proved to be a significant challenge, as all athletes produced a negative $R^2$ value.




To further illustrate the sensitivity of the model to different temporal training blocks, Figure 2 visualizes the grid search landscape for Athlete 01. This athlete was selected because their performance matrix best shows the separation between optimal and suboptimal training windows. As observed in the heatmap, models trained on 90-day chronic windows resulted in a severe degradation of performance, with the value dropping below -20 in extreme cases. Note that values displayed in the heatmap are rounded to two significant figures for readability.

As athletes have different biological capacities to fatigue and recovery, forcing a singular temporal

Table 2: Optimal Temporal TRIMP Windows and Maximum Cross-Validated $R^2$ per Subject

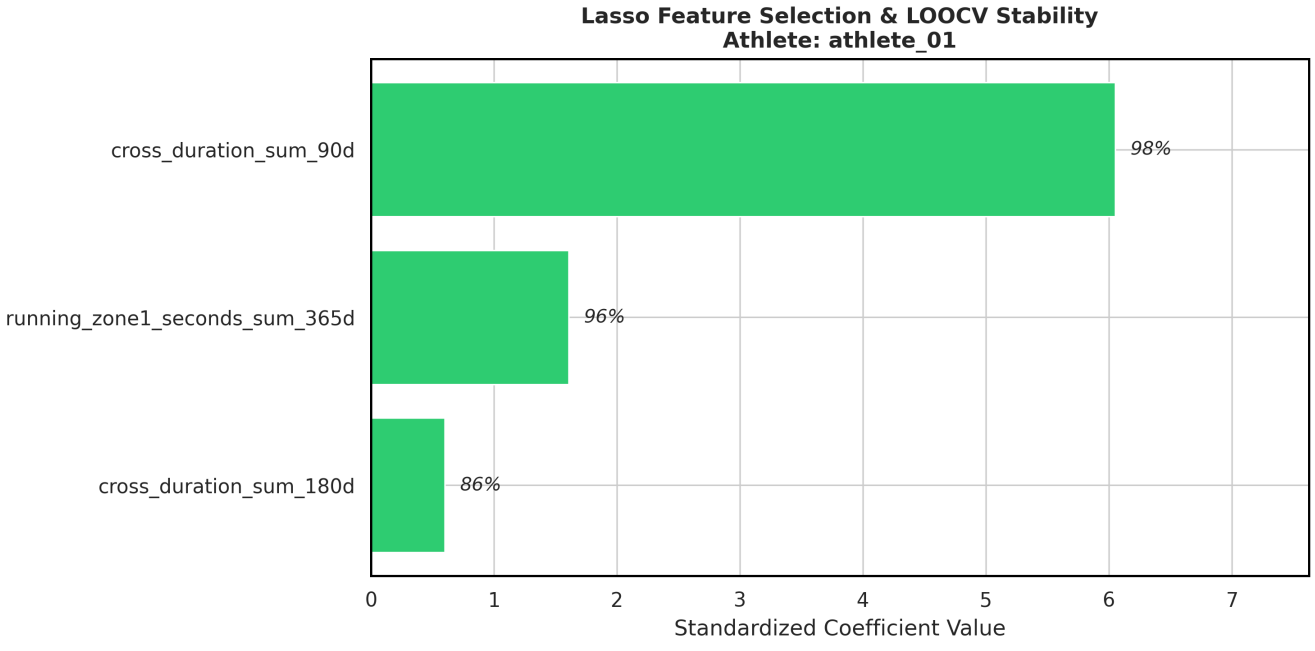
| **Subject** | **Optimal Acute Window** | **Optimal Chronic Window** | **Max CV $R^2$** |
|---|---|---|---|
| Athlete 01 | 7-Day Cross-Training TRIMP | 28-Day Cross-Training TRIMP | -0.0018 |
| Athlete 02 | 14-Day Running TRIMP | 28-Day Cross-Training TRIMP | -0.1129 |
| Athlete 03 | 14-Day Running TRIMP | 28-Day Running TRIMP | -0.1397 |
| Athlete 04 | 14-Day Running TRIMP | 28-Day Cross-Training TRIMP | -0.0781 |

lag on all athletes would lead to suboptimal results for the individual athletes. The table above clearly illustrates this, as not all athletes have the same optimal lag.

