Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Statistics and Computer Science

Machine Learning to go *nyoom*

Using Machine Learning to evaluate rowing training and predict training outcomes and performances

Liam Junkermann

Supervisor: Dr. Lucy Hederman

April 2024

A Final Year Project submitted in partial fulfilment
of the requirements for the degree of
B.A.(Mod.) in Computer Science

Declaration

I hereby declare that this Final Year Project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

I consent / do not consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Liam Junkermann
April 15, 2024

Machine Learning to go *nyoom*

Using Machine Learning to evaluate rowing training and predict training outcomes and performances

Liam Junkermann, B.A.(Mod.) in Computer Science

University of Dublin, Trinity College, 2024

Supervisor: Dr. Lucy Hederman

Abstract

Modelling human performance has been a challenging task for many years. The complexity of human physiology and the variety of factors that can influence performance make it difficult to develop accurate models. Machine learning, specifically the use of neural networks, has proven particularly promising in applications to predict human performance. This project explores the approaches to data collection, analysis, and machine learning to predict rowing training outcomes and performances. The project aims to use machine learning techniques to predict performance using rowing data and explore the use of high-quality data to predict athlete injury.

A web application was developed to collect data from rowers as part of the project. This application was used to collect data on training sessions by allowing participants to connect data sources already used to track training with the project. The data collected was then analysed to develop features for a model, and visualisations were produced as feedback for participants. Unfortunately, the data collected was insufficient to develop a model to predict performance. Instead, the project explored the use of machine learning to predict injuries in runners, noting the approach and its potential application to rowing data.

Finally, this project suggests a protocol by which an improved data collection and analysis process can be implemented to train and produce effective machine learning models. These can be personalised and applied to understand individual athlete and crew training, injury, and illness behaviours. As a result of a more involved collection approach, deeper analysis can be completed to provide valuable feedback in the absence of individual performance models. This improved data collection and analysis will provide more potential features to train a variety of models, beginning with the injury prediction and performance models suggested in this project. Each of these suggestions can be used to develop stronger athlete and crew performances by reducing injury and illness rates and adapting training to an individual's unique training response. Furthermore, this approach can deliver a more robust analysis of training sessions and athletes' qualitative and quantitative responses, leading to even greater success. This can also be adapted and applied to other sports, particularly endurance sports.

Keywords: Data Analysis, Data Visualisation, Machine Learning, Personalisation, Human-Computer Interaction, Human Performance

Acknowledgements

I would like to start by thanking all my friends and family for their support throughout this project. In particular, Mama and Papa, thank you for your continued love and support, especially through my educational career. You have always supported whatever wild ideas and curiosities I've chosen to pursue, so it is only fitting that I am finishing my Computer Science degree by talking more about rowing.

A massive thank you is also due to the group of boys I have been rowing with for the last four years of college. I owe so much of the success of the last four years to you all, you really have helped deepen my love for our shared sport. Especially, Tom Stevens for enthusiastically enduring three years of "look at this cool graph", don't worry, many more cool graphs are coming. Thank you as well, to John Harman, for enduring my endless questions for the last four years, and your continued guidance and support as I continue to try and balance rowing and life.

I would also like to thank the participants who took part in the study. Without your data, this project would not have been possible.

Finally, I would like to thank my supervisor, Dr Lucy Hederman, for your support and guidance throughout this project. Without your guidance, I undoubtedly would have lost the run of myself and gotten too wrapped up in getting excited about sports science instead of completing this project. So thank you for keeping me on track.

Contents

List of Figures

List of Tables

Vocabulary

Heart Rate Metrics	
HRV	Heart Rate Variability
HR	Heart Rate
HR_{max}	Maximum Heart Rate
Training Zones	
UT2	Basic Oxygen Utilisation Training
UT1	Oxygen Utilization Training
AT	Anaerobic Threshold Training
TR	Oxygen Transport Training
AN	Anaerobic Capacity Training
LT	Lactate Threshold
General Training Terms	
Stroke Rate	The number of strokes completed per minute (like cadence in cycling or running)
Split	The standard split in rowing is time to complete 500m
ACWR	Acute Chronic Workload Ratio
TRIMP	Training Impulse
Data Providers	
Strava	An exercise tracking platform with social media features commonly used by endurance athletes.
Concept2	The manufacturer of the most commonly used rowing machines. They provide an API for data access for sessions completed on the machines.
Polar	A manufacturer of heart rate monitors, GPS watches, and other fitness tracking devices.
Garmin	A manufacturer of heart rate monitors, GPS watches, and other fitness tracking devices.

Chapter 1

Introduction

This project explores the use of Machine Learning to model and predict rowing training outcomes and performances. Developing models for human performance is a complex and challenging task. Many factors can influence the outcome of a training session or block, and a variety of factors can also influence performance when it matters. As early as 1975, linear models for performance were explored with reasonable success. Since then, many models have been developed to predict performance, with the most recent models implementing machine learning techniques and algorithms for this purpose. This project aims to use some of these techniques which have been developed to predict performance using rowing data. This chapter will introduce the general approach and motivation for this project, the expected outcomes, and a structure for the remainder of the report.

1.1 Motivations

The inspiration for this project is fuelled by a personal passion for rowing and performance alongside an academic interest in data analysis and machine learning. As a competitive rower for the last 12 years, I have produced a huge amount of data. In the last three years, I have been tracking my training and recovery data continuously with several wearables. At the moment, I wear at least three heart rate monitors at all times. These devices capture much of the same data but provide it to different platforms, such as Whoop and Oura, which each have its own hardware. These different platforms provide me with different feedback, of varying value, on training, recovery, and readiness to train more. I have always been interested in using the data I produce to improve my training and recovery habits to perform optimally and avoid illness and injury – the largest deterrents to my athletic growth in recent years.

Despite the amount of data I have been producing, I have yet to consistently analyse it to provide feedback on my training. I am motivated by the prospect of developing a more

effective approach to feedback, one that capitalizes on the wealth of data I have accumulated. By doing so, I hope to not only improve my performance but also contribute to the broader understanding of how data-driven approaches can be utilized in training prescription and personalisation.

This project was an opportunity to blend my love for rowing and sports science with my academic interest in data analysis and machine learning. Ultimately aiming to produce tangible improvements in training and performance, through the power of data analysis and machine learning.

1.2 Goals

There are several steps to this project, with the main goal being to develop a model which can predict rowing performance. This will be done by collecting data from rowers, performing initial analysis on this data to help develop features for a model, and then producing visualisations as feedback for participants. Finally, the ultimate goal of this project is to use this data to train a model which can predict performance, ideally predicting test scores, such as a 2km test on the rowing machine.

In developing a data collection method, the goal is to collect data from rowers in a way that is minimally invasive and requires almost no extra effort from participants (beyond some initial setup). It is notoriously difficult to collect data from athletes, especially those working or in college. As a result of the time-intensive nature of the sport, athletes want solutions which require minimal setup, time and effort. The goal is to develop a system which can collect data from participants, and provide feedback on their training data easily.

When providing training feedback, the analysis must be relevant to rowers and the visualisations need to effectively convey the analysis of their data without requiring an understanding of the underlying data analysis. The goal is to provide feedback on training data in a way that is easy to understand and actionable for athletes.

The ultimate goal is to develop a model which could predict performance. This is the most ambitious goal of the project, due to the amount of data required to train a machine learning model. This goal was adapted to explore the use of machine learning to predict injuries in runners.

1.3 The Report

The report begins in Chapter 2 with a review of the literature on rowing training and performance and the use of machine learning in sports science. This will provide the sports science-related knowledge necessary to understand the approach used for data collection,

analysis, and model development. Chapter 3 discusses the approach used to collect and manage participant data, including the ethical considerations taken. Next, Chapter 4 outlines the approach to developing a data analysis pipeline to provide participants with feedback on their training data. This will cover steps taken to clean and standardise the incoming data, the general approach which guided analysis, and how visualisations were developed and deployed. Chapter 5 explores the machine learning approach taken. Unfortunately, due to limitations in data collection and time, the target model was not able to be fully developed. The machine learning goals were adapted to explore an approach to predict injuries in rowers. The report concludes with a discussion of the results and potential future work in Chapter 6.

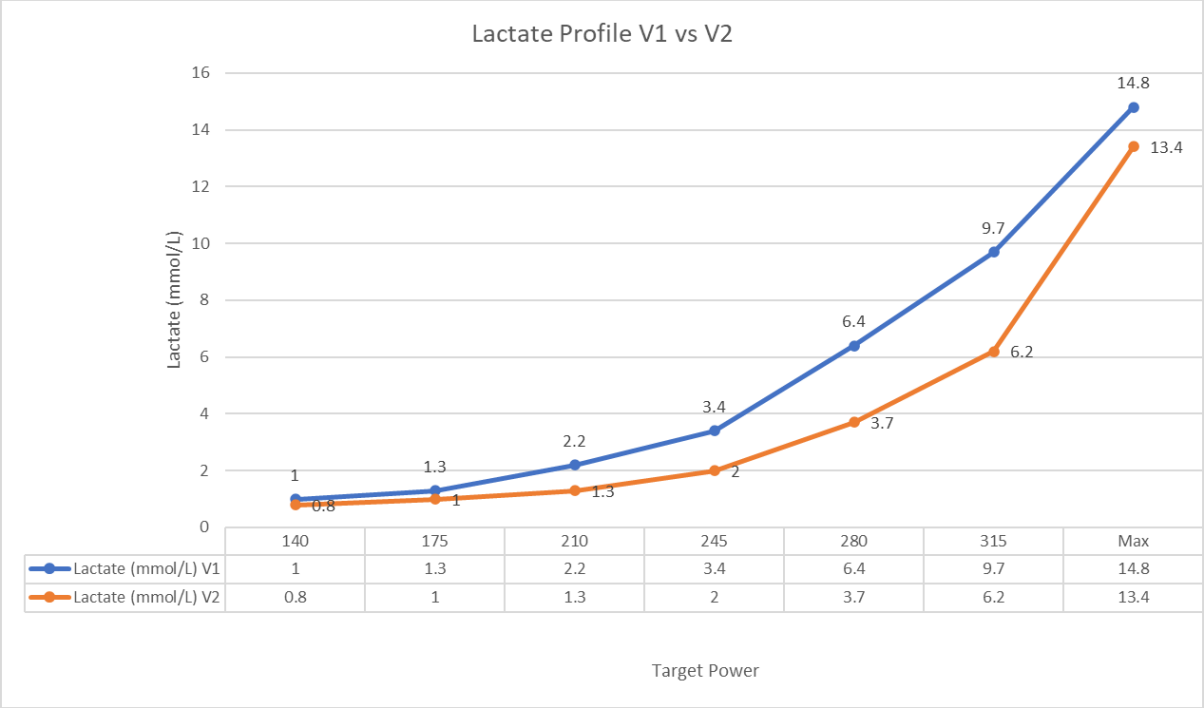
Chapter 2

Background

This chapter will cover the basic background of rowing, the sports science which guides rowing training, and how an athlete's body responds to training stimulus. Next, a review of performance modelling will explore the evolution of human performance modelling since the introduction of the basic Bannister model in 1975 [1]. The section will outline how training load and performance can be quantified, and explain the way these approximations are used in various performance models to date.

2.1 A Brief Introduction to Rowing

Rowing is an Olympic sport, raced across a 2,000 metre course, typically lasting six to seven minutes. It is classed as a power-endurance sport, this means training is focused on building aerobic, anaerobic, and power while also developing rowing technique [2]. Most training time is spent building endurance, next most time is spent building anaerobic capacity, with some remaining training time spent building strength and power through strength and condition sessions [3]. The importance of power is more significant in rowing than in cycling, another power-endurance sport for example, given the relatively short duration of exertions, with longer distance racing typically only covering five to seven kilometers, or fifteen to twenty-five minutes. Conversely, road cycling, tends to last for a longer period of time, where the shortest races might last two hours. There are many different approaches to how training is conducted and which energy systems are targeted for improvements in efficiency or strength. This section will discuss the basic training principles which guide training, the way athletes respond to different kinds of training loads, and how performance is evaluated in rowing.



	140	175	210	245	280	315	Max
Lactate (mmol/L) V1	1	1.3	2.2	3.4	6.4	9.7	14.8
Lactate (mmol/L) V2	0.8	1	1.3	2	3.7	6.2	13.4

Figure 2.1: Two lactate profiles completed 98 days apart following the protocol described above. Test one completed January 24, 2023. Test Two completed May 2, 2023

2.1.1 Training Principles

Generally, when a coach builds a training plan they have a few variables they can work with: volume, the amount of mileage or total time spent training, session intensity, how hard the given session is meant to be, and the frequency of, or the time spent in, different intensity zones. There are various ways to measure intensity, including heart rate, blood lactate concentration, velocity at maximal oxygen uptake (VO2 max), and relative perceived exertion (RPE) [4].

Blood lactate - A brief description

Blood lactate acts as a biomarker used regularly to determine muscular fatigue during exercise. Lactate is constantly produced by the body during the day. The concentration of blood lactate, measured in millimoles per litre (mmol/L), does not increase until the rate of lactate production surpasses the rate of lactate removal. Many things can affect the rate of lactate removal. Training can improve the rate of lactate removal, with certain sessions targetting that adaptation. During an exercise lactate test, a lactate profile is generated. Figure 2.1 shows a chart showing two lactate profiles generated during two lactate tests about 14 weeks apart (98 days) in January and May of 2023. The test conducted saw the rower complete seven, four-minute, efforts, at prescribed wattage targets. Details of each set of intervals can be found in A.1. These curves can be used to prescribe training as described below.

Training zones

Rowers tend to use heart rate zones or blood lactate concentration depending on access to the equipment to test blood lactate. Typically when a rower uses calculated aerobic zones each zone will be a percentage of HR_{max}. Following this approach, zones are defined as follows:

Z1 "Very Light" intensity, 50% - 60% of HR_{max}

Z2 "Light" intensity, 60%-70% of HR_{max}

Z3 "Moderate" intensity, 70%-80% of HR_{max}

Z4 "Hard" intensity, 80%-90% of HR_{max}

Z5 "Maximum" intensity, 90%-100% of HR_{max}

The exact definition of these zones varies in the literature, as does the method to calculate HR_{max}without a specific test. However, most high-level athletes will have completed some kind of stress test to determine their HR_{max}in order to train more effectively in their prescribed zones. A rower who uses lactate-based training zones might use the following zones:

T1 basic oxygen utilization training (UT2) [lactate = 0-2 mmol/L]

T2 oxygen utilization training (UT1) [lactate = 2-3.5 mmol/L]

T3 anaerobic threshold training (AT) [lactate = 3.5-4.5 mmol/L]

T4 oxygen transport training (TR) [lactate = 4.5-6 mmol/L]

T5 anaerobic capacity training (AN) [lactate $\geq$ 6 mmol/L] [5]

Depending on how rigorous the testing protocol was, Heart Rate zones may be calculated for each zone, these may vary from the aerobic zones calculated from HR_{max}.

The most basic zone approximation approach uses three zones based around certain physiological thresholds, such as lactate thresholds (LT_1 and LT_2) and ventilatory thresholds. The zones become simply low-intensity, moderate-intensity, and high-intensity.

Training prescription and distribution

There are a few different approaches for distributing intensity for endurance training. The three main methods are polarised training, sweet spot or threshold training, and pyramidal training. This directs the final factor a coach evaluates when building a training plan: frequency. To compare polarized training (POL), threshold training (THR), and pyramidal training (PYR), the more basic three zones of intensity will be used. The breakdown per zone for each training method is as follows:

Polarised Training Far more time spent in the low-intensity zone [6].

Low-Intensity 75%-85% of total training volume

Medium-Intensity 5%-10% of total training volume

High-Intensity 5%-10% of total training volume

Threshold Training More time spent in the medium-intensity zone [6].

Low-Intensity 45%-55% of total training volume

Medium-Intensity 35%-55% of total training volume

High-Intensity 15%-20% of total training volume

Pyramidal Training Most time spent in low-intensity zone with progressively less time spent in higher zones [7].

Low-Intensity 75%-85% of total training volume

Medium-Intensity 15%-20% of total training volume

High-Intensity 5%-10% of total training volume

This report will not compare the effectiveness of different training distributions. Different distributions tend to be used by different sports, or depending on which energy system is being targeted. The use of polarized training is most common in rowing [4], although other distributions are used, especially around competition time.

2.1.2 Energy Systems

In the human body, there are two major types of skeletal muscle fibres: fast twitch, and slow twitch. Slow twitch muscles are used for longer, slower contractions where relative strength is low. Conversely, fast twitch muscles are shorter, faster contractions where strength is relatively higher. These types of fibres store energy differently and respond to training differently. Slow twitch muscles can be adapted to longer, lower-effort, sessions and become resistant to fatigue, while fast twitch muscles fatigue easily, due to their lower glycogen capacity. Slow-twitch muscle fibres have a low anaerobic capacity, while fast-twitch muscle fibres have high anaerobic capacities. These two kinds of fibres draw on two different energy systems: anaerobic and aerobic systems. These systems provide muscles with Adenosine Triphosphate (ATP) which is used by the mitochondria in the muscle fibre cells to produce energy, allowing the muscles to contract [8].

Anaerobic System

The anaerobic system is used to provide energy to muscles to produce power without the use of oxygen. The anaerobic system typically provides energy for shorter periods of time through the immediate and short term energy systems. The immediate energy system can provide energy for 1-2 seconds of maximal work, this is typically used for resistance-based strength training, such as Olympic weightlifting. More commonly in rowers, the short-term energy system is used to provide energy to muscle fibres under significant strain. However, each time the energy system produces ATP, lactate is also produced. This system is limited by an athlete's ability to flush lactate from muscles. If the muscles are unable to flush the lactate quickly enough, the muscles will fatigue until failure forcing a stop to exercise.