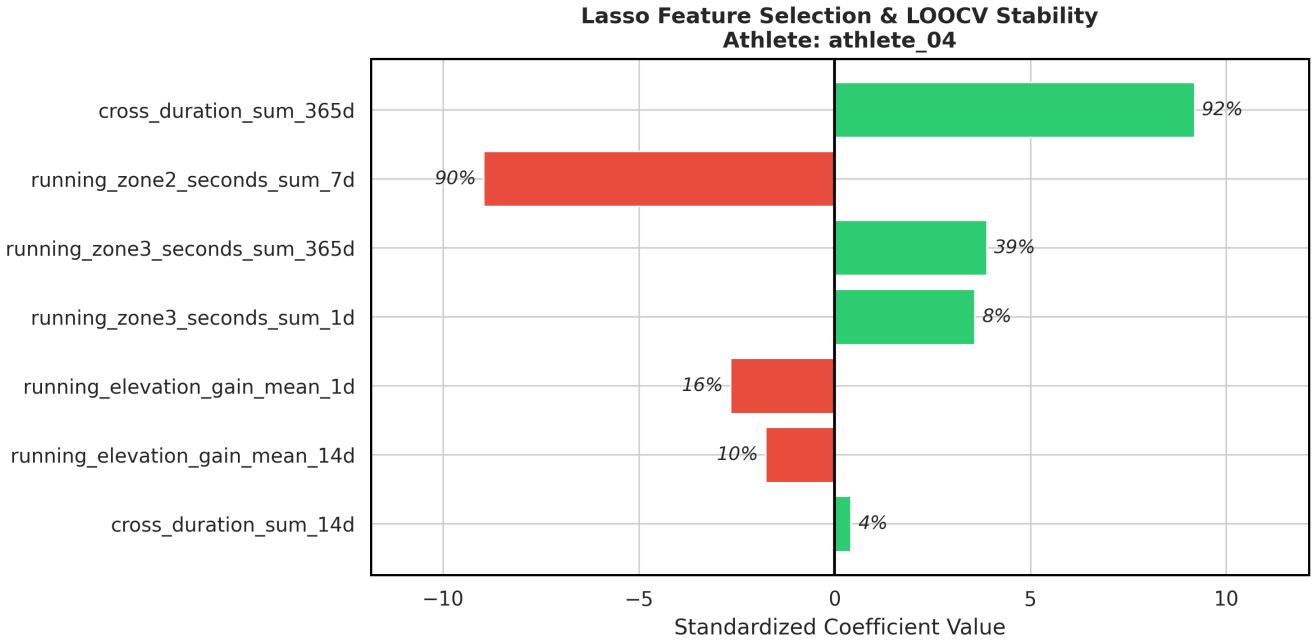
While the 14-day acute and 28-day chronic split is a clear winner, Athlete 01 has a shorter acute load at just 7 days. This confirms the need for an individualized approach, as athletes clearly have different responses to load depending on the time lag from a performance.

## 5.3 Lasso Regression

When applying the Lasso model on Athlete 02 and Athlete 03, the model failed to converge; therefore, their results will not be presented in this section. This failure was due to the model not finding any features. Instead, this section will focus on the valid results of Athlete 01 ($n = 49$) and Athlete 04 ($n = 55$). Athlete 01 achieved an out-of-sample LOOCV $R^2$ of 0.0553 and Athlete 04 produced a lower cross-validated $R^2$ of -0.0827. The performance of the models improved slightly for Athlete 01 for the Lasso model and actually decreased for Athlete 04. It should be noted that the number of features digested by Lasso was 116 for both athletes, with the final selected features detailed below.