Aerobic System

The aerobic, or long-term energy, system is the slowest at providing energy to muscles. This system uses sugars, fats and oxygen to produce ATP. This system only works when large amounts of oxygen are available, which for rowers is typically during lower-intensity sessions. Longer "steady state" sessions rely on the aerobic system to provide energy to muscles, with these sessions also being used to build the aerobic system. By spending time in the appropriate training zones the aerobic system builds efficiency by creating new capillaries to support the slow twitch fibres.

Rowers develop both their anaerobic and aerobic systems. Much of the winter season (typically September-March) is spent building the aerobic system, this is due to the longer length of any races done during this period, and to build a larger base on which to build a strong anaerobic system. As the sprint racing season begins, or shortly before, more anaerobic sessions will be introduced to build the system. Rowers will build their lactate tolerance, and spend more time at just-below-maximal efforts to prepare for the shorter race format.

2.1.3 Training Considerations

Training Cycles

When building a training plan many things are considered. Plans are typically built in macro-, meso-, and micro-cycles [9]. The macro-cycle in rowing is normally a yearly cycle, with the overall target of performing at a championship event at the end of the training year.

Meso-cycles are typically seasonal or monthly cycles. For example, the autumn season is focused on building base fitness and forming a crew's technical style, there may be small races that a squad may attend, but any specific preparation for these events is unlikely. The winter months will be focused on building the aerobic base, with some race preparation

towards the end of the season being introduced. The spring season will see the first race events. Typically these are "Head races", longer races (5-6km) allowing the longer length, lower intensity sessions from the winter to be leveraged. This season will also have an increase in interval and sprint race preparation in anticipation of the summer months. The summer season is the primary racing season, with the macro-cycle peak in performance. This meso-cycle will see a reduction in lower-intensity aerobic sessions and an increase in anaerobic intensity work to build fast-twitch muscle performances and the necessary racing performances of the shorter events.

Within these meso-cycles, there are micro-cycles typically lasting 7-14 days. It is crucial that these cycles are "periodised" appropriately. Periodisation is the budgeting of load across micro-cycles, ensuring athletes are experiencing the appropriate increase in intensity, followed by an appropriate length recovery period.

A coach will develop a training plan at the start of a new year, with a general idea of each cycle, and make adjustments as necessary throughout the year. If a coach had access to a system which could model the individual training responses for each athlete, individualised plans could be generated for each athlete, allowing the entire squad to train, and perform, at their physiological maximum.

Overtraining and burnout

Training for endurance sports puts a large amount of strain physically, and mentally on athletes. Endurance athletes are conditioned to consistently push their bodies, making it difficult to identify when overtraining has occurred. Spending too much time in hard training zones without providing enough time to rest and recover leads to overtraining. Overtraining can be diagnosed when a performance decrease, as a result of training fatigue, has not gone away within two weeks of relative rest [10]. Overtraining can manifest symptoms physically and emotionally. Athletes may become depressed, chronically fatigued, and lose their appetites before a drop in performance. Ensuring training plans are appropriately periodised (load is managed in sustainable cycles of intensity and rest), and with enough variety in sessions to avoid monotony can help avoid overtraining. Athlete stress management outside of training is also crucial to managing overtraining.

Overtraining can often lead to burnout. Burnout is when an athlete gives up a sport entirely. With many athletes attaching their self-esteem to success in their sport, they increase the risk of burning out. This is particularly problematic when an athlete may be suffering from overtraining, or underperformance [11].

Managing both overtraining and burnout are important when training, and can be considered when building a model of performance. Recognising overtraining can potentially prevent a promising athlete from retiring from their sport. By developing individual

performance models, it can be easier to recognise overtraining and adjustments can be made to help the athlete recover more quickly.

Tapering

The concept of a taper is common among endurance athletes. Training blocks are physically exhausting and typically leave athletes in an "over-reaching" state, where they are constantly fatigued. This encourages physiological adaptations but also negatively impacts performance. A taper period sees an athlete reduce their training load ahead of a competition [12]. Timing of the taper is important as athletes want to strike a balance between not losing too much training time, and consequently, fitness. However, by not having a long enough taper, there is a possibility the athlete will still experience some training-induced fatigue. Different taper strategies can be used, but typically session volume decreases, rather than session count. Typically as part of a taper, a very short period (2-3 days) of high-intensity, interval, sessions is completed. This is followed by a period of recovery, or relative rest, in order to induce a "super-compensation" effect. This effect is a result of the body continuing to recover beyond the original baseline of fitness in anticipation of another period of high-intensity interval sessions. This results in a minor performance boost [13].

2.1.4 Performance

There are two primary ways to measure performance in rowing. First, on the rowing machine, a 2,000-meter, 6,000-meter, or 30-minute at stroke rate 20 (30 r20) test is normally used, depending on when during the season the test is done. This can be used to determine a rower's fitness and can be used to track the progress of an athlete. These sessions are typically maximum effort sessions with some degree of taper (typically 4-7 days maximum) to elicit a peak performance. On the water times can also be used to judge an entire crew, with external factors like wind direction and intensity, water current flow-rate, and temperature considered. Some boats and squads may use on the water telemetry to quantify the impact and power output of each rower. On the water performances may not always be peak efforts depending on a squad's seasonal focus or an event's progression (e.g. a maximal effort will rarely make an appearance in the heat stage of a heat-semi-final progression). Ergometer scores tend to be considered the purest way of quantifying a rower's performance, however, a combination of approaches can be used to quantify on the water performance.

2.1.5 Summary

Rowing training is quite time-consuming for athletes. Many rowers will learn much of what has been discussed in this section to make better decisions themselves or understand why

training is done in a given way. The focus for the majority of the training cycle is to build a strong aerobic base to allow for a higher "peak", ideally coinciding with the peak event for a season of training. For international athletes, this will be a yearly cycle targeting the World Championships, which is part of a four-year cycle targeting the Olympics. For national-level athletes, national championships are typically the target, with Henley Royal Regatta potentially featuring as another season target. Coaches will normally take into account season targets, rower proficiency and base fitness when generating training plans. The target of this project is to make it easier for coaches and more accessible for athletes who may be more independent.

2.2 A Review of Performance Modelling

The first widespread approach to modelling performance was developed by Banister *et al.* [1] with the Banister Fitness–Fatigue model, and further refined by Morton *et al.* [14] when defining training impulse to determine fitness and fatigue. The use of machine learning, specifically artificial neural networks has since been introduced in approaches to model performance. In preparation for the 2000 Olympics in Sydney, Edelmann-nusser *et al.* [15] successfully predicted Olympic swimming performances within an error of 0.05 seconds across a total time of 2:12.64 (min:sec) for the 200m backstroke event. Edelmann-nusser *et al.* [15] specifically consider the limitation of using a linear model, such as the model proposed by Banister *et al.* [1], on training adaptation and performance; adaptation to training is inherently a complex non-linear process.

This section will review basic approaches to systems modelling, exploring metrics which are commonly used to guide models and some non-linear systems models which have been explored.

2.2.1 Quantifying Training Load (Fatigue)

Rate of perceived exertion (RPE)

Rate of perceived exertion (RPE) is a scale, originally introduced by Gunnar Borg, to measure an athlete's effort. The original scale ranged from 6-20, where 6 would be no exertion at all, and 20 is maximal effort. This scale ranges from 6-20 to correspond more easily with heart rate, as the scale is used beyond just athletic settings. Therefore another scale, ranging from 0-10, was developed by Borg as well for use with "extreme intensity of activity", this is the scale normally used in athletic studies. The second scale is called a Category-Ratio (CR) scale where the reference anchor is RPE CR10 of 10, meaning maximal effort/pain. These two scales, can be seen in Figure 2.2 and Figure 2.3. Session RPE (sRPE), using CR10, can be used, in conjunction with duration to determine session load

Borg RPE	
Score	**Level of exertion**
6	No exertion at all
7	
7.5	Extremely light
8	
9	Very light
10	
11	Light
12	
13	Somewhat hard
14	
15	Hard (heavy)
16	
17	Very hard
18	
19	Extremely hard
20	Maximal exertion

Figure 2.2: The original Borg 6-20 RPE Scale [16]

Borg CR10 scale	
Score	**Level of exertion**
0	No exertion at all
0.5	Very, very slight (just noticeable)
1	Very slight
2	Slight
3	Moderate
4	Somewhat severe
5	Severe
6	
7	Very severe
8	
9	Very, very severe (almost maximal)
10	Maximal

Figure 2.3: The Borg Category-Ratio 10 RPE Scale [16]

The two Borg RPE scales

[16] using the equation Load = sRPE $\times$ Duration.

Training impulse (TRIMP)

Training Impulse (TRIMP) is a method to calculate training load and is defined as TRIMP = Training Volume $\times$ Training Intensity [17]. There are many methods to calculate TRIMP, using different metrics to calculate training volume and training intensity. For the purposes of this project volume will simply be minutes, and intensity will be average heart rate (bpm), as the simplest TRIMP method outlined. Other modifications may apply a weighting against a heart rate metric to normalise longer sessions completed at a lower heart rate. The weighting factors considers the documented difference in heart rate responses to training in male and female athletes [14]. Alternatively, the load can be calculated by using the product of RPE and duration (ie. RPE $\times$ Duration (minutes)).

Acute chronic workload ratio (ACWR)

Acute Chronic Workload Ratio (ACWR) can be used to monitor load experienced by an athlete. It compares the training load accumulated, in arbitrary units (AU), over the last seven days (acute workload) to the training load accumulated over the last twenty-eight days (chronic workload). The exact calculations used to determine acute and chronic workload depend on what type of ACWR model is selected. Additionally, the exact time periods considered for the acute and chronic load can change depending on the application [18]. Typically ACWR is used for injury management in team sports, but some articles have found it may not be effective for preventing injury as it is an inaccurate way of

approximating load [19]. Training load estimation in rowing tends to be more objective given the objective internal load measurement of heart rate measured during a session. There is no clear consensus in literature about the effectiveness of using ACWR in injury prevention. As a relatively new concept, first published in 2016, more research into its efficacy and to provide validation needs to be completed [20]. Regardless of its effectiveness in reducing injury, ACWR charts are easy to produce and provide easy feedback to rowers to see how the training load can change week-to-week and ensure the load is not increased too quickly to begin to risk overtraining or injury. In practice, this metric can be used to determine if a rower is at risk of overtraining, or if they are not training hard enough. An ACWR in the range of 0.80 - 1.30 is considered optimal, with values above 1.5 considered the highest relative risk for injury, and values below 0.8 suggest an athlete is undertraining and is at a relatively high risk of injury.

2.2.2 Impulse-Response Models

Impulse-response models are a group of models, built on the model initially introduced by Banister *et al.* [1]. These models assume that a single exercise, or training session, produces two responses: fitness, or a positive performance response, and fatigue, or a negative performance response. A simplified version of this model is defined as

$$\textit{Performance} = \textit{Fitness} - \textit{Fatigue} \tag{2.1}$$

This simplified model is commonly called the Fitness-Fatigue impulse response model (FFM) and research since its introduction in 1975 has sought to refine how the *Fitness* and *Fatigue* components of this equation are defined.

Banister fitness-fatigue model

In the FFM model introduced in 1975 [1], and built upon by Morton *et al.* [14] in 1990, the functions for fitness and fatigue were defined, considering a cumulative training load and a decay factor for each response. This decay factor is different for both the fitness and fatigue responses, resulting in a longer decay period for fitness than fatigue, but also differs from person to person, offering a parameter to be tuned for each athlete. First, a session needs to be quantified. A session is defined as:

$$w(t) = D(\Delta\text{HR ratio}) \tag{2.2}$$

where

$$\Delta\text{HR ratio} = \frac{\text{HR}_{\text{exercise}} - \text{HR}_{\text{rest}}}{\text{HR}_{\text{max}} - \text{HR}_{\text{rest}}}$$

Morton *et al.* [14] acknowledged the disproportionate influence longer sessions at a low heart rate could have on resulting sessions $w(t)$, in arbitrary units (AU), so introduced a weighting factor Y. They determined $Y = e^{bx}$, where b is a pre-selected value determined for men and women, and $x = \Delta$HR ratio. So each training session is calculated as:

$$w(t) = D(\Delta\text{HR ratio})Y \tag{2.3}$$

Next, Fitness, $g(t)$, and Fatigue, $h(t)$, can be calculated. These two functions are defined as

$$g(t) = g(t-1)e^{\frac{-i}{\tau_1}} + w(t) \tag{2.4}$$

and

$$h(t) = h(t-1)e^{\frac{-i}{\tau_2}} + w(t) \tag{2.5}$$

where each function states the fitness and fatigue at the end of the day t, i is the number of days between the training session being added and the last training session, and τ_1 and τ_2 are the decay time constants for fitness and fatigue respectively. Finally, to model performance, two more constants are introduced: k_1 and k_2. These constants are weighting factors for fitness and fatigue, they have "no direct physiological interpretation" [14] but can be used to adjust the model for athletes who recover more quickly to heavy training load or require more time to recover from sessions across a taper period. Finally, performance is modelled as

$$p(t) = k_1 g(t) - k_2 h(t) \tag{2.6}$$

The model was originally developed using a swimmer. When evaluated using linear regression for the athlete-specific parameters (τ_1, τ_2, k_1, and k_2), the model reasonably estimates the performance outcome of a swimmer with relative confidence. No r^2 value is provided for the swimming application in 1975 [1]. In the 1990 model from Morton *et al.* [14], two of the researchers participated in a training and testing regimen and a "good degree of fit" was observed, with r^2 values of 0.71 and 0.96 calculated for a slightly trained and an untrained runner respectively, with the former being given a starting 1000 AU points given some pre-existing running fitness. The impulse-response model, if trained appropriately can help guide the structure and timing of a taper period for each athlete [14]. Versions of this kind of impulse-response model are used in basic analysis across many consumer fitness apps today, such as Training Peaks and Strava.

Limitations to the impulse-response model

The Banister FFM simplifies human physiology quite effectively. However, to tune the parameters of the model, regular "criterion" testing needs to occur. These are maximal efforts done at regular intervals. During the testing of the Morton *et al.* [14] model

refinements, both participants tested roughly once per week, something which is not feasible for high-level endurance athletes. This dependency on regular testing to tune parameters makes it particularly difficult to implement effectively as a holistic model for rowers. Furthermore, the use of data to adjust model parameters results in less data being available to model training behaviour.

The model only considers training load, specifically endurance training load, as an indicator of performance. Given the power dominance in rowing, a training block with seemingly less training load, as a result of more time and energy spent doing strength training. This can result in a stronger performance when tested next which may not be explained by changes in endurance training load. Additionally, the model does not consider other factors such as recovery and non-training strain an athlete might experience.

Finally, if the model is not constantly updated, it cannot predict future performances effectively. For the purposes of this project, a model would require an understanding of training behaviours to model the physiological response, predicting future performances. By modelling physiological responses, suggestions can be made to improve certain energy system efficiency or effectiveness if needed.

Impulse-response models are still widely used to model training and adaptation for athletes, their simplicity being a key factor for this. To more effectively model performances, without the need for constant testing, alternative models can and should be used. Impulse-response models can still be used in machine learning solutions, though, but as part of a larger ensemble approach [21]. These models can also be used to validate and evaluate machine learning-based models.

2.2.3 Alternative Models

There are models which employ more complicated mathematical approaches. This section will explore one of the most developed models, Performance Potential (PerPot), and explore artificial neural network approaches used by some researchers in recent years.

Performance potential (PerPot)

Performance Potential (PerPot) was first introduced by Perl [22] in 2001. It is a performance potential meta-model which models responses to training input through a flow model. PerPot approaches performance modelling similarly to the impulse-response approach, each training load has an antagonistic response, and two contradicting effects: response potential and strain potential. These can be compared to the impulse-response fitness and fatigue responses, respectively. Strain and response potentials impact the performance potential with differing decay factors, like impulse-response models as well. There have been reports of PerPot being effective in modelling cycling training [23]. PerPot,

unfortunately, suffers from one of the key issues with impulse-response modelling: repeated testing is needed to tune the model's parameters. Additionally, due to the nature of the decay factors for strain and response potentials, the super-compensation effect is not considered with the basic PerPot model [23].

Artificial neural network (ANN) approaches

Artificial neural networks (ANN) have been used to effectively model performance based on training data as early as 2000 by Edelmann-nusser *et al.* [15] where (200m - Backstroke) swimming performances at the 2000 Olympic Games were predicted based entirely on training data. In building their model, Edelmann-nusser *et al.* [15] considered the use of linear models like the impulse-response model, but commented that biological adaptations are inherently non-linear, and therefore set out to "demonstrate that the adaptive behaviour of an elite female swimmer can be modelled by means of the non-linear mathematical method of artificial neural networks".

Machine learning techniques require vast quantities of data in order to be effective. Edelmann-nusser *et al.* [15], however, used a single athlete, collecting 95 weeks of training data and 19 competitive performances. This is quite a small training set for a machine learning application. With only 19 performances to compare a performance model to, the risk of overfitting is significant. Edelmann-nusser *et al.* [15] completed three data analysis steps to generate a model. The first two steps involved determining the impact of a 2-week taper period and the impact of a 2-week intense period immediately before the taper (3-4 weeks before competition) for steps one and two respectively. The final step combined the first two steps to determine the influence of the four weeks leading into the competition period. The result of this neural network approach was quite successful and Edelmann-nusser *et al.* [15] concluded that neural networks can be effectively used on small datasets to predict performances.

Churchill [23] sought to build on the success of using ANNs by Edelmann-nusser *et al.* [15] to model performance outcomes, but targeting professional cycling performances. Churchill, however, struggled to collect data for peak performances due to the strategic nature of cycling. Churchill did develop techniques to smooth noise in small datasets using an ensemble approach, citing the larger the number of neural networks used, the better smoothing was applied to noise. Artificial neural networks have consistently shown promise in predicting performances in swimming and cycling, even addressing small dataset issues which are prevalent in machine learning applications. This project aspires to build on these successes and apply ANN and ensemble techniques to individual rowing training data to predict ergometer performances.

Chapter 3

Data Collection and Management

The first step in building any machine learning model is collecting, or acquiring, data with which to train and develop a model. To collect user data there were also ethical considerations to take into account and ethics approval was required. This chapter will discuss the data collection process, how the data was managed and stored for use in the project, and the ethical considerations and concerns that were a part of that data collection and storage development.

3.1 Research Ethics Considerations and Approval

When collecting data from participants, it is important to consider the ethical implications of that data collection. This is especially true when the data being collected is personal data, such as the data collected in this project. The data collected in this project was personal data, as it included information about the user's training sessions, including in many cases heart rate and GPS data, and their details such as their name and email address. As such, it was important to consider the ethical implications of collecting this data and to ensure that the data was collected in a way that was respectful of the user's privacy and rights. Furthermore, all of the participants are active competitors, many of them competing against the researcher in the same events. It was important to ensure that the data collected was not used in any way that could be seen as an unfair advantage. Some squads have strict policies surrounding the sharing of training data and scores; the same data privacy requirements were required to begin conversations with these squads and their members to collect data.

When applying for ethics approval it was important to consider how user data could be protected both from unauthorised access and, in the unlikely event of a data leak, ensuring the exposure a user sees is minimal.

To ensure user data protection, much of the connection to an actual person is removed, the data is stored in a database with a unique identifier, and the user's email address is stored in a separate table. This means that if the database is accessed, the user's email address or other identifying information is not immediately available. The data is also stored in a secure database, with access restricted to only the researcher and the serverless functions used for analysis. To further ensure data protection for participants during the project, all data received is encrypted at rest and in transit. A table linking the unique identifiers generated for each user and identifiable user data was stored securely on the researcher's personal machine making the de-pseudonymisation of the data more difficult. This key was kept to provide users feedback through the website and to allow the researcher to delete any collected information if a user elected to leave the project early.

There was some GPS data available in some training sessions submitted as part of the project. It was determined that this data, if leaked, would not be a significant risk to the user, as the data was only collected during on the water training sessions where many rowers train and compete normally. As most rowers in a squad train together, if the raw data for multiple sessions were leaked it would be nearly impossible to pinpoint which sessions belonged to which athletes given the number of total athletes training at any one time.

All of the analytical approaches were tested using the researcher's training data. This ensured that the researcher did not get an unfair advantage when compared to other athletes' data; any further analysis done for individual users was done using serverless functions meaning the researcher never had direct access to the raw data. These policies and procedures were included in the ethics application, approved by the College Research Ethics Committee, and shared with participants during recruitment.

Following some difficulties with the new College Research Ethics Approval Management System (REAMS), an ethics approval application was submitted on November 2, 2023, and approval was granted on December 6, 2023. It is available to view in the appendix A.5.

3.2 Data Collection

When developing the data collection pipeline, there were some key requirements to consider. Firstly, the user effort per activity logged was to be as minimal as possible. Rowers typically have a lot of data to log, so the more effort required to log each activity, the less likely they are to do so. In many cases they were already logging their training elsewhere as a part of their squad's training program, therefore making the process of logging their sessions for this project as seamless as possible became a priority. Next, for what little interaction was required, the platform needed to be easy to use and intuitive. This was important as it reduced the workload of maintaining the platform, by responding to user queries, and

encouraged users to continue using the platform due to its ease of use.

Considering the minimum requirements for a successful data collection pipeline, a website was developed to manage data provider connections and view training feedback. This website worked in conjunction with a series of serverless functions to automatically collect and analyse user sessions, and generate the feedback which was presented in the website.

3.2.1 The Frontend

The website was developed using Next.js, a React framework for building server-side rendered websites, it was developed using Typescript, which is a statically typed superset of JavaScript. This allowed for type checking and better code quality and made it easier to develop and maintain the backend. It was designed to be simple and easy to use, with a focus on the user experience. This meant that a mobile-first approach to design was needed, as many users would be accessing the website from their mobile devices, it also needed to be fast and lightweight meaning many of the data-heavy components of the web app are rendered with server-side components to reduce the client-side workload ensuring a smooth and consistent app experience. The web app is also fairly basic, including only three screens once logged in.

Login and register

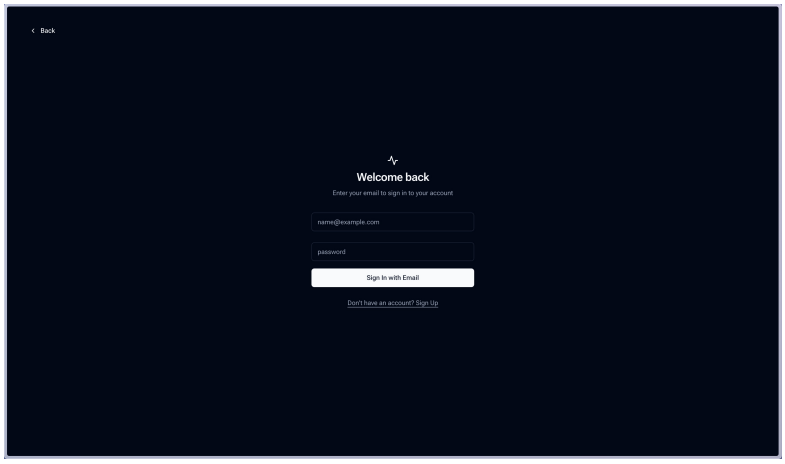


Figure 3.1: The login screen for the web app

When users first navigate to the website, they are greeted with either the login screen (Figure 3.1) or the register screen (Figure 3.2). The login screen is simple, with only two fields, one for the user's email address, and one for their password. The register screen is similarly simple, with fields for the user's name, email address, and password. Users are also required to agree to the consent form and acknowledge they have read the information sheet provided by a PDF link. The consent form and information sheet were written as part of the

ethics approval obtained to collect user information. Once the user has registered, they are automatically logged in and redirected to the dashboard screen.

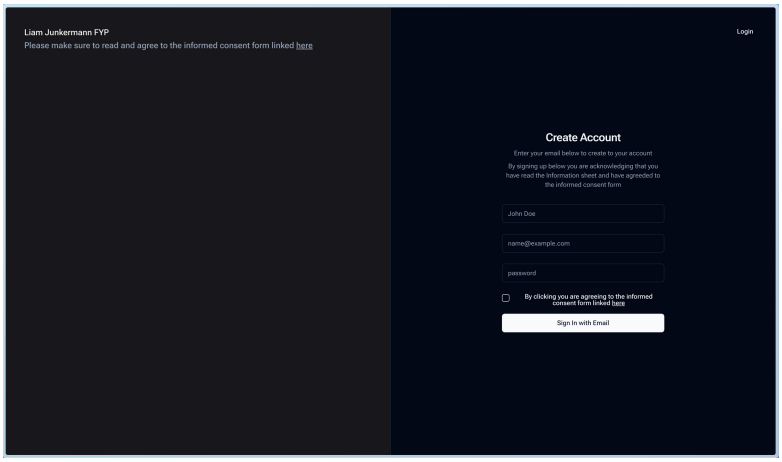


Figure 3.2: The user registration screen for the web app

Dashboard

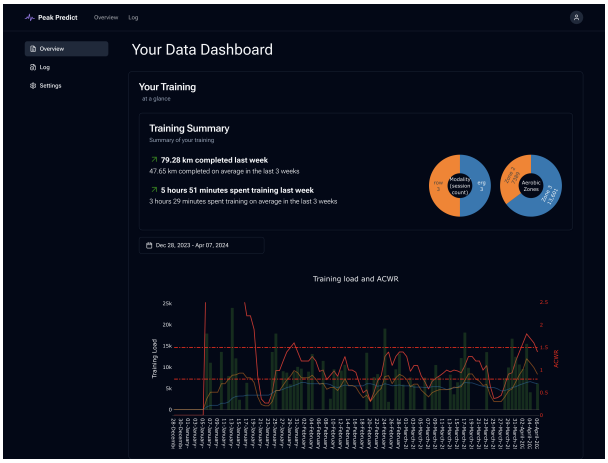


Figure 3.3: The dashboard screen for the web app

Once logged in, users are greeted with a dashboard screen (Figure 3.3). This screen gives a summary of training in the last seven days, as well as some visualisations across a selected time period of the user's choice. At a glance, athletes can identify if they are training sustainably, and compare, week-to-week, their mileage and modality breakdown. Managing training load is particularly important for rowers as overtraining has even greater knock-on effects in other aspects of a rower's life, like work or college.

Training log

The training log (Figure 3.4) provides a detailed view of all the sessions a rower has completed. This project did not require athletes to provide additional feedback for sessions,

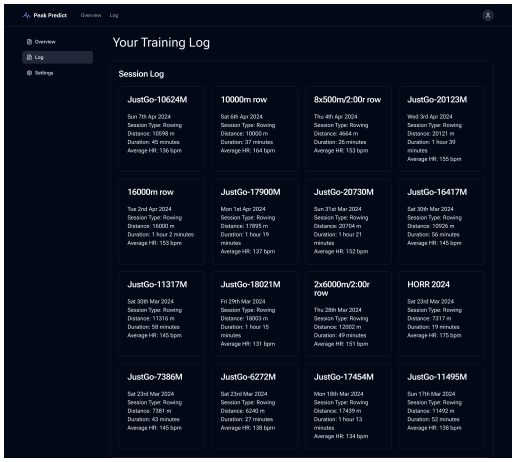


Figure 3.4: The training log screen for the web app with 16 sessions loaded

however, this page would host a form to allow athletes to add sessions and session feedback in a more developed version of the web app, this will be discussed more in the Discussion, Chapter 6. The user can also click on a session to view more detailed information about that session, including heart rate data, GPS data, and any other data collected during the session.

Connections manager

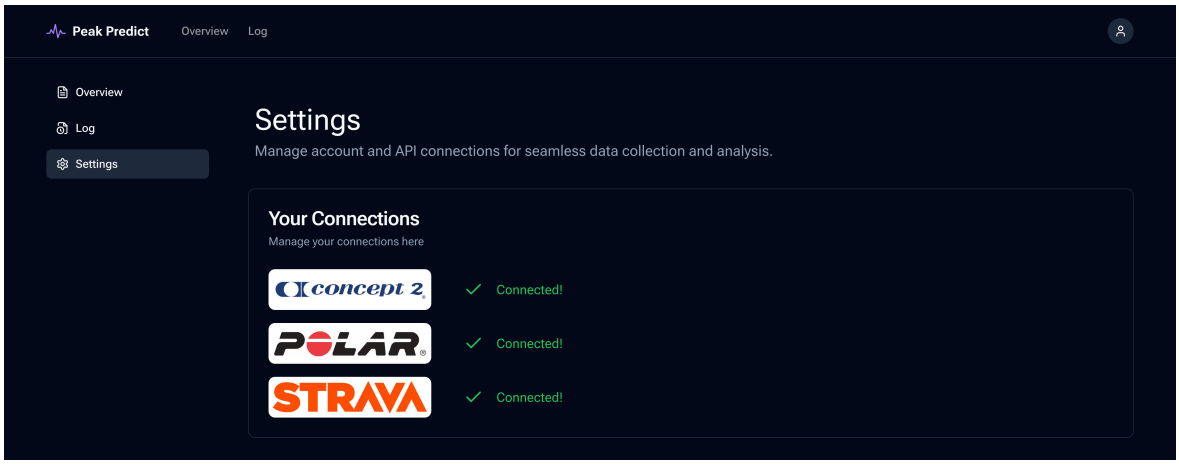


Figure 3.5: The connections manager screen for the web app

The final screen in the app is the settings page (Figure 3.5), which is home to the connections manager. Here users can see which data providers they have connected. This page is where the most user interaction happened throughout the project. Given the goal of minimal user intervention and maintenance, when logging in for the first time they were directed to this page to connect their rowing data providers, and using the OAuth protocol, shared the necessary keys with the project to allow data to flow automatically in the backend.

3.2.2 The Backend

The backend services were crucial to developing a system that required minimal user and maintainer intervention. The backend was developed using serverless functions, which are small, single-purpose functions that are run in response to events. These functions were used to collect data from the user's data providers, analyse the data, and generate feedback for the user. The backend was developed using the Serverless framework, which is a framework for building serverless applications. The backend was hosted on AWS and used a combination of AWS Lambda, AWS API Gateway, AWS DynamoDB, and AWS S3 to provide serverless functions and store the data. The backend functions were primarily built using Node.js and Typescript. This section will only discuss the data collection and management functions. Analysis functions will be discussed in the Data Analysis and Visualisation chapter, Chapter 4.

There are 4 major functions, which were separated into serverless functions to manage and handle data. These functions are the Provider Connection, Provider Webhooks, Data Ingestor, and Data Processor. Each of these functions is responsible for a different part of the data collection and management process.

The general web app backend was mostly handled by AWS Amplify. This service made it possible for the researcher to focus on providing valuable features to users, rather than spending extra time developing boilerplate authentication and database interactions. The Amplify service also provided a GraphQL API, which was used to interact with the DynamoDB database, and provided a simple way to interact with the database from the frontend in a secure way.

Provider connection

When connecting to a service using the OAuth protocol, the user is redirected to the service's login page, where they log in and authorise the project to access their data. Once the user has authorised the project, the service redirects the user back to the project's website and provides an access token that can be used to access the user's data. This access token is then stored in the database and is used to access the user's data in the future. This function is responsible for handling the OAuth flow and storing the access token in the database. The provider connection function handles the various data formats which different providers use and connects access and refresh tokens to the users for use later on in the pipeline when activities are generated. This particular function was built as part of the Next.js server, leveraging the server-side rendering capabilities of the framework to handle the authentication of users using cookies making the linking of tokens with users easier.

Provider webhooks

Typically when a user completes a workout, the data provider can query a webhook (if it has been set up), with a payload containing basic information about the workout, and in some cases, the details to retrieve the full session. For each of the supported data providers, Polar, Concept2, and Strava, webhooks were setup and registered with the providers. Webhooks were then handled through the Next.js server, retrieving any additional data needed. The full session data was then shared with the Data Ingestor function for processing and storage in the database. Different providers had different requirements for the webhook payload, and the data was stored in different formats, so the webhook function was responsible for handling these differences and normalising the data before sending it to the Data Ingestor function, while also limiting the unnecessary sharing of client secrets and keys across multiple services.

Data ingestor

The data ingestor handled the actual storage of the data in the database. The data ingestor first validated any data provided. This was done by validating the provider user ID and comparing a key which was shared with the Next.js server. Then it stored the raw session data, provided by the webhook handler in the Next.js server, in AWS S3, this meant that any issues with data processing could be rectified after the fact with the raw data available. Finally, the ingestor triggered the Data Processor function. This extra level of data processing before storing the data ensured that data ingested into the data pipeline was legitimate and provided through the proper channel. This function was a standalone serverless function, built using Node.js and Typescript, and deployed to AWS Lambda using the Serverless framework.

Data processor

Given the variety of data and session types provided, a standard exercise model was needed to make analysis and machine learning easier. This standard exercise model will be discussed in more detail in the data cleaning section of the data analysis chapter (4.1), with a schema introduced in the Data management section of this chapter (3.3). The data processor function was responsible for taking the raw session data, and converting it into a standard format. The standard format included a session's duration, distance, average heart rate (if applicable), time of completion, session type, and session modality. These features would be used to complete the analysis and produce the weekly training summaries shown on the front end. This was also a standalone serverless function, built using Node.js and Typescript, and deployed to AWS Lambda using the Serverless framework.

3.3 Data Management

Given the expected volume of data collected, the design of how to store and manage this data was important to ensuring analysis could happen quickly, and AWS costs remained low. Almost all the data stored was already provided in a structured JSON format, originally these data responses were saved and stored in AWS S3. This allowed the researcher time to understand what data was being collected, and what was important. From this, the Standard Exercise Model was derived. The process for how this model was developed will be detailed in the next chapter 4. This model was then used to create a DynamoDB table to store the data in a structured format. The data was stored in a table with the schema described in Table 3.1.

Column Name	Column Data Type	Description
`owner`	`string`	The users uid
`exercise_id`	`string`	An autogenerated id. The tables primary key
`data_source`	`{source: string, path: string}[]`	The raw data source(s) and linked data in S3
`datetime`	`number`	The unix timestamp for the session
`type`	`"endurance" | "interval" | "strength" | "other" | "unknown"`	The session type, an enumerated list
`modality`	`"row" | "erg" | "other"`	The session modality, an enumerated list
`duration`	`number`	The duration, in seconds, of the session
`distance`	`number`	The distance, in meters, of the session
`avg_hr`	`number`	The average HR, in bpm, of the session

Table 3.1: The schema for the "Standard Exercise Model"

This schema made it particularly easy to perform analysis on each athlete's training, distilling the key metrics into a small data size. DynamoDB also allows for robust querying of large datasets, allowing this solution to scale well with lots of data. The data was stored in a single table, with the `exercise_id` as the partition key, and the user's unique id (`owner`) as the sort key. This allowed for easy querying of the data and made it easy to retrieve data for a specific user, or a specific session. With connections to the raw data included in this data format, more detailed analysis can still be completed automatically, without unnecessary S3 reads to find the session amongst a larger, less dynamic data set. Future work could continue to grow this standard model, adding more features which may be commonly used for analysis, such as time in zones, calculated training impulse, or more. However, for the purposes of providing data for the frontend and the initial analysis, this model was sufficient.

Chapter 4

Data Analysis and Visualization

Collecting and managing data is a key part of this project, although simply collecting data does not provide any benefit to athletes. Data analysis and visualization are essential to extract useful information from the data. This chapter will discuss the data analysis and visualization process, including cleaning of the data collected, the analysis performed, and the visualizations generated and presented as feedback to users.

4.1 Data Cleaning

The data collection step was developed to reduce any cognitive load for users. As a result, there are many data sources, and therefore many data formats, which need to be cleaned and transformed into a uniform format for further analysis. Furthermore, some athletes might use several devices to track a session, so it is also necessary to match multiple session results to a given session. All the raw data from the data providers was stored throughout the project in case the data cleaning steps changed; this allowed database tables to be as small, and queryable, as possible.