The figure above visually represents the features that were extracted as the main drivers of performance for Athlete 01. While these long-term training variables demonstrated high stability, their practical predictive capacity is minimal. This is evident by the model's overall low performance of just 0.0553. Thus, these features should be interpreted as statistically persistent features rather than actual strong drivers of performance.




While Athlete 01's performance was heavily dictated by chronic volume, Athlete 04 exhibits a slightly different physiological profile, as shown in the figure above. The coefficients identified show a clear relationship between tapering and long-term training. This is shown by the LOOCV stability percentage being greatest for chronic cross-training volume and 7-day Zone 2 running. To clarify, the athlete did not have any lactate test data; therefore, Zone 2 corresponds to HRR between 80 and 90 percent.

Ultimately, the Lasso model struggled to capture the linear biological drivers of performance, showing a generally lower performance for the two successful results. This is to be expected, as standard linear coefficients assume that adjusting these values would lead to an increase or decrease of performance by a fixed amount; this is, of course, not the case. There are limits to these rules imposed by human biology; therefore, to investigate the failure points, subgroup discovery was critical.

## 5.4 Subgroup Discovery

This section presents the results of subgroup discovery for all athletes. To ensure sufficient context when interpreting the results, the significance thresholds and resulting subgroups are presented sequentially by search depth. It should be noted that not all athletes produced statistically

significant subgroups; thus, their results are omitted from the final tables. Additionally of the subgroup discovered only the top 5 that do not repeat the same conditions will be presented.

- **Athlete 01:** 1033
- **Athlete 02:** 968
- **Athlete 03:** 968
- **Athlete 04:** 903

To properly evaluate the significance of a subgroup, it is important to understand the athletes' baseline career averages. By comparing their average baseline to the average of a subgroup, it helps identify the magnitude of the conditions impact on performance.

### Depth 1 Results

Table 3: Significance Thresholds per Athlete (Depth 1)

| Athlete | 5% Threshold ($Z$) | 1% Threshold ($Z$) | Number of Subgroups |
|---|---|---|---|
| Athlete 01 | 2.92 | 3.12 | 1 |
| Athlete 02 | 2.86 | 3.24 | 0 |
| Athlete 03 | 2.86 | 3.05 | 0 |
| Athlete 04 | 2.91 | 3.15 | 3 |

Table 3 details the specific 5% and 1% significance thresholds ($Z$) for each athlete at a search depth of 1.

Table 4: All Significant Subgroups for Athletes 01 and 04

| Athlete / Rank | Cov. | Qual. | Avg. | St. Dev. | Condition |
|---|---|---|---|---|---|
| **Athlete 01** | | | | | |
| 1 | 9 | 2.99 | 1080 | 30.26 | `discipline = '10000m'` |
| **Athlete 04** | | | | | |
| 1 | 8 | 3.20 | 969 | 22.97 | `cross_duration_hours_180d` $\geq$ 218.17 |
| 2 | 14 | 3.12 | 952 | 29.44 | `running_duration_hours_365d` $\geq$ 416.44 |
| 3 | 13 | 3.09 | 953 | 30.70 | `cross_duration_hours_365d` $\geq$ 430.87 |

Table 4 outlines the all subgroups discovered at Depth 1 for Athletes 01 and 04. As for Athlete 04, surprisingly one of subgroup exceeded the 1% threshold of 3.15, namely total training in all modalities being greater the 218 hours within 180 days of race. With further inspection of the subgroup, the relationship of massive chronic training volume and high performance is true irrespective of the athletes maturity, with improvements manifesting as early as 2021 and as late as 2025.

**Depth 2 Results**

Table 5: Significance Thresholds per Athlete (Depth 2)

| Athlete | 5% Threshold ($Z$) | 1% Threshold ($Z$) | Number of Subgroups |
|---|---|---|---|
| Athlete 01 | 3.63 | 3.73 | 37 |
| Athlete 02 | 3.58 | 3.72 | 0 |
| Athlete 03 | 3.16 | 3.25 | 0 |
| Athlete 04 | 3.69 | 3.79 | 40 |

Moving to a deeper search space, Table 5 shows the updated thresholds and subgroup counts for Depth 2. As expected, the significance thresholds naturally increased to account for the larger search space.

Table 6: Top Non-Repetitive Depth-2 Subgroups for Athletes 01 and 04

| Athlete / Rank | Cov. | Qual. | Avg. | St. Dev. | Conditions |
|---|---|---|---|---|---|
| **Athlete 01** | | | | | |
| 1 | 13 | 3.85 | 1079 | 25.33 | `cross_duration_hours_365d` $\geq$ 558.43 & `cross_acwr` $\leq$ 0.93 |
| 2 | 14 | 3.83 | 1077 | 24.30 | `running_duration_hours_1d` $\leq$ 0.77 & `cross_duration_hours_90d` $\geq$ 132.91 |
| 3 | 25 | 3.76 | 1060 | 38.57 | `wind_speed_kmh_14d` $\geq$ 10.17 & `cross_acwr` $\leq$ 0.93 |
| 4 | 20 | 3.68 | 1064 | 31.89 | `running_zone1_hours_28d` $\geq$ 13.41 & `cross_banister_trimp_7d` $\leq$ 2811.90 |
| 5 | 18 | 3.67 | 1067 | 31.49 | `running_max_hr_60d` $\leq$ 201.0 & `running_duration_hours_1d` $\leq$ 0.82 |
| **Athlete 04** | | | | | |
| 1 | 18 | 4.05 | 959 | 19.87 | `running_duration_hours_180d` $\geq$ 203.36 & `running_zone2_hours_7d` $\leq$ 3.11 |
| 2 | 21 | 3.96 | 953 | 26.02 | `running_avg_hr_365d` $\leq$ 131.30 & `running_elevation_gain_1d` $\leq$ 18.0 |
| 3 | 27 | 3.92 | 947 | 37.34 | `running_zone3_hours_365d` $\geq$ 29.89 & `running_elevation_gain_1d` $\leq$ 64.10 |
| 4 | 15 | 3.90 | 962 | 24.27 | `running_zone3_hours_365d` $\geq$ 29.89 & `wind_speed_kmh_14d` $\leq$ 13.46 |
| 5 | 21 | 3.79 | 951 | 28.07 | `running_duration_hours_180d` $\geq$ 203.36 & `humidity_pct_1d` $\geq$ 51.55 |

For the sake of brevity, only the top five non-repetitive results are presented in Table 6. both Athlete 01 and 04 produced subgroups that safely exceed the 1% significance threshold. From the presented conditions, Athlete 01 has 3 subgroups exceeding the 1% threshold and all of Athlete 04's subgroups either match or exceed the threshold.

### 5.5 Comparison

Table 7: Cross-validated $R^2$ for all global models per athlete.

| **Athlete** | $n$ | **Naive Baseline** | **LMM** | **Lasso** |
|---|---|---|---|---|
| Athlete 01 | 49 | -1.3467 | -0.0018 | **0.0553** |
| Athlete 02 | 30 | -0.2838 | **-0.1129** | — |
| Athlete 03 | 22 | -0.3728 | **-0.1397** | — |
| Athlete 04 | 55 | -7.1270 | **-0.0781** | -0.0827 |

The table above summarizes the performance of each of the global models. Lasso results are only reported if any stable features were discovered; otherwise, they are indicated by dashes. The pooled LMM $R^2$ macro average was -0.0955. The highest score for each athlete is presented in bold.

## 6 Discussion

This section will discuss and contextualize the results and answer the research questions of the study. Additionally, it covers the key limitations that potentially caused the poor performance of the global models.

### 6.1 Interpretation

In this section, we revisit the four research questions driving this study and answer them based on the results of each of the models.