4.1.1 Identifying Valuable Data

In order to begin data cleaning, it was important to understand what information was valuable. To do this, collected data was analysed for what information was available across each session provided, then understanding what basic analysis and feedback would be most useful to athletes. The most basic information needed is the date and time of a session, what modality of session it is (row, ergometer, or other), and what type of session it is (endurance, interval, strength). These modalities and types were intentionally kept quite simple, with only three options, to make analysis easier to start. Next, to do any kind of fatigue or training impulse analysis, the duration of a session needed to be considered, along with any available heart rate or RPE information. Finally, rowers measure progress in

training through mileage so a session's mileage, where appropriate, was included in this standardised model. As a result of this brief macro-analysis of the data and expected analysis results, the "Standard Exercise Model" described in the previous chapter (Table 3.1) was generated.

4.1.2 Developing a Cleaning Process

The data collection pipeline ingests sessions regardless of type. This required a series of functions to handle data from different providers and perform analysis on session data to appropriately classify each session, and normalise to the standard exercise model developed in the previous chapter.

Identifying session modality

The first step in cleaning data was identifying the session's modality. Was a session a row, an erg, or something else? Typically the sessions were tagged, with sessions coming from Strava being classed as "rowing", regardless of if they were on the water or the rowing machine, causing some issues. Sessions not explicitly tagged by Strava, for example, or ingested from Concept2 were initially flagged as other. This approach, however, did result in some issues. One member of the Commercial Rowing Club men's senior squad, for example, uses a running watch which classifies on the water sessions as runs. He also happens to be an avid runner so it was necessary to analyse sessions to determine if they occurred on land or on water. As a result of this case, and the issues with differentiating on the water or the machine "rowing" from Strava, it became clear a more developed approach was necessary to accurately identify session types. Differentiating on the water rows and erg sessions was much easier. In many cases, erg sessions were provided through the Concept2 API and were easily identified. On the water rows were typically provided through Strava and were identified by the activity type and generally the presence of GPS data. Users who did not have a GPS in the boat could manually insert sessions and the name of the session could be used to identify the modality.

Identifying session type

The next challenge in the data cleaning process was determining if a session was an endurance or interval session. Typically, the distance and duration can be used to identify a session. For erg sessions provided through the Concept2 API, distance, time, watts, and pace were included at a minimum. It is possible to differentiate endurance or interval sessions when comparing pace with other sessions from that user. The ideal session data from Concept2, though, includes highly detailed information about a session such as distance, duration, average wattage, pace, and heart rate, and detailed stroke-by-stroke data including wattage, pace, and heart rate. Many athletes, though, choose not to connect a

heart rate monitor to their rowing machine, so it was necessary to match heart rate data, recorded through Polar, with the stroke data, when available. Using heart rate it is very easy to identify if an erg session is an interval or endurance session.

Identifying the type of an on the water row was more difficult. Some on the water sessions may be a mix of interval and endurance. A common training session for the researcher throughout this project was 10 kilometres of endurance paddling, followed by 10 kilometres of interval work. In this case, the session could be classified as an interval session. The researcher, however, would separate this into two sessions, classifying the first 10 kilometres as endurance training, considering the duration and average heart rate, and the second 10 kilometres as interval training, again considering the duration and average heart rate to calculate the training impulse. This is a clear example of the limitations of the data collected. The researcher could manually classify this session as an endurance session, but this would not be scalable. A model could be developed to classify this session, but unfortunately, not enough data was collected and classified to warrant the time needed to develop the model. As a result, the approach used to classify sessions relied on distance and average heart rate, if available. Sessions greater than 18 kilometres in distance were considered endurance sessions, then average heart rate was used to identify endurance or interval sessions below this threshold. This is a limitation of the data collected and is a clear area for future work.

Finally, strength sessions needed to be identified. When a session was ingested as a strength session this was typically already tagged. Logging strength sessions using a heart rate watch typically is not effective. Strain experienced during strength sessions rarely has a direct correlation to heart rate as cardio sessions do. Furthermore, many users did not log strength sessions using one of the supported data sources as most athletes made use of an app like TeamBuildr[1], to track these sessions. As a result, strength sessions were not normally included in the training impulse analysis. This is another clear limitation of the data collection process implemented and an area for future work.

4.2 Data Analysis

The data collected and cleaned was then analysed to provide feedback to athletes. Several analysis metrics were generated. Starting with basic metrics such as weekly mileage and duration, time spent in given training zones, analysis of time and mileage completed across different modalities, and basic training load analysis. This section will discuss the approach used to generate analysis metrics for the project, and how the Standard Exercise Model was used and adapted.

[1]TeamBuildr is an online strength and conditioning platform for coaches to build S&C plans for their athletes and for athletes to track their weights sessions

4.2.1 Analysis Development Methods

Using the Standard Exercise Model developed in the previous chapter, analysis was performed. In order to generate metrics, the data available first needed to be explored. This exploration and initial analysis was done using Python, in a Jupyter Notebook, with the help of packages like pandas and numpy. To do this, core session information from the researcher's sessions was fetched from the Standard Exercise Model table, and basic analysis was performed. This included calculating weekly mileage and duration, time spent in training zones, and training impulse. Then, using the original data source, modality analysis could be refined to produce more effective feedback for athletes. This process was repeated for each analysis metric generated. Once a series of analysis metrics were generated in a Jupyter notebook, functions were developed to automate the process. These functions were then used to generate analysis metrics for all users in the database. In some cases, the analysis functions were ported to Typescript to be run directly on the client side to generate metrics based on different timeframes selected by users on the frontend. Any data that was ever interacted with was data the researcher had generated as a part of training. This was done to ensure the data was clean and accurate and that the researcher was not interacting with sensitive data.

4.2.2 Analysis Metrics Generated

The metrics generated for analysis were based on the Standard Exercise Model developed in the previous chapter. The model was used to generate some analysis metrics for rowers. These metrics were then used to provide feedback for rowers in the application. The metrics generated were:

Weekly mileage and duration

Calculating weekly mileage and training duration metrics was very straightforward. Every session ingested by the platform required, at a minimum, duration, and if the session was a row or erg session, mileage. Using the pandas library in Python, the data was grouped by day, week, and month and the sums of mileage and duration were calculated. This data was then used to generate a graph of weekly mileage and duration for each rower. This graph was used to track progress over time and to provide feedback to rowers on their training volume. This is a key metric for rowers to track as it is a key indicator of progress in training.

Training zone mileage and duration

Training zones were calculated using the heart rate data available for each session. The researcher used the heart rate data to calculate the time spent in each training zone. This was done using the average heart rate described in the Standard Exercise Model. Athletes

were asked to input their maximum and resting heart rates, which were used to determine training zones based on HR ratio. The time and mileage spent in each zone was calculated similarly to the weekly mileage and duration metrics. A pandas pivot table was used to group the data by zone and sum the mileage and duration. This data was then included in the daily, weekly, and monthly Dataframes and passed on to the further analysis functions and in turn the visualisation functions. Understanding how much time is spent in each zone throughout the week allows rowers to ensure they are training effectively. If they are following a polarised or threshold training plan, this metric can be used to ensure they are executing the correct split of training zones.

Modality analysis

Modality analysis was performed to understand how much time was spent on the water versus on the erg. This was done by grouping the data by modality and summing the mileage and duration. This data was also included in the duration Dataframes which were passed to the visualisation functions. This metric is important for rowers to track as it can indicate if they are spending too much time on the erg or on the water. This can be important for rowers to track as it can indicate if they are at risk of injury from overtraining on the erg or if they are not spending enough time on the water to develop their technique. When looking at a rower as a coach, more time on the water can lead to more efficiency in technique. If a rower finds they are not performing as expected on the water, they may need to spend more time on the water to develop their technique. Additionally, when considering features for a model to predict injury, extensive time on the erg may potentially be a factor. This can be a potential issue particularly if a rower changes from doing no erg work to almost exclusively erg work. Deeper analysis can explore a wider range of potential modalities, such as running and cycling, and exploring how these modalities interact with rowing training and performance.

Training load analysis

The most in-depth analysis performed was training impulse analysis. As described in the training quantification subsection (subsection 2.2.1), training impulse is an approximation of the strain a training session places on an athlete. For this project, it is calculated by multiplying the duration of a session by the average heart rate of the session. This is a simple way to quantify the stress a session places on an athlete. The training impulse of a session can be used to calculate the training load of a week, month, or year. This can be used to track fatigue and fitness over time. Due to the limitations in data available, fitness and fatigue scores for rowers were not calculated, this will be discussed further in the Data Analysis section of the Discussion Chapter (6.2). Each session with heart rate data available had a training impulse (`trimp`) value generated. This value could then be used to calculate

acute and chronic workload, which in turn were used to calculate the acute chronic workload ratio (ACWR) a rower was experiencing. Presenting this analysis to rowers can indicate if they are at risk of injury by overtraining, as a result of increasing load, or undertraining for a period of time beyond a productive taper. This is a key metric for rowers to track and is a key part of the feedback provided to rowers in the application.

4.3 Data Visualisation

The data visualisation built upon the analysis done, representing the outcomes of the analysis graphically making them easier to understand at a glance. To develop visualisations, Python was again used in a Jupyter Notebook, with the help of the matplotlib library to generate plots. The plots generated in the Jupyter Notebook were then mostly ported to Typescript, using the Plotly.js library, to be used in the application. This section will discuss the approach used to generate visualisations for the project, how the analysis metrics discussed in the previous section were used and adapted, and the process for bringing visualisations to the frontend.

4.3.1 Visualisation Development Methods

As a result of the analysis steps taken, a set of pandas DataFrames were generated, these could be used to easily generate basic plots for visualisations. In isolation though, certain plots were not as useful as they could be. For example, a bar plot of daily training impulses is not particularly useful, especially given the arbitrary nature of the units it is measured in. A more useful plot may overlay acute workload, chronic workload, and ACWR lines plots to add some more useful context for a rower. This iterative approach of visually exploring the data, determining if a metric in isolation was useful, or finding a suitable pairing of metrics to plot together, was used to generate the visualisations for the project. Each plot generated was made into a function which including any data transformations necessary to generate the visualisation. These functions were then ported to Typescript to generate the graphs in the application, leveraging the Plotly.js library.

4.3.2 Visualisations Generated

This subsection discusses the visualisations generated. All the data used is from the researcher's training and is up to date as of April 7, 2024.

Acute, Chronic, TRIMP, ACWR Chart

The primary chart implemented on the frontend is the ACWR chart. The version shown in Figure 4.1, is generated from the Jupyter notebook used to develop visualisations, using the

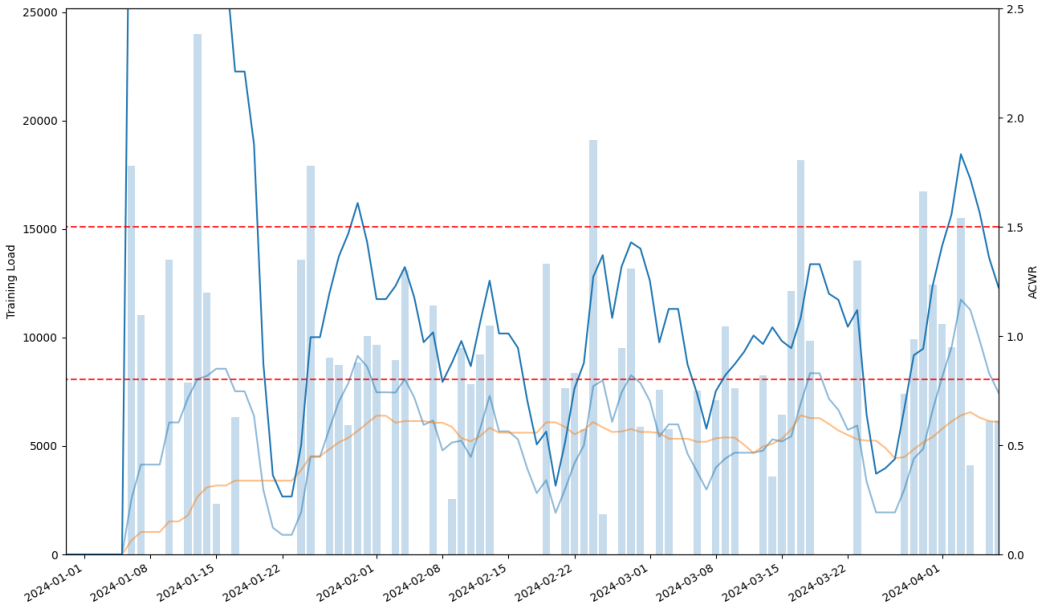


Figure 4.1: The Acute, Chronic, TRIMP, ACWR chart

researchers training data on April 7, 2024. This chart looks at a roughly 100-day period from December 28 to April 7, 2024. During this period the researcher recovered from an ankle sprain, and two periods of illness, accounting for gaps in training (trimp bars) and the erratic nature of the ACWR line.

This chart displays daily training impulse (arbitrary units), as bars, acute and chronic workload (arbitrary units), as low opacity lines (acute workload in blue, chronic workload in orange), and ACWR on a second axis (solid blue line). This chart is useful for rowers to track their training load over time and to ensure they are not at risk of injury from overtraining. The two red lines overlayed on the ACWR axis at 0.8 and 1.3 indicate the optimal training load range.

When implementing this chart in the frontend, a date picker was added to allow users to explore how their training over different periods was balanced. This resulted in a more interactive chart, allowing users to explore their training load over time. This chart was the most complex to implement, as it required the most data to be passed to the frontend, and given the multi-axis nature of the data, it was not possible to create a generalised, reusable, chart component.

Figure 4.2: The weekly training overview widget, comparing the last week's mileage and training time to the average of the previous 3 weeks

Weekly training overview

The weekly training overview widget (Figure 4.2) was the only visualisation not initially developed in the Jupyter Notebook, a text version was created, but due to the styling needed, it made more sense to develop this directly in Typescript. This visualisation shows the mileage and training duration for the last week, which updates each day, and compares that with the rolling average of the last 3 weeks. This chart is useful for rowers to track their progress over time and to ensure they are hitting their weekly mileage targets. This visualisation was the simplest to implement, as it required the least data to be passed to the frontend, and was an easy-to-generalise component in react.

Modality and Zone Donuts

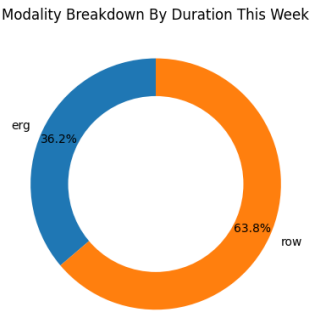


Figure 4.3: A donut graph showing the distribution of training duration across different modalities

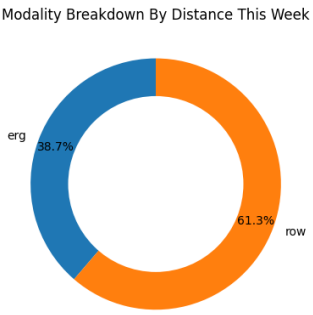


Figure 4.4: A donut graph showing the distribution of training distance across different modalities

Rowers tend to split training across a few modalities, most training is completed either on the erg or on the water. The donut charts in Figure 4.3 and Figure 4.4 show the distribution of training duration and distance across different modalities. The data shown is the training distribution for the week starting on April 1, 2024. One drawback of the data collection and cleaning implemented for this project was the limited number of training modalities logged. Including modalities such as cycling, running and swimming could provide a more complete

picture of training modality distribution, particularly in the winter season when mileage, in any form, is the target.

This visualisation is useful for rowers to track as it can indicate if they are spending too much time in the "other" category currently supported by the platform rather than on the erg or on the water. Furthermore, rowers with less rowing experience might look to optimise time spent on the water to improve their technique.

This visualisation was relatively easy to implement, Plotly.js library being quite lenient with how data could be passed to the donut chart. This visualisation was also generalised to be reusable, so it could be used for both duration and distance data.

Modality Duration and Distance Charts

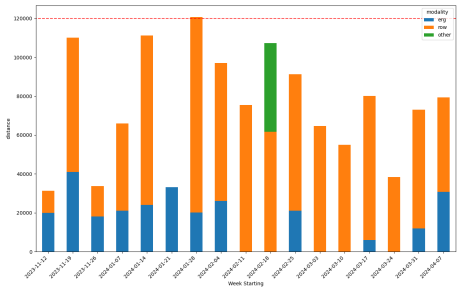


Figure 4.5: A stacked bar chart showing the weekly distance completed across different modalities

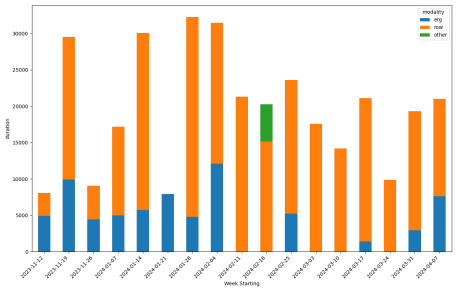


Figure 4.6: A stacked bar chart showing the weekly duration completed across different modalities

Full-sized versions of these charts are available in the appendix:
Distance vs Modality and Duration vs Modality

The stacked bar charts in Figure 4.5 and Figure 4.6 show the weekly distance and duration completed across different modalities since the second week of November 2023. This chart only includes weeks where any training was completed, so gaps that would exist as a result of illness or injury are not shown. Similar to the donut charts, this provides rowers with a snapshot of how they are distributing their weekly mileage and time. If a significant difference in how a given week's time and distance is observed, a rower can prioritise a certain modality to balance their training. This is particularly useful for rowers who may be particularly busy, as they can prioritise erg sessions which are typically a shorter time commitment for the same mileage. These visualisations can also include overlays, such as the dotted red line in the duration chart (Figure 4.6). This dotted line delineates the weekly 120km target the researcher has set, again due to injury and illness this target has not been hit during many of the weeks of training shown. Rowers could also select to include a duration target as well.

This visualisation did not make it to the frontend as there were many formatting and styling issues encountered. With more time, this visualisation could be implemented, but it was not a priority for the project.

Chapter 5

Machine Learning Applications

This chapter will discuss the machine learning approach taken in this project. First, a review of the data collected will be completed, this will outline the next steps taken. Next, the machine learning approach taken will be discussed, this will include an introduction to the adapted machine learning goal, identifying injury risks. This will be executed by replicating an approach developed for runners used by Lövdal *et al.* [24], using the data provided by the literature.

5.1 Rowing Data and Training Review

As discussed in Chapter 3, the data collected was limited to a small number of participants. This section will review the data collected, including the data sources used, the data collected, and the limitations of the data collected.

5.1.1 Review of Data Available

When the goals for the project were outlined, it was not clear what data was available, what data would be needed to produce a useful model, or how much data. In an attempt to have a larger set of data, the data collection system was built to require as little additional input from athletes as possible. As a result, the degree of context available for the data collected was limited. For example, session goals were not outlined, making it difficult to determine if a session was a hard session or an easy session. Additionally, gaps in training data could be present as a result of injury, illness, or athlete vacations. There was no way of collecting this information to discard these portions of data or account for gaps. Finally, most data was collected throughout the winter season. This part of a rower's yearly cycle is notorious for long periods of low-intensity training, with very few options to test performance. No clear performance "criterion" were recorded, making it difficult to even determine if a model was successful in predicting performance.

5.1.2 Initial Machine Learning Plan

Based on existing performance prediction literature, and before it became clear building a performance model with the data available was not feasible, an initial plan was formed to develop a model to predict performance. First, existing models would be adapted to rowing and implemented using the data collected. This would begin with implementing the more basic Banister fitness-fatigue model using the process outlined in Morton *et al.* [14]. Then a more complex model, such as PerPot [22], would be implemented and evaluated. These two implementations of performance models would then be used to evaluate the effectiveness of any machine learning-based models developed.

5.1.3 Limitations of the Data and the Plan

When this plan was devised, despite the lack of data available, there was a tentative strategy to combat the relative lack of data. As noted by Churchill [23], who only had three participants, the use of a hybrid ensemble of neural networks was effective in managing the relative lack of data. Edelmann-nusser *et al.* [15] was also able to predict with strong accuracy a swimming performance using data from a single athlete. This optimistic approach to managing a small dataset further proved to be too ambitious. The "relative lack of data" experienced by Churchill still contained 250, 870, and 1107 days of training data for each of the three participants. Meanwhile, Edelmann-nusser *et al.* [15] had 95 weeks of data for a single athlete, including 19 competitive performances. The data collected for this project was significantly less than these examples, with only 8 participants actively providing data throughout the data analysis stage, which only lasted about 3 months. While there were some erg tests completed during this period, some athletes were unable to complete these tests due to injury or illness, while other athletes did not log these efforts in detail, a limitation of the data collection system developed. Even if the data collection pipeline had been running for the full course of this project with the adjustments necessary to have a full set of features, not enough data would have been collected to develop a model to predict performance to any degree of accuracy.

5.2 Developing an Injury Classification Model

The limitations of the initial goal and plan described in the previous section resulted in a pivot to a new goal. Given the limitations of the data collected for training a machine learning model, it became necessary to look for a dataset that could be used to develop a modeling approach that could be applied to rowing data. Fortunately, Lövdal *et al.* [24] explored the use of machine learning to predict injuries using binary classification, with the full dataset published alongside the paper.

This section will outline the steps taken to replicate the model, including the data used, the model iteration and development, the evaluation of the model, and a reflection on its potential applicability to rowing data.

5.2.1 Predicting Injury in Runners

Data Review

The data provided by Lövdal *et al.* [24] the training log of a high-performance Dutch running squad over 7 years (2012-2019) containing 77 middle- and long-distance athletes. These athletes compete at distances between 800m and full marathon distances, the training for both disciplines is largely endurance-focused so training programs are relatively the same. Furthermore, the training program remained the same as the team's head coach remained consistent throughout the 7 years. The dataset has also been pre-cleaned. Injuries were flagged in the dataset by athletes being either unable to start or complete a session. These flags were then considered an injury event if an athlete had been training injury-free for the previous 3 weeks. Healthy events were those where an athlete was fully fit 3 weeks before and 3 weeks after the event day. Events that contained missing or anomalous data were removed from the dataset. Recurring injury events, those occurring within 3 weeks of a previous injury event, were also removed from the dataset.

This cleaning process resulted in a dataset containing 74 athletes, 42,183 healthy and 583 injury events for the day-to-day model, and 42,223 healthy and 575 injury events for the week-to-week model. The number of injuries per athlete ranges from 0 to 35.

It is evident this data set is quite comprehensive, with detailed information about training behaviours, and a clear definition of injury; this data set is ideal for developing a model to predict injury risk. Furthermore, with many of the same potential features as a complete rowing dataset would contain, the approaches used to predict injury risk in this dataset could be applied to rowing data.

Model Development

There are many approaches which can be taken when developing a classification model. First, an understanding of the problem is crucial. Different classification models and algorithms are better suited to different kinds of problems or expected outcomes. For example, a decision tree model is better suited to problems where the relationships between features are non-linear, while a logistic regression model is better suited to problems where the relationships between features are linear. Given the non-linear relationships between training features and injury risk, and the importance of understanding feature ranking (ie. what behaviours lead to higher injury risk), a decision tree model approach is best suited for this problem. When using a decision tree approach, using ensemble learning (ie. an

approach that combines multiple models to produce the most accurate result), is also recommended. This is because decision trees are prone to overfitting, and ensemble learning can help to reduce this risk and illicit more general results.

Based on the considerations above, the Extreme Gradient Boosting, or XGBoost, algorithm was used. This decision tree boosting system has performed well in several classification problems [24], and is well suited to the problem of predicting injury risk.

The next step in developing the model was to select training and validation data. Given the unbalanced nature of injury data in training, two steps were taken to ensure the model was trained on balanced data. First, a balanced subset of the data was selected at random, with replacement, from the training set. Next, a bagging approach was used to train the model on the balanced subsets of the data. This approach was used to ensure the model was not biased towards predicting healthy events and was able to predict injury events with a high degree of accuracy.

When implementing the XGBoost algorithm, several parameters are passed to the classifier. The only parameters explicitly passed to the classifier are the learning rate, the maximum depth of the tree, and the number of trees, with the remaining parameters being left at default values. A low learning rate of 0.01 was passed, a choice of $2, 3$ was selected for the maximum depth of the tree, and the number of trees $256, 512$ was selected. Lövdal *et al.* [24] did not fine-tune any of the participating models, only adjusting the values described.

The final step of making a usable model was to perform some calibration. This step transforms the output of the model so that a risk score can be generated for each athlete event. This risk score can then be used to determine the likelihood of an injury event occurring. This was done by fitting a logistic regression model – using `CalibratedClassifierCV` from the `scikitlearn` package – to the output of the XGBoost model. This model was then used to generate a risk score for each athlete event. Finally, a threshold needed to be selected to classify a predicted event as an injury event or not. When selecting the threshold value, it is important to consider the cost of false positives and false negatives. In this case, a threshold value that results in more false negatives (ie. predicting an injury event when there is none) is preferred, as the cost of a false positive (ie. not predicting an injury event when there is one) is higher. Practically, taking a day off training to mitigate a potential injury is less costly than training resulting in an injury. An injury will result in more days taken off, likely a longer recovery period, and as a result a loss of performance.

Model Evaluation

Evaluating the accuracy and effectiveness of a classification model is arguably the most important step in developing the model. When evaluating this model it is important to understand what the model is predicting, and how well it is predicting it. To evaluate this model, the area under the receiver operating characteristic curve (AUC-ROC) was used. This metric is a measure of how well the model can distinguish between injury and healthy events. The AUC-ROC ranges from 0 to 1, with a value of 0.5 indicating the model is no better than random, and a value of 1 indicating the model is perfect. In sports science, an AUC of 0.7 is considered strong evidence of a model's predictive ability [24].

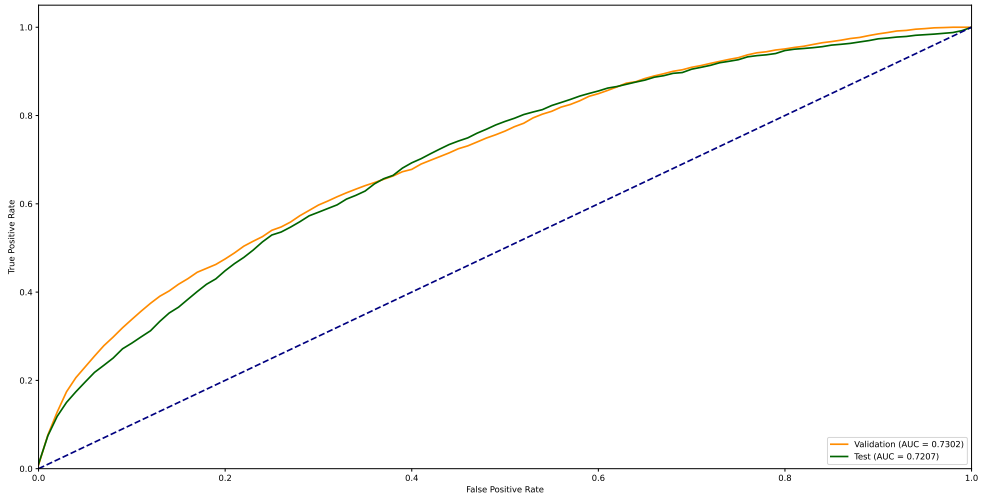


Figure 5.1: The receiver operating characteristic curve for the runner injury classifier trained on daily data leading into the injury event.

For this implementation of the XGBoost Classifier, both approaches taken by Lövdal *et al.* [24] were implemented, the day-to-day model and the week-to-week model. The day-to-day model was trained on data considering the 7 days leading up to an injury event, while the week-to-week model was trained on data considering the 4 weeks leading up to an injury event. The ROC curves for the day model and week model can be seen in 5.1 and 5.2 respectively. The average AUC values for the day model were 0.73 and 0.72 for the validation and test data, while the average AUC values for the week model were 0.79 and 0.71 for the validation and test data, respectively. The consistency between the observed AUC for both models demonstrates the ability of each model to generalise, and new (unseen) data is correctly classified.

To further evaluate the models, reviewing the specificity and sensitivity of the model is important. The specificity of a model is the proportion of true negatives that are correctly identified, while the sensitivity of a model is the proportion of true positives that are

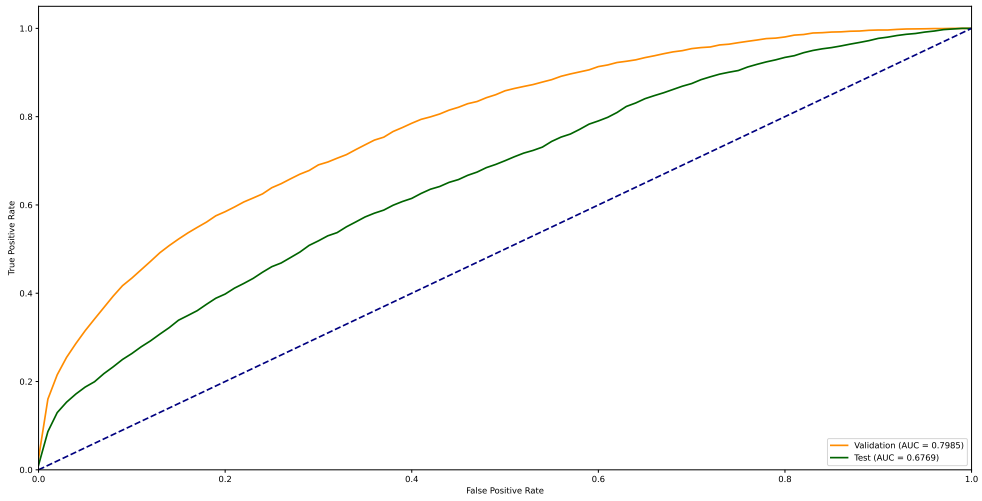


Figure 5.2: The receiver operating characteristic curve for the runner injury classifier trained on weekly data leading into the injury event.

correctly identified. Simply put, the specificity ratio shows: of all healthy events predicted, how many were correct, while the sensitivity ratio shows: of all injury events, how many were identified. A breakdown of the results can be seen in Table 5.1. Based on these results, the day model performs better. This is likely due to the accumulated impact of acute training load on injury risk, which is more accurately captured by the day model. The week model, on the other hand, may not capture the acute training load as well, leading to a lower sensitivity and specificity. The day model has a noticeably higher AUC, indicating better potential performance on new data.

Finally, as a result of using a decision tree-based approach for classification, a feature ranking for each model was generated. This feature ranking can be used to identify which features are most important in predicting injury events, and can be used to inform future training programs, this is available in the appendix (Figure A.1, Figure A.2).

Table 5.1: Mean (SD) Test Scores Obtained by 5 Experiments, With a Threshold of 0.445 for the Day Approach and 0.462 for the Week Approach

Approach	Specificity	Sensitivity	AUC
Day	0.746 (0.02)	0.564 (0.07)	0.73 (0.002)
Week	0.723 (0.01)	0.545 (0.04)	0.68 (0.006)

5.2.2 Applying the Model to Rowing Data

Based on the evaluated effectiveness of the above model on running training data, the potential application for rowing data is promising. Furthermore, a brief anecdotal analysis of the feature ranking for the more effective day model shows that the most important features are related to the acute training load. The ability of this model to uncover complex relationships between training features and injury risk is promising for the application of this model to rowing data. Finally, the training data available to be collected from rowing training is quite similar to the training data collected from running training, with the same potential features available. This model could be applied to rowing data with minimal changes and could be used to predict injury risk in rowers.

Chapter 6

Discussion

This project provided many opportunities for learning and growth and, now, at its conclusion, there is an opportunity to reflect on the proposed goals of the project and its subsequent execution. Firstly, the initial target of developing a comprehensive machine learning model for rowing performance was quite ambitious. This became clear in January when a data collection system was still being developed. Furthermore, the data collection goal of minimal user input directly prevented the development of a meaningful model. Nevertheless, the project was still successful in developing a data collection system, and a data cleaning pipeline. This then facilitated providing analysis and visualisations to the athletes who kindly volunteered their training data. Additionally, there was a successful exploration of a potential machine learning approach using a more complete data set which demonstrated a strong ability to produce an injury risk score and predict injury. This foundational work could serve as a basis for future development of an effective model.

This chapter will discuss and evaluate the methods used for data collection, management, cleaning, analysis and visualization. It will also discuss the potential for future work, particularly in developing a machine learning model for predicting athlete performance. Finally, a protocol will be outlined for the effective implementation of a system to develop and deploy a performance model, based on the knowledge gleaned throughout this project.

6.1 Data Collection

6.1.1 Evaluation

The data collection system, based on the goals outlined at the start of the project, was successful in collecting data from several athletes, across four different data sources. Roughly 40 athletes were approached to provide data. Of that number, 12 athletes signed

up, with around 8 consistently providing data and feedback on developments. This drop-off between the number of athletes registered for the platform and providing data was due to injuries or illness, which prevented consistent training. The data collection system was effective in providing a user-friendly platform for athletes to provide data with the help of the data provider's APIs.

In total, since athlete recruitment was completed at the end of January, roughly 500 activities were ingested through the data collection pipeline. This number was lower than expected due to a recent, and poorly documented, change by Strava in how developer apps are handled. This meant that roughly 6 weeks of training data for all but one user were not shared with the application. As a result, many user data sets were somewhat incomplete and there was not enough time to fully remedy this issue by backfilling athlete data. Analysis for these users was supported by data provided from Concept2, Polar, and, in the last 4 weeks, Garmin. The researcher's data set was preserved and used for developing the analysis and visualisation steps of the project.

6.1.2 Discussion

The data collection approach, while effective in minimising the effort of individual athletes, made the development of any kind of performance model difficult. The advertised automation of data collection made it easier to recruit athletes, both due to the minimal effort required to participate in the project and because of the clear security by obfuscation for competitors. This did, unfortunately, result in a dataset which was not as complete as it could have been. The lack of heart rate data for many sessions, and the lack of detail when recording strength sessions, limited the analysis which could be performed. In particular, the lack of heart rate data made it difficult to classify trainings as an endurance or interval session and also to calculate training impulse. For further advancement of this research, a more involved group of participants would be required. Athletes who are willing to dedicate the time to provide feedback for sessions, and ensure their training logs are complete, make it easier to develop a model which can be used to predict performance. These individuals perhaps facilitate the low-effort data collection and analysis approach this project strove for.

There were also many different training plans being executed by participants, which made it difficult to develop a model which could be used to predict performance. A more consistent training plan, or a more detailed analysis of the training plans being executed, would be necessary to develop a model which could be used to predict performance. If it were possible to onboard an entire squad or have participants commit to completing a specific training session each week, there would be a clear metric for improvements, or decline, in fitness and therefore a baseline by which to compare an athlete's performance. This will be further

discussed in the final section of this chapter, section 6.4.

The data collection method developed throughout this project was successful in providing an easy, and secure, way for athletes to provide training data. However, the data collected was not as complete as it could have been. This limited the analysis and model development which could be performed. This is an easy opportunity for future work. Introducing session matching and a more detailed onboarding and feedback process, could help to ensure the data collected is more complete and useful for analysis. This also introduces the opportunity to develop daily monitoring surveys for athletes to complete to give a more comprehensive picture of their training, recovery, and general well-being, which is beneficial for more effective analysis.