**RQ1: Does historical wearable data combined with environmental data carry statistically significant predictive power for elite race performance?** Looking at the results, it becomes evident that the predictive power of the selected models is limited. Both the pooled and individualized global models produced negative $R^2$ values; this performance is worse than if the model made a trivial mean-based prediction. Although the models did outperform the naive baseline for all athletes, this could be due to the naive baseline being a poor model rather than evidence of predictive value. One standout result was Athlete 01, who achieved an out-of-sample $R^2$ (0.0553) for the Lasso model. The model selected three highly stable features (86–98% across LOOCV folds) for cross-training and Zone 1 volume features. This result provides direct but limited evidence for the predictive capabilities of historical training data. The answer to RQ1 is therefore that predictive power does exist, but it is small in magnitude and dependent on the athlete and sample size.

**RQ2: Can a Mixed-Effects regression architecture effectively model physiological load and preserve individual baselines for athletes with sparse performance data?** The LMM model was able, in some capacity, to preserve the individual baseline for each athlete, as the model was able to significantly reduce the overall error for each of the athletes in comparison to the naive model. The grid search converged to a population optimum of 14 days for the acute load and 28 days for the chronic, suggesting that the temporal representation for acute and chronic load could exist across the cohort. However, the model has a glaring shortcoming: its inability to predict results greater than the naive mean-based guess, as evidenced by all athletes producing a negative $R^2$. This definitively proves that the current model essentially explains none of the variance in the athlete data. The answer to RQ2 is largely negative, as while the model has been proven in other studies to be well suited for sparse athletic data, it failed in the case of this cohort.

**RQ3: Which specific temporal windows (lags) and engineered features are the most critical drivers of race-day success?** The answer to this question varies per athlete and is not clear based on the negative results of the Lasso model. For Athlete 01, long-term cross-training duration and Zone 1 volume proved to be the most dominant, stable features. This indicates that the performance of Athlete 01 is partly dictated by their long-term preparation, as around 5% of the variance of the final race outcome was attributed to the predicted features. This finding is of course supported by the literature and could be considered common knowledge, as training more for a long period of time would be expected to improve performance, making the utility of this insight questionable. As for Athlete 04, it would be misleading to present the features selected by Lasso as anything other than noise, as the model produced a negative $R^2$ value.

Additionally, the results of the matrix search cannot be used as evidence due to suffering from a similar issue of a negative $R^2$. In conclusion, there is no single temporal window or feature that was generalizable across the cohort; however, long-term chronic training volume consistently outperformed acute load as a predictor of performance.

**RQ4: What localized, non-linear physiological thresholds exist within the data, and how can they be translated into actionable training constraints for coaches?** Subgroup discovery produced the clearest answer and presents the strongest empirical results. In contrast to the global models, SD produced a significant number of validated results for Athletes 01 and 04. The significance of the subgroups was especially high for depth 2, with Athlete 01 and 04 both producing a substantial number of subgroups exceeding the 1% threshold. To keep this section as verbose and clear as possible, only the top subgroups at each depth will be discussed.

Athlete 01's strongest condition was that if the athlete completed more than 558 hours of total training in the year and reduced their average weekly load to 93 percent of their monthly average, their IAAF score would increase by 46 points. This finding is especially useful to coaches, as it gives a clear value for a tapering protocol; given that the athlete maintains consistent training, they can maximize their performance by reducing load in the week leading up to a race. While Depth 1 did produce a subgroup, it is somewhat trivial as it simply captured the athlete's specialization being 10000m.

The same pattern can be observed in Athlete 04; although it has manifested differently, the same relationship of consistent training alongside sufficient tapering can be seen. This is illustrated

by the most significant subgroup of Athlete 04, where if Athlete 04 trained more than 203 hours across all modalities within 180 days of the race and less than 3 hours of Zone 2 a week before competition, performance jumped 56 points. For clarification, Zone 2 corresponds to 70% to 80% of the athlete's HRR.

As for Athlete 01's results for Depth 1, all the subgroups sit comfortably above the thresholds, with the greatest performing subgroup exceeding the 1 percent threshold. The subgroups paint a consistent pattern of chronic training being a critical part of the athlete's preparation for them to excel. This, of course, could be swept away as common sense, but subgroup discovery's ability to provide specific thresholds is especially useful. It shows clearly that the athlete has more room to grow their total volume and that 218 hours of training should be the floor when planning for a 6-month build-up to a competition.