There is indeed an opportunity for further exploration, particularly in the integrated tracking and management of athletes. Specifically, a system could be developed that benefits both coaches and athletes by providing useful data and feedback without overwhelming them. For example, a system could generate insights about an athlete's sleep and steady-state session enjoyment. While this may be an interesting metric, it is not particularly useful and would not help the athlete or coach improve training. Therefore, it would not be included. Meanwhile, an insight such as steady-state zone and the following day's recovery could prove useful and help an athlete maintain a sustainable training volume. This insight would be included on the dashboard. Originally, the emphasis on data collection was not prioritized in this project. However, there is a clear direction for future work to be explored on how technology can support training and athlete management.

6.2 Data Analysis and Visualisation

6.2.1 Evaluation

The data analysis and visualisation step of this project was successful in providing concise feedback to rowers on their training. First, the data cleaning step effectively cleaned and transformed data from multiple data sources and formats into one concise and uniform format allowing more efficient analysis and visualization. Next during the analysis step, data was analysed to provide appropriate and useful feedback and insights on training habits. This analysis was guided by metrics typically used by rowers to define a successful training week. These included reducing the amount of time and effort needed for rowers to understand their training, allowing rowers to identify trends in training behaviours, and easily identifying minor improvements to help motivate future training. Finally, the visualisation step successfully provided rowers with a way to understand the analysis graphically. The presentation of this information was particularly important to provide athletes with an at-a-glance overview of their training outcomes and habits, and this generation of visualisations was largely

successful in completing this objective. This approach to presentation, allowed rowers to quickly identify trends in their training and make adjustments as necessary.

Each of these sub-steps was developed in a way that allows them to be dynamically called through serverless functions, allowing them to be run automatically when new training events were ingested by the data collection pipeline.

6.2.2 Discussion

The data analysis and visualisation portion of this project was the most straightforward part of the entire project. This was made possible due to a variety of reasons. First, the data cleaning pipeline was developed to transform data into a standard format, which made analysis and visualisation significantly easier. Second, the analysis and visualisation steps were developed to provide feedback on the most basic metrics which rowers use to define a successful training week. This meant that there was no minimum amount of data required to begin exploring potential analyses. Furthermore, due to the researcher's background in rowing, the metrics used to define a successful training week were already well understood and the approach to developing an analytics pipeline had already been considered before the commencement of this stage of the project. Finally, inspiration for visualisations was drawn from other researchers' work in sports science, particularly in rowing, which made it easier to develop useful visualisations for rowers.

The primary limitation of the data analysis and visualisation step was the lack of data available for analysis. Although analysis of the data available was largely successful, the lack of live heart rate data in many cases made it difficult to calculate training impulse and provide more detailed feedback on training sessions. This is a clear area for future work, and the development of a more complete data collection system, as discussed in the previous section, would help to alleviate this issue. Additionally, the lack of contextual data, such as a rower's injury or illness status, made it difficult to provide feedback on why a rower may have missed a session, or why they may have performed poorly in a session. This is another area for future work, and the development of a daily monitoring survey, as discussed in the previous section, would help to provide this context.

One of the target analyses where this project fell short was comparing "like" sessions. For example, comparing steady-state (endurance) sessions throughout a season. These endurance sessions can vary in distance and target intensity. Some athletes may target a session as a UT1 (higher intensity, typically lower volume) rather than a UT2 (lower intensity, typically higher volume) session. Without the additional context of the session, provided by the user through a note or an RPE rating, it is very difficult to identify the exact type of session, making matching and comparison of these sessions impossible. With adjustments to the data collection pipeline, this would be possible, allowing athletes to more

easily identify trends in their training, and make adjustments as necessary.

A variety of visualisations were generated for the project as well. As described in section 4.3, the visualisations were first developed using a Jupyter Notebook to more interactively work with the data. The content of the Jupyter Notebook was then condensed to produce the pipeline that was used to generate the appropriate datasets for use on the frontend. Unfortunately, when converting the graphs and visualisations developed using Python into a format that could be easily displayed on the frontend was not as straightforward as initially thought. This was due to the limitations of the Plotly library, which was used to generate the visualisations. The documentation for the library was quite sparse, and a huge amount of effort was spent to ensure that data was reformatted correctly and styling was applied appropriately. In the end, three of the four visualisations were implemented in the frontend, with the modality duration and distance charts not being implemented due to styling issues. This is a clear opportunity for future work, and could likely be a full project in itself. Exploring how athletes interact with data, and what visualisations are most useful to them, could provide a wealth of information for future development of a data analysis and visualisation pipeline.

6.3 Machine Learning

The machine learning goals of this project were not fully realised. The initial goal of developing a machine learning model to predict rowing performance was overly ambitious, given the data and time available. However, the exploration of a machine learning model to predict injury in runners was successful. This chapter will evaluate the methods used for developing the machine learning model, and discuss the potential for future work in developing a machine learning model for rowing performance.

6.3.1 Evaluation

Despite the data collection and availability issues, an alternative approach to exploring machine learning was developed. This approach was based on the work of Lövdal *et al.* [24], who developed a machine learning model to predict running injuries. This approach was successful in developing a model which could predict injury events in runners with a specificity of 74.6% and sensitivity of 56.4%. These rates would be acceptable to predict injury risk and inform training decisions for high-risk athletes and sessions.

6.3.2 Discussion

The largest potential for future work is within the machine learning applications. Assuming enough data has been collected, through the development of an effective data collection

system. The ability to apply the processes demonstrated in this project to rowing training data, given the similarity in the expected data collected, is straightforward. This model can then be offered through the data collection dashboards to provide coaches and athletes with a tool to gauge injury risks and make appropriate changes, if necessary, to manage potential injuries.

A novel approach will need to be used for predicting performances, such as expected 2km erg test times. Future advancements for developing this model should look at first developing a method to obtain an appropriately large, and detailed, dataset. Then using the artificial neural network approaches explored by Edelmann-nusser *et al.* [15] and Churchill [23], a model could begin to be developed and refined. The potential approach to how this process might be developed will be discussed in the next section.

6.4 An Ideal Protocol to Develop an Effective Model

In order to successfully develop an effective performance model, several steps need to be taken. The following protocol outlines the steps which could be taken to develop, validate, and leverage a performance model which could be used to predict, and prevent, athlete burnout and injury. This protocol is based on the insights and the research conducted throughout this project.

6.4.1 Squad Onboarding

Based on the difficulty in recruiting participants described in Chapter 3, collecting long-term training and recovery data from a large number of athletes necessitates participation from all members of senior-level squads starting with senior leadership. Approaching coaches and club administrators means that assurances can be provided concerning data protection and privacy. This would ensure squad training is not leaked to other competitors and establishes trust which can trickle down to each athlete. With this consideration, the number of eligible squads for a partnership is limited to university squads, high-performance groups in open clubs, or national teams. These squads have athletes training nearly year-round following a strict training plan. Furthermore, any squad which is onboarded needs to have complete buy-in from the coaching staff and athletes. This is necessary to ensure that all athletes are providing data and that this data is complete and accurate. It will also ensure collaboration between the research team and the coaching staff, meaning a training plan can be developed with the necessary procedures. A strong relationship between the research team and the partner squad will allow for more frequent and effective iterations of the athlete and coach dashboards and apps. With potential improvement in how the analytics and statistics are collected, generated and visualised. If future endeavours progress as anticipated, models for performance prediction, injury prevention, and more could be established.

Once a squad has been identified and onboarded, each rower should complete a series of physiological tests, this will include lactate and VO_2 max tests (much like what is described in subsection 2.1.1), strength tests such as the maximum deadlift and bench press, power tests such as a countermovement jump, resting metabolic rate test, and body composition tests. This will provide rowers with more accurate zones to train in, based on both heart rate and wattage, and provide context for a rower's physiology. These tests should be completed at least once per season where possible. Sub-maximal versions of the tests can be completed to dial in steady-state training zones as each season progresses or when recovering from illness or injury.

Squad onboarding will also include registering all coaches and rowers to an online system. Through this system, coaches will be provided an overview of their athlete's training adherence, recovery metrics, and general well-being. There would also be tools to generate or adapt training programs, build crews for on the water sessions and otherwise manage their squads. Rowers will be able to see their scheduled sessions and track their training and well-being through the platform, with suggestions for personalisation based on their recovery and these metrics. These features need to be implemented to encourage engagement from athletes and coaches, by providing useful tools to track, manage, and analyse training and also collect as much data as possible for machine learning model training. The potential analyses available as a result of collecting and collating each of these metrics can further support improvements in training behaviour, crew selection, and more. As has already been extensively discussed, providing a potential model with as much training data, and context, as possible gives the model the best chance for success in predicting performance and potential injuries, illness, or overtraining, thereby, encouraging consistent and productive training.

6.4.2 Data Collection

With a squad onboarded and initial rower data collected, the next step is to develop a data collection system. There are two main parts to consider for this system: recovery and wellness tracking, and training tracking.

Recovery and wellness tracking

Every morning athletes should complete a monitoring questionnaire, this questionnaire will include questions about sleep quality, sleep duration, muscle soreness, fatigue, stress, and mood. This questionnaire can be supported by data from wearables such as a Whoop, Oura ring, or a smartwatch. These devices provide heart rate metrics (resting heart rate and heart rate variability) which can be analysed to determine recovery and readiness to train. Rowers can also note if they are unfit to train due to illness or injury. Providing this context is crucial for explaining gaps in training data when in the analysis step.

An additional metric to track could be nutrition, which could also be built into the system. A nutrition tracking system would need to track what food was eaten, how much food was consumed, and at what time. This data can be used to determine if an athlete is eating enough to support their training, and if they are eating the right foods to support their training.

Training tracking

While automated tracking and analysis of sessions would be an ideal scenario for many athletes, to effectively train a model, it is important to provide context for sessions which have been completed. An ideal training tracking solution would leverage session data provided by APIs from Concept2, Strava, Garmin, and Polar, and provide a simple interface for athletes to provide feedback on their sessions. Through the online platform, rowers would provide feedback on sessions, including how they felt during the session, how they felt after the session, and any other notes they might have. If athletes have augmented a prescribed session for any reason, this can be noted as well and feedback can be relayed to the coach. This feedback is essential for developing a model which can be used to predict performance and identify burnout to prevent injury and illness.

Strength sessions can also be prescribed by the coaches platform, and then tracked through the rower platform. By logging lifts used, the number of reps and sets completed, and tonnage moved, a coach can track the progress of their athletes, and provide feedback on their strength training. This will also provide context for the strain experienced during strength sessions, and provide a more complete picture of the training load experienced by an athlete each week.

6.4.3 Analysis and Visualisation

With data collected from a squad, the next step is to develop a data cleaning pipeline, and analysis and visualisation tools. The data cleaning pipeline can be iterated from the one developed for this project. Given the increased athlete feedback, session classification will be much more straightforward, allowing for more effective analysis. With a more complete dataset of heart rate and RPE data, alongside more detailed strength session information, the analysis and visualisation tools can be developed to provide more effective feedback to athletes and coaches. Additionally, due to the addition of recording recovery data, athletes and coaches can explore the impact of different session intensities, duration, and modality on recovery and wellbeing metrics. They can also adjust squad or personal training as a result of that feedback. For example, at the start of the season, an athlete may have lost fitness as a result of time off. Within six weeks they may see a decrease in resting heart rate as a result of beginning training again. Conversely, if a trained athlete begins to see their resting heart rate trend upward, this could be an indicator of overtraining or illness and the

need for a rest period. This basic feedback can be used to adjust training plans and prevent burnout and injury.

Algorithms can also be developed to suggest different training sessions based on the feedback provided by athletes. For example, if an athlete is feeling fatigued, a session with a lower intensity can be suggested. If an athlete is feeling good, a session with a higher intensity can be suggested. These suggestions can be generated as a result of morning monitoring results. Additionally, if a coach wants to maintain a certain training load, sessions can be suggested with options to account for time constraints or perceived intensity.

6.4.4 Model Development and Validation

With a complete dataset, and effective analysis and visualisation tools, the next step is to develop a performance model. There will be several iterations of models. The first, and likely easiest, to develop would be a model to identify injury or overtraining. Leveraging the methods described in Chapter 5 in replicating the prediction or running injuries [24], a similar approach can be applied to rowing training. Ideally, considering the volume of training and recovery data collected a more effective model can be developed. This can be used to help prevent injury and illness and provide feedback to athletes and coaches on how to adjust training plans to prevent burnout.

Next, to build the holy grail: a model for rower performance. This model will be developed iteratively, with feedback from athletes and coaches. The standard weekly training session, outlined in subsection 6.4.1 can be analysed for changes in fitness, based on heart rate and wattage metrics. With these reference points for athlete performance, a model can be trained to predict performance based on the training load experienced by an athlete. With enough training and recovery data, this can then be used to build training plans for an entire macro-cycle and applied to individual athletes based on their performances each season.

Based on the success of this model, a model for crew performance can be developed. This model will be based on the performance of individual athletes, and how they perform in a crew. This model can be used to build crews for on the water sessions and provide feedback to coaches on how to adjust crews to improve performance. This model will be developed iteratively, with feedback from athletes and coaches, and will be based on the performance of individual athletes in a crew, and how they perform in different crew combinations.

6.4.5 The Potential

Assuming the success of this approach over 5-10 years, a truly data-driven approach for coaching and training can help a squad find success and look to optimise every part of

training and recovery. The only limitation to success is how much data is available. With progress in the development of hardware to track on the water power and the development of a more complete data collection system, the potential impact is significant. Athlete development and training can be tuned to prevent injury and illness, which is one of the biggest hindrances for athletes in their pursuit of success. Furthermore, crew selection can be made more data-driven and transparent, identifying the most physiologically capable athletes, and the most effective crew combinations. This approach can then be applied to other endurance sports, like cycling and running very easily. With additional adaptation and research to quantify training load, applications in field and combat sports are also possible.

The potential for further research and development in applying machine learning to sports and human performance is vast and exciting. Developing and providing these tools can free up time and energy for coaches and athletes to allow them to engage with and enjoy their sports more. This project has only scratched the surface of what is possible, and with the right approach, and the right data, the potential for success is limitless.

Chapter 7

Conclusion

This chapter provides the conclusion of the project, with a review of objectives set out in Chapter 2.

7.1 Objectives Completed

This project aimed to explore the use of machine learning to predict rowing training outcomes and performances. The project was largely successful in its data collection and analysis goals but was unable to develop a model to predict performance. This section will include a synopsis of the work completed and a summary of future work which could be completed.

The project successfully collected data from rowers through a web application developed throughout the project. This application facilitated the collection of data with no extra interaction needed from participants. In total 12 participants were onboarded to the platform with data collected from 8 participants. This discrepancy is due to some participants becoming heavily injured or ill, or suffering other technical issues resulting in being removed from the project.

After successfully developing a data collection pipeline (Chapter 3), a data analysis pipeline was developed (Chapter 4) to provide participants with feedback on their training data. This produced several data analysis metrics, including training load, acute and chronic workloads, daily ACWR, time spent in zones for each session as well as on daily, weekly, and monthly intervals, and summaries of daily, weekly, and monthly training duration and mileage. Finally, data visualisations were drafted using Matplotlib in Python and then developed for the frontend using Typescript and Plotly.js. These visualisations were used to provide feedback to participants on their training data by presenting the analysis completed in a digestible format.

While the project aimed to produce a model to predict performance, it became clear that this was an overly ambitious goal, due in part to the limited time for the project, but also the limited data collected as a result of the automatic data collection goals. Instead, the project explored the use of machine learning to predict injuries in runners, noting the approach and its application to appropriately collect rowing data.

This exploration into the use of machine learning to predict injuries in runners was largely successful. The project replicated the approach taken by Lövdal *et al.* [24] to predict injuries in runners. This involved developing a model to predict injury risk using a binary classification model. The model was developed using the data provided by the literature, and the model was evaluated using the same metrics as the original paper. The applicability of this approach to appropriately collected and tagged rowing data is significant due to the similar training nature of both sports producing similar data.

7.2 Final Thoughts

There are some opportunities for future work as a result of this project, as described in Chapter 6.

The most immediate opportunity is to continue developing a data collection model which balances the need for minimal effort from athletes, with the need for endless data for machine learning training. Further work could then develop a machine learning model to predict injuries in rowers, which could advance to producing performance models. Beyond machine learning, an exploration into how to leverage technology to support coaches in managing their athletes' training and recovery could be a valuable first step. Based on the volume of potential data sources available in rowing, building a framework to collect and analyse data from these sources could provide a more comprehensive picture of an athlete's performance and recovery.

The potential of taking the insights from this project and applying them to other sports is significant. Trying to build integrated recovery and training tracking for athletes alongside team management and analysis can be incredibly useful. This solution can help coaches and athletes understand the impact of training on performance and recovery at an individual level.

To conclude, this project was largely successful in its data collection and analysis goals but was unable to develop a model to predict performance. This project also produced many insights and learning opportunities which can be applied to future work to deliver more effective feedback to athletes and coaches across numerous disciplines.

Bibliography

[1] E. W. Banister, T. W. Calvert, M. V. Savage, and T. Bach, "A systems model of the effects of training on physical performance," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 2, pp. 94–102, Feb. 1976. DOI: 10.1109/tsmc.1976.5409179.

[2] J. Mäestu, J. Jürimäe, and T. Jürimäe, "Monitoring of performance and training in rowing," *Sports Medicine*, vol. 35, no. 7, pp. 597–617, Jul. 2005. DOI: 10.2165/00007256-200535070-00005.

[3] 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, ISSN: 0905-7188. DOI: 10.1111/j.1600-0838.2004.00418.x.

[4] M. A. Rosenblat, A. S. Perrotta, and B. Vicenzino, "Polarized vs. threshold training intensity distribution on endurance sport performance: A systematic review and meta-analysis of randomized controlled trials," *Journal of Strength and Conditioning Research*, vol. 33, no. 12, pp. 3491–3500, Dec. 2019. DOI: 10.1519/jsc.0000000000002618.

[5] A. Das, U. S. Kaniganti, S. J. Shenoy, P. Majumdar, and A. K. Syamal, "Monitoring training load, muscle damage, and body composition changes of elite indian rowers during a periodized training program," *Journal of Science in Sport and Exercise*, vol. 5, no. 4, pp. 348–359, Nov. 2022. DOI: 10.1007/s42978-022-00197-7.

[6] 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 and Science in Sports*, vol. 16, no. 1, pp. 49–56, Oct. 2004. DOI: 10.1111/j.1600-0838.2004.00418.x.

[7] S. Selles-Perez, J. Fernández-Sáez, and R. Cejuela, "Polarized and pyramidal training intensity distribution: Relationship with a half-ironman distance triathlon competition," *Journal of Sports Medicine and Science*, vol. 18, no. 4, pp. 708–715, Nov. 2019.

[8] A. S. Göktepe, "Energy systems in sport," in *Amputee Sports for Victims of Terrorism*. IOS Press, 2007, pp. 24–31.

[9] Apr. 2018. [Online]. Available: https://exercise.trekeducation.org/principles/periodisation/.

[10] B. Kayser and G. Gremion, "Chronic fatigue and loss of performance in endurance athletes: Overtraining," *Schweizerische Zeitschrift fur Sportmedizin und Sporttraumatologie*, vol. 52, no. 1, pp. 6–10, 2004.

[11] H. Gustafsson, "Burnout in competitive and elite athletes," Ph.D. dissertation, Örebro Univeristy, 2007.

[12] J. Lawton, *Athlete tapering faqs all endurance coaches need to know*, May 2023. [Online]. Available: https://www.trainingpeaks.com/coach-blog/endurance-coaching-tapering-faqs/.

[13] N. Kanwetz, *The science of supercompensation and how it makes you fast*, Jun. 2016. [Online]. Available: https://www.trainerroad.com/blog/science-of-supercompensation/.

[14] R. H. Morton, J. R. Fitz-Clarke, and E. W. Banister, "Modeling human performance in running," *Journal of Applied Physiology*, vol. 69, no. 3, pp. 1171–1177, Sep. 1990. DOI: 10.1152/jappl.1990.69.3.1171.

[15] J. Edelmann-nusser, A. Hohmann, and B. Henneberg, "Modeling and prediction of competitive performance in swimming upon neural networks," *European Journal of Sport Science*, vol. 2, no. 2, pp. 1–10, Apr. 2002. DOI: 10.1080/17461390200072201.

[16] N. Williams, "The borg rating of perceived exertion (rpe) scale," *Occupational Medicine*, vol. 67, no. 5, pp. 404–405, Jul. 2017. DOI: 10.1093/occmed/kqx063.

[17] M. Kent, *Trimp method*, 2007. DOI: 10.1093/acref/9780198568506.013.7418. [Online]. Available: https://www.oxfordreference.com/view/10.1093/acref/9780198568506.001.0001/acref-9780198568506-e-7418.

[18] R. White, *Acute:chronic workload ratio*, Oct. 2023. [Online]. Available: https://www.scienceforsport.com/acutechronic-workload-ratio/.

[19] F. M. Impellizzeri, M. S. Tenan, T. Kempton, A. Novak, and A. J. Coutts, "Acute:chronic workload ratio: Conceptual issues and fundamental pitfalls," *International Journal of Sports Physiology and Performance*, vol. 15, no. 6, pp. 907–913, Jul. 2020. DOI: 10.1123/ijspp.2019-0864.

[20] H. Zouhal, D. Boullosa, R. Ramirez-Campillo, A. Ali, and U. Granacher, "Editorial: Acute: Chronic workload ratio: Is there scientific evidence?" *Frontiers in Physiology*, vol. 12, May 2021. DOI: 10.3389/fphys.2021.669687.

[21] F. Imbach, N. Sutton-Charani, J. Montmain, R. Candau, and S. Perrey, "The use of fitness-fatigue models for sport performance modelling: Conceptual issues and contributions from machine-learning," *Sports Medicine - Open*, vol. 8, no. 1, Mar. 2022. DOI: 10.1186/s40798-022-00426-x.

[22] J. Perl, "Perpot: A metamodel for simulation of load performance interaction," *European Journal of Sport Science*, vol. 1, no. 2, pp. 1–13, 2001. DOI: 10.1080/17461390100071202.

[23] T. Churchill, "Modelling athletic training and performance," Ph.D. dissertation, University of Canberra, Australian Capital Territory, Australia, 2014.

[24] 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, Oct. 2021. DOI: 10.1123/ijspp.2020-0518.

Appendix A

Appendix

A.1 Lactate Test Results

Table A.1: Lactate Test 1 Interval Details

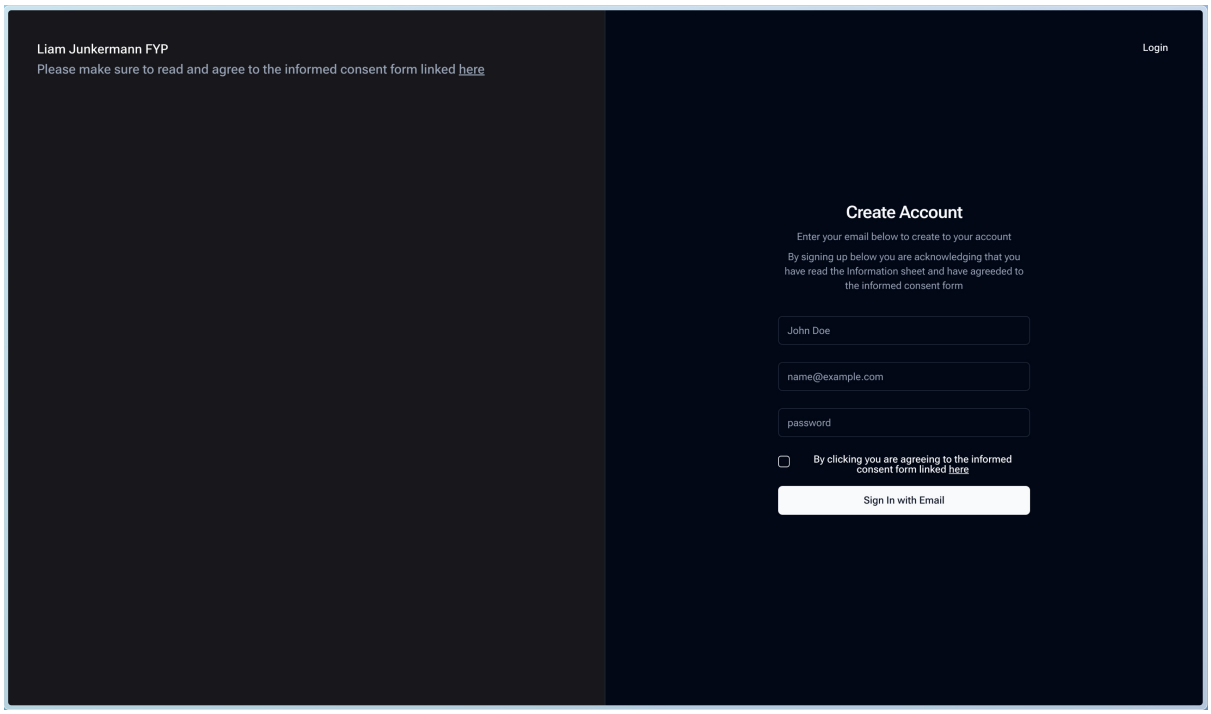
Stage	Target Power (Watts)	Target Time	Relative VO2 (ml/kg/min)	HR (BPM)	RER	RPE	Actual Power (Watts)	Actual Time	Lactate (mmol/L)	Stroke Rate	Distance (m)	Notes
1	140	02:15.9	38.25	132	0.87	1	148	02:13.3	1	17	900	
2	175	02:06.2	39.88	141	0.87	2	178	02:05.3	1.3	19	957	
3	210	01:58.7	56.64	157	0.93	4	212	01:58.1	2.2	19	1016	
4	245	01:52.7	58.86	165	0.94	5	250	01:51.9	3.4	21	1072	
5	280	01:47.8	64.3	175	0.99	6	287	01:46.8	6.4	24	1123	
6	315	01:43.7	68.54	183	1.03	7	323	01:42.7	9.7	29	1168	
7	Max	Max	76.47	189	1.01	9	354	01:39.6	14.8	57	1204	
									16.3			1 minute post
									17.4			2 minutes post

Table A.2: Lactate Test 2 Interval Details

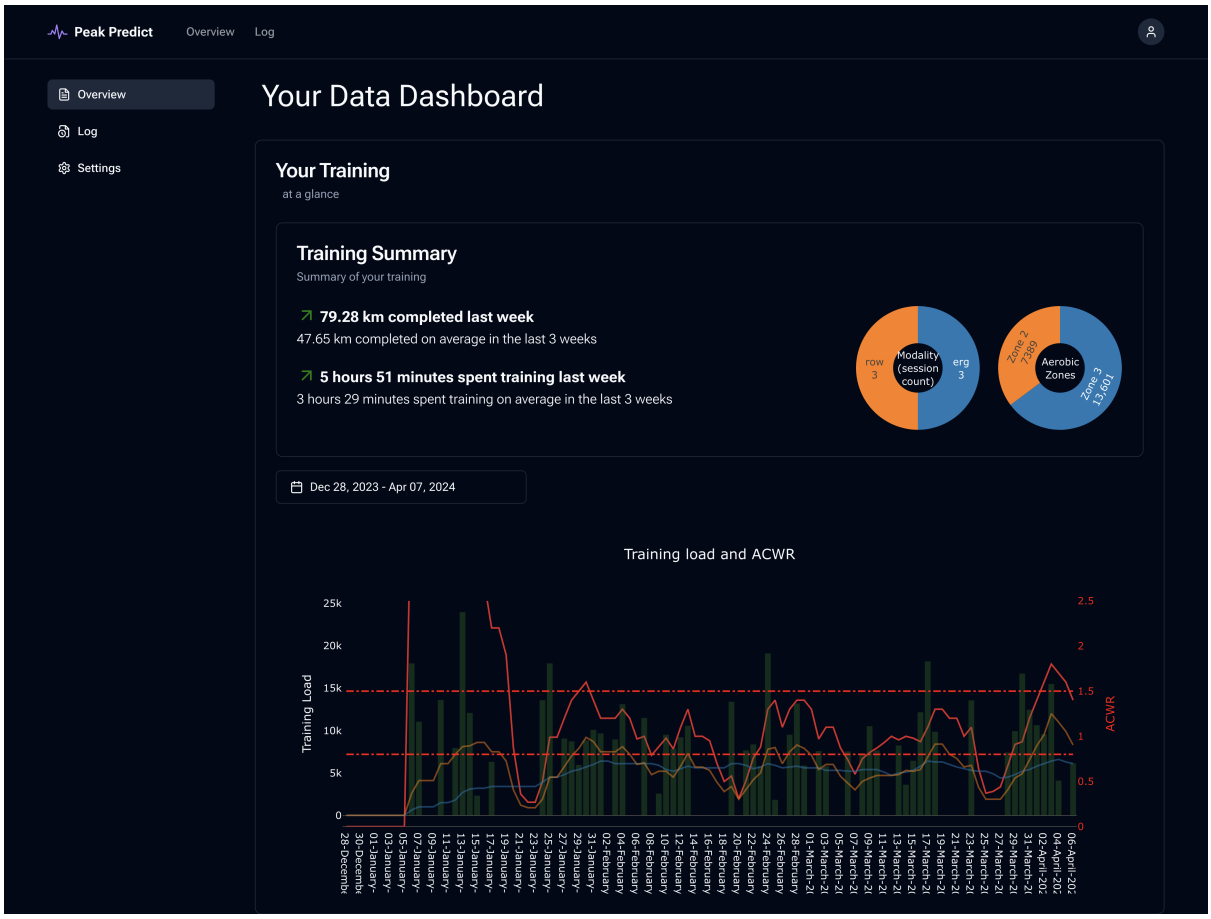
Stage	Target Power (Watts)	Target Time	Relative VO2 (ml/kg/min)	HR (BPM)	RER	RPE	Actual Power (Watts)	Actual Time	Lactate (mmol/L)	Stroke Rate	Distance (m)	Notes
1	140	02:15.9	35.1	130	0.82	0	143	02:14.8	0.8	16	890	
2	175	02:06.2	41.9	142	0.81	1	178	02:05.2	1	17	958	
3	210	01:58.7	47.8	153	0.88	3	212	01:58.2	1.3	20	1015	
4	245	01:52.7	54.7	162	0.88	4	247	01:52.2	2	22	1069	
5	280	01:47.8	59.8	172	0.92	5	285	01:47.1	3.7	24	1120	
6	315	01:43.7	64.2	178	0.99	6	318	01:43.2	6.2	28	1162	
7	Max	Max	74.4	192	1.1	10	398	01:35.7	13.4	43	1253	
									13.3			2 mins post

A.2 Data Collection Website Screenshots

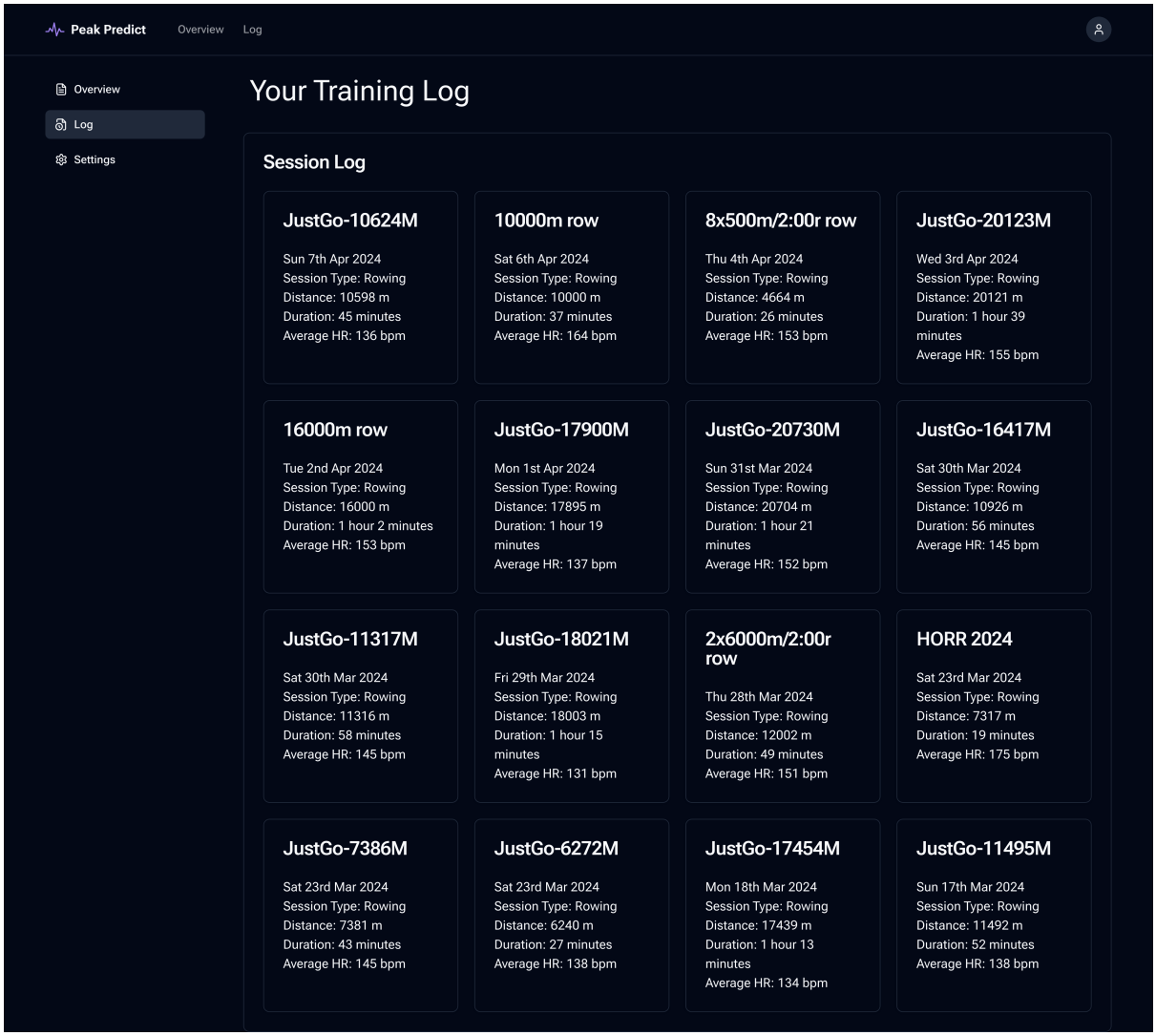
Full-size screenshots of the data collection website can be found below.