RQ4 was only successfully answered for the two athletes with the largest datasets, reinforcing that individualized models require sufficient data to function. To succinctly answer RQ4, finding an actionable threshold is possible, but requires that athletes have enough data.

In summary, the four research questions point to a consistent pattern: predicting elite athletic performance is possible with wearable and environmental data, but it is neither large nor accessible to all athletes. Additionally, using LMM to overcome sparsity failed due to its inability to find a cohort-wide relationship and individual linear patterns in training load. Both feature importance and subgroup discovery were successful for the two athletes with the greatest number of race observations. This reinforces the existence of a data-sufficiency boundary at $n \approx 30$ observations across the two unrelated modeling techniques.

## 6.2 Limitations

The critical failure of the global models is likely due to a number of factors. One of the main issues was data scarcity, which makes it difficult to model global patterns from the data due to the extreme risk of overfitting. This likely caused the low performance across the models.

Beyond this, our study fails to capture all known factors that affect performance. The current features exclusively focus on cardiovascular and environmental factors, completely ignoring critical variables such as recovery and stress, which have repeatedly been shown to play a role in performance. Additionally, watch data can only capture training that is directly coupled to heart rate, making it difficult to capture anaerobic mechanical work such as resistance training or short, high-intensity sprints. These exercises cause significant neuromuscular fatigue but fail to increase heart rate and are thus excluded.

Another significant challenge is the quality of the data and the reliability of the measuring apparatus. As this study heavily relies on commercial wearable sensors and does not enforce a standard model, there is a risk of introducing systematic bias due to differing sensor accuracies, potentially skewing the internal load calculations across the cohort.

Finally, the weak performance of the LMM model could be primarily attributed to three limitations. The first and most glaring issue is the size of the cohort, with the pooled model consisting of only 4 athletes with a total of 156 performances. The second limitation is that within the cohort, athletes compete in various events that demand different physiological adaptations. This limits the

utility of the global effect, as athletes with the greatest amount of data were long-distance runners and those with few data points were middle-distance (1500 m–5 k) specialists, likely reducing the utility of pooling athlete data. The final limitation was the absence of standardized lactate step-test data for the entire cohort, making it impossible to calculate the individualized TRIMP (iTRIMP) coefficients as done in Avalos's study [1]. The standard TRIMP score likely overestimates the physiological strain for the elite athletes, which could potentially introduce inaccurate load representation for each of the athletes. The reason that standard TRIMP overestimates strain is that the coefficients used are derived from the general population, who experience significantly more strain than an elite athlete at the same percentage of HRR. Thus, these inaccuracies could possibly fail to capture the exponential variance between general load and the true load experienced by the athlete. While this limitation is not directly supported by experimental evidence in this study, the success of Avalos's method [1] illustrates that using iTRIMP is a possible improvement to the model's final performance.

# 7 Conclusions and Further Research

The primary objective of this thesis was to determine the capability of consumer-grade wearable trackers in predicting race outcomes for competitive runners. The findings hints towards a personalized, non-linear approach when predicting highly sparse race outcomes. The global continuous models (Linear Mixed-Effects and Lasso Regression) largely failed to produce substantial results across the cohort, yielding an average $R^2$ around or below 0. This underscores the difficulty of finding global relationships within such a sparse dataset and the possibility that the true nature of the relationship being non-linear and would therefore require a different approach. Additionally, it shows that pooling athletes under the umbrella of endurance running, without further separation into long-distance and middle-distance, will produce no improvements when attempting to artificially increase the dataset, as athletes' physiological differences are too great to find global patterns that are true for the entire cohort. Ultimately, the identified data threshold for an individualized approach ($n \approx 30$ observations) dictates that sports analytics must prioritize individualized, threshold-based frameworks to safely and effectively alter elite training cycles.

The limitations encountered in this study provide a clear roadmap for future research to build upon. Future studies utilizing the LMM model could investigate a homogeneous cohort that only contains athletes who compete in similar brackets of endurance running—for example, only middle-distance (800 m–5 k) or only long-distance (10 k–marathon). This would ensure that global effects could capture more relevant and generalizable physiological truths during modeling. Finally, attempting to directly map daily training to infrequent maximal race efforts remains a statistical challenge due to the lack of data points. Instead of predicting the athletes' final performance, future research should focus on predicting rich, daily-recorded target variables such as resting heart rate or heart rate variability. These metrics would still be useful to coaches and athletes, as they provide insights into when an athlete is over- or under-training and ways they could adjust their training accordingly.

# References


[1] Marta Avalos, Philippe Hellard, and Jean-Claude Chatard. Modeling the training-performance relationship using a mixed model in elite swimmers. *Medicine & Science in Sports & Exercise*, 35(5):838–846, 2003.

[2] V. L. Billat and J. P. Koralsztein. Significance of the velocity at $\dot{V}O_{2\,max}$ and time to exhaustion at this velocity. *Sports Medicine*, 22(2):90–108, 1996.

[3] Jill Borresen and Michael I. Lambert. The quantification of training load, the training response and the effect on performance. *Sports Medicine*, 39:779–795, 2009.

[4] Nathan Broughton. Exploring the relationship between behaviour and health: A comparative analyses of time series prediction models and feature explainability techniques. Master's thesis, Leiden University, Leiden, The Netherlands, November 2023.

[5] TW Calvert, EW Banister, MV Savage, and T Bach. A systems model of the effects of training on physical performance. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(2):94–102, 1976.

[6] A. Casado, F. González-Mohíno, J. M. González-Ravé, and C. Foster. Training periodization, methods, intensity distribution, and volume in highly trained and elite distance runners: A systematic review. *International Journal of Sports Physiology and Performance*, 17(6):820–833, 2022.

[7] Víctor Cerezuela-Espejo, Javier Courel-Ibáñez, Ricardo Morán-Navarro, Alejandro Martínez-Cava, and Jesús G. Pallarés. The relationship between lactate and ventilatory thresholds in runners: Validity and reliability of exercise test performance parameters. *Frontiers in Physiology*, Volume 9 - 2018, 2018.

[8] Matthew R. Ely, Samuel N. Cheuvront, William O. Roberts, and Scott J. Montain. Impact of weather on marathon-running performance. *Medicine & Science in Sports & Exercise*, 39(3):487–493, 2007.

[9] Franco M Impellizzeri, Samuele M Marcora, and Aaron J Coutts. Internal and external training load: 15 years on. *International Journal of Sports Physiology and Performance*, 14(2):270–273, 2019.

[10] M. J. Joyner and E. F. Coyle. Endurance exercise performance: the physiology of champions. *The Journal of Physiology*, 586(1):35–44, 2008.

[11] Arno Knobbe, Jac Orie, Nico Hofman, Benjamin van der Burgh, and Ricardo Cachucho. Sports analytics for professional speed skating. *Data Mining and Knowledge Discovery*, 31:1872–1902, 2017.

[12] Benjamin D Levine and James Stray-Gundersen. "living high-training low": effect of moderate-altitude acclimatization with low-altitude training on performance. *Journal of Applied Physiology*, 83(1):102–112, 1997.

[13] Daniel Maupin, Ben Schram, Elisa Canetti, and Robin Orr. The relationship between acute: chronic workload ratios and injury risk in sports: a systematic review. *Open access journal of sports medicine*, 11:51–75, 2020.

[14] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[15] Stephen Seiler. What is best practice for training intensity and duration distribution in endurance athletes? *International journal of sports physiology and performance*, 5:276–91, 09 2010.

[16] Bojidar Spiriev. *World Athletics Scoring Tables of Athletics*. World Athletics, 2022.

[17] P. Szot. Evolution of sport wearable global navigation satellite systems' receivers: A look at the garmin forerunner series. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, 240(1), 2024.

[18] Hirofumi Tanaka, Kevin D Monahan, and Douglas R Seals. Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology*, 37(1):153–156, 1 2001.