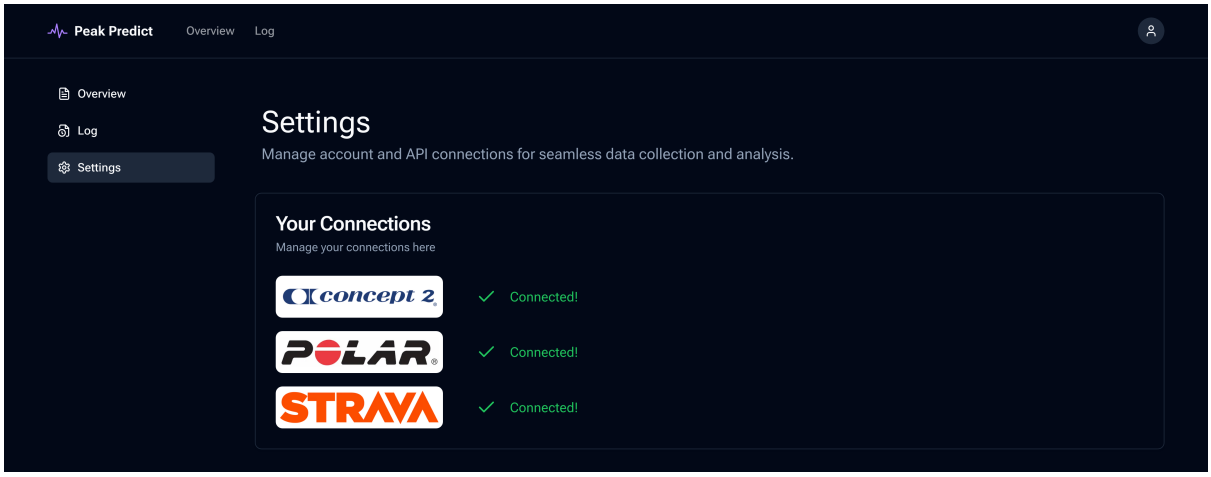
The user registration screen for the web app



The dashboard screen for the web app

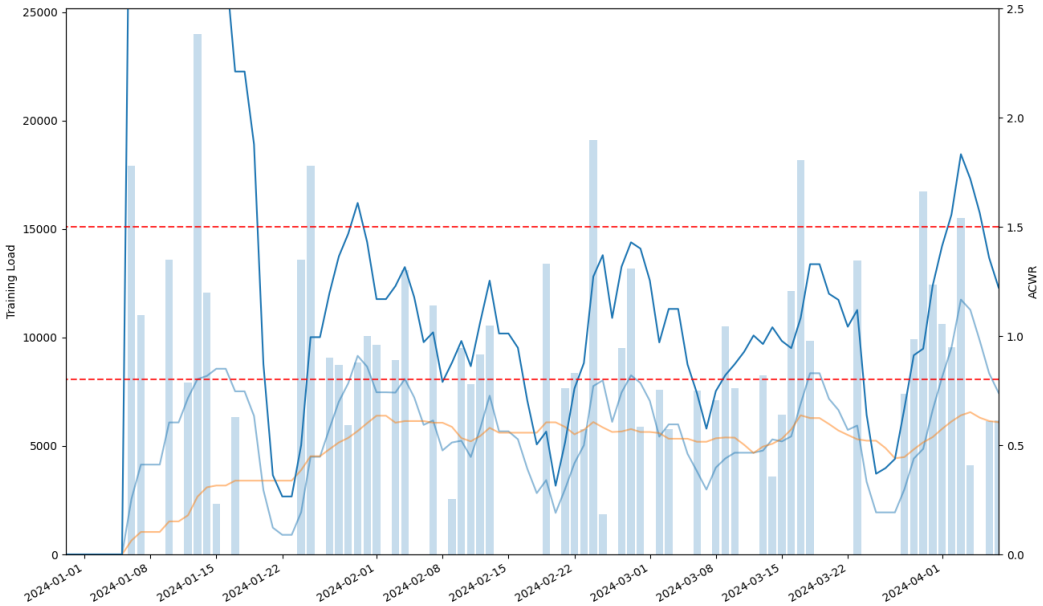


The training log screen for the web app with 16 sessions loaded



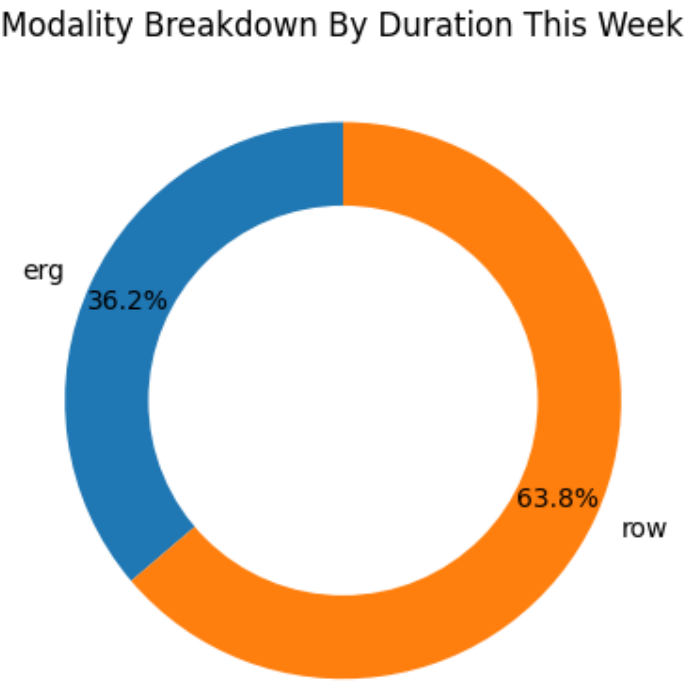
The connections manager screen for the web app

A.3 Visualisations

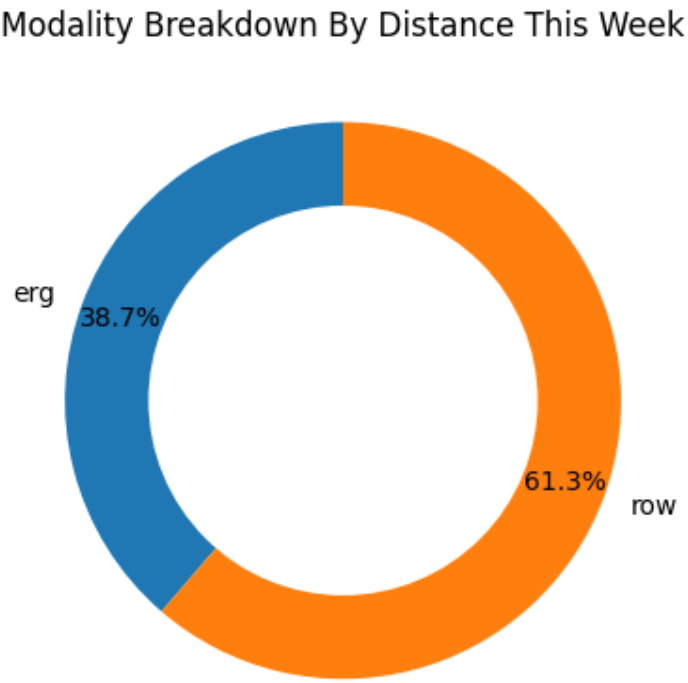


The Acute, Chronic, TRIMP, ACWR chart

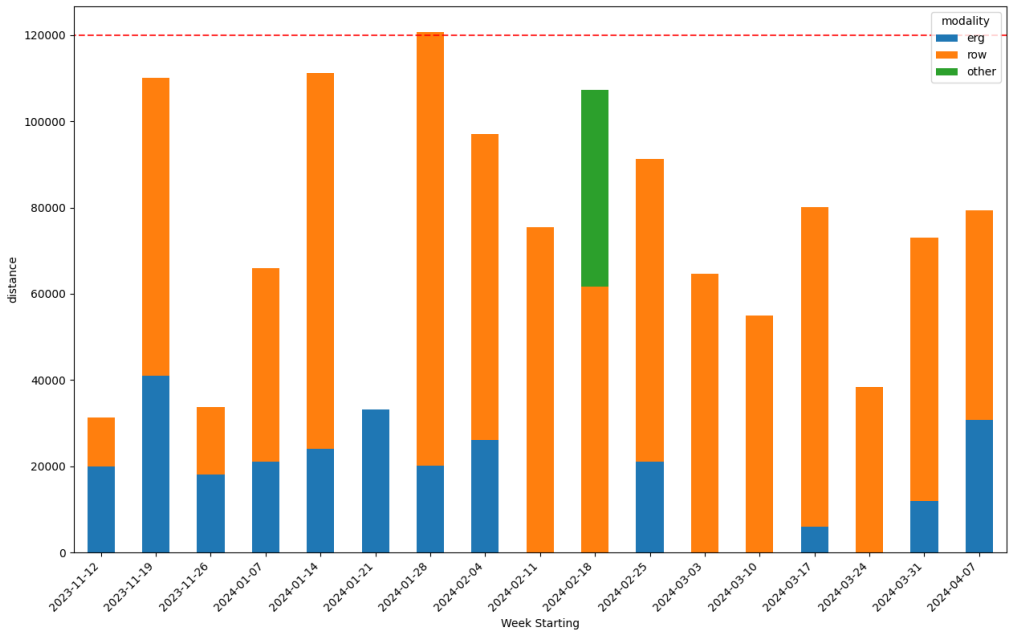
The weekly training overview widget, comparing the last week's mileage and training time to the average of the previous 3 weeks



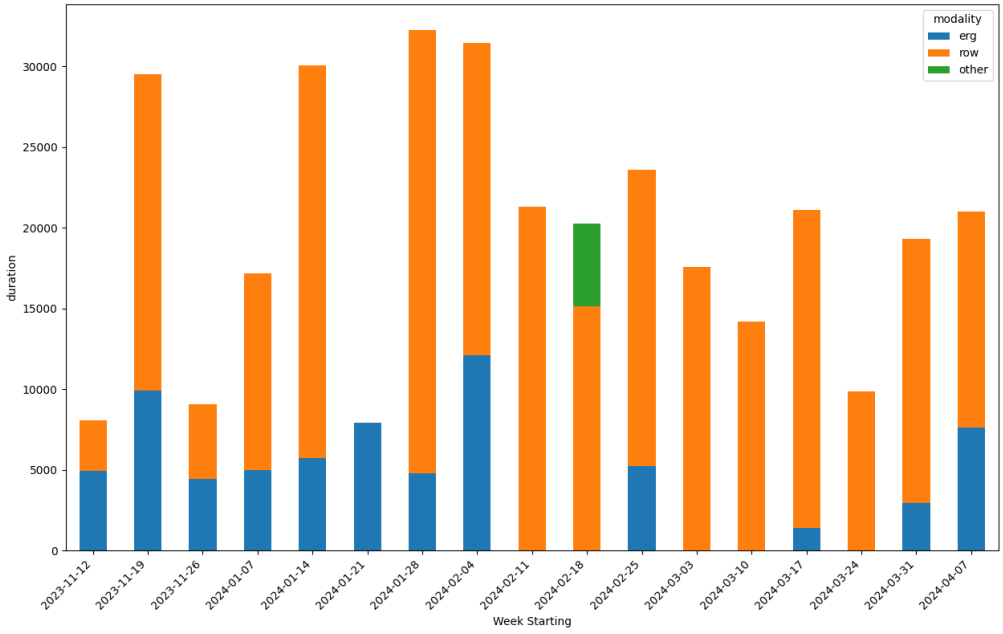
A donut graph showing the distribution of training duration across different modalities



A donut graph showing the distribution of training distance across different modalities

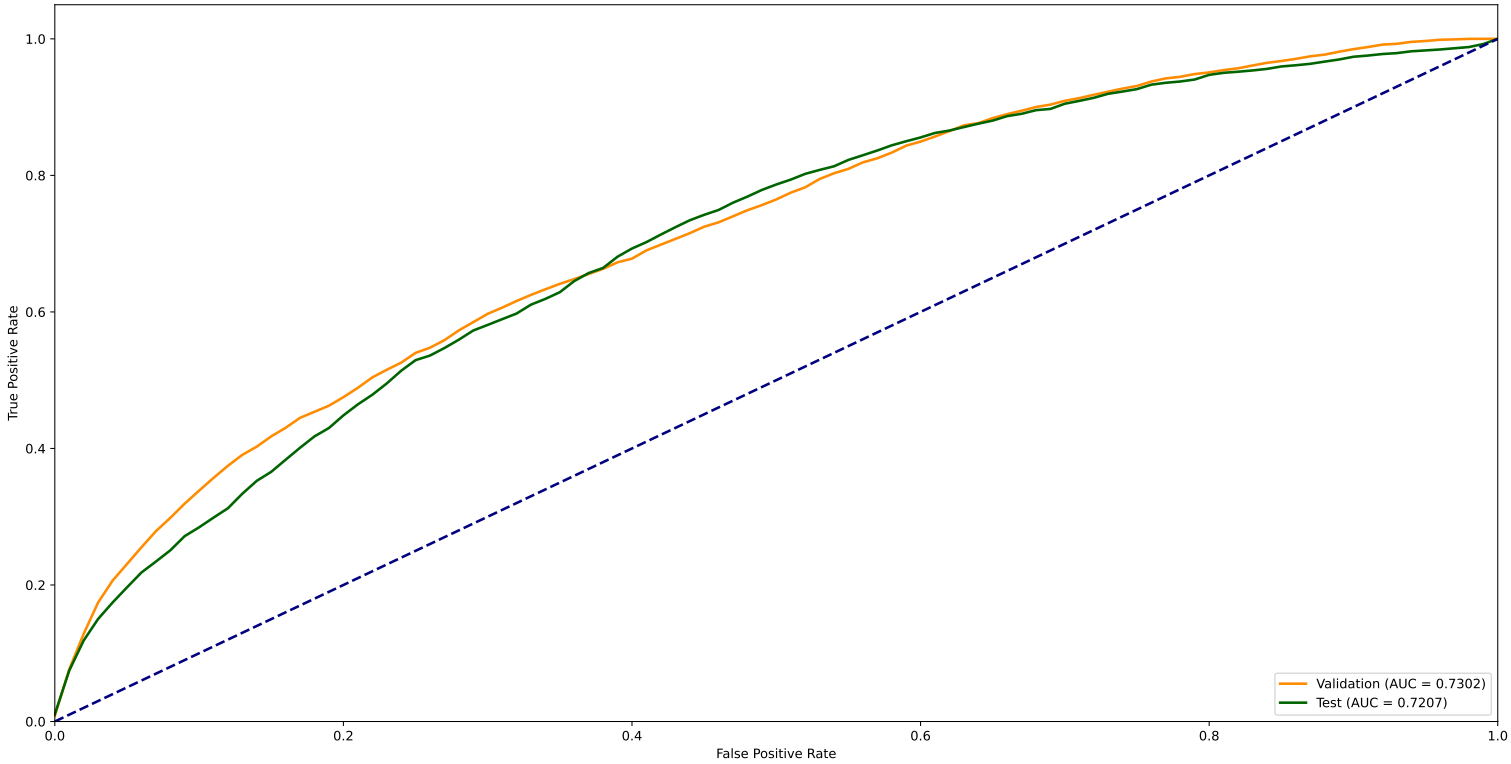


A stacked bar chart showing the weekly distance completed across different modalities

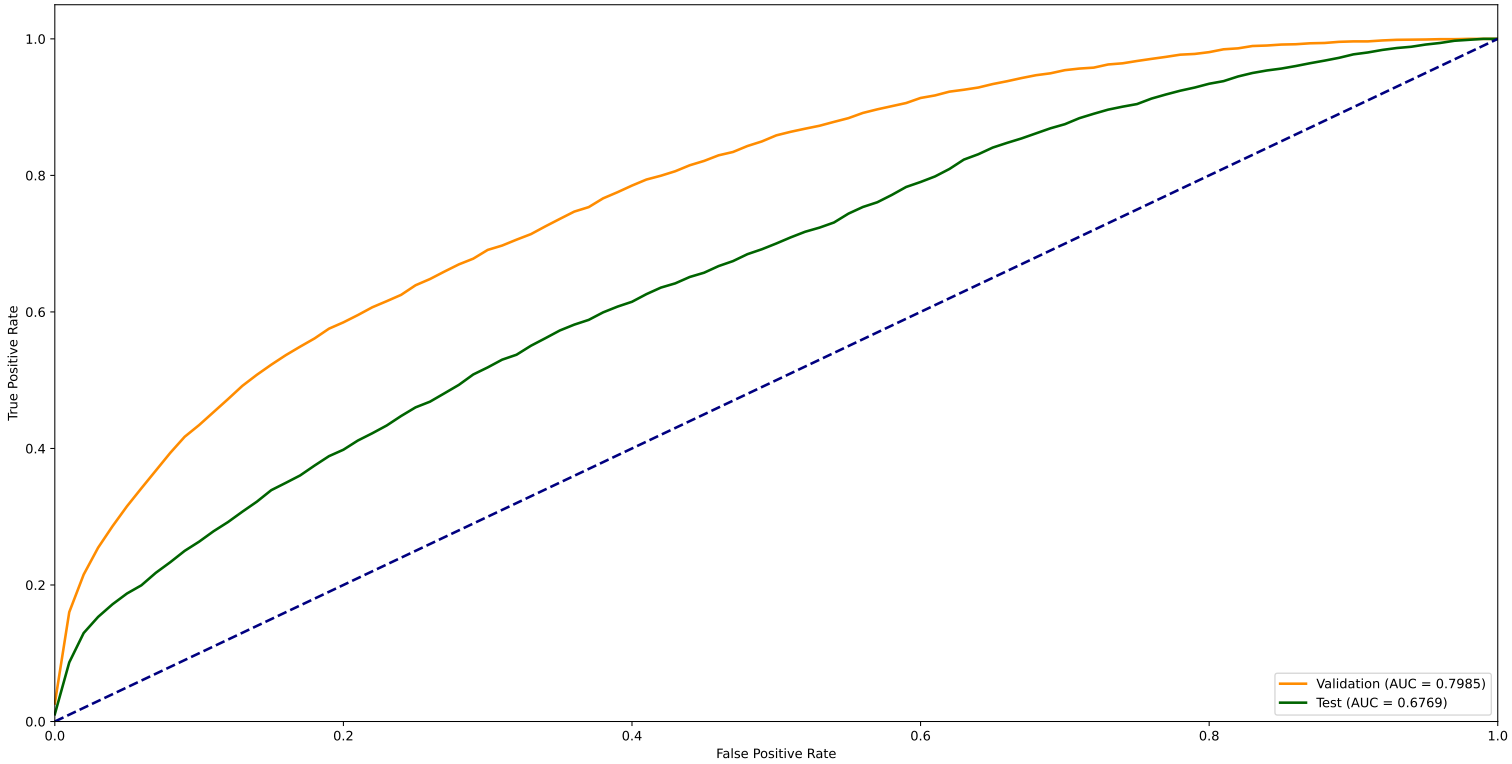


A stacked bar chart showing the weekly duration completed across different modalities

A.4 Machine Learning Results



The ROC curve for the day model



The ROC curve for the week model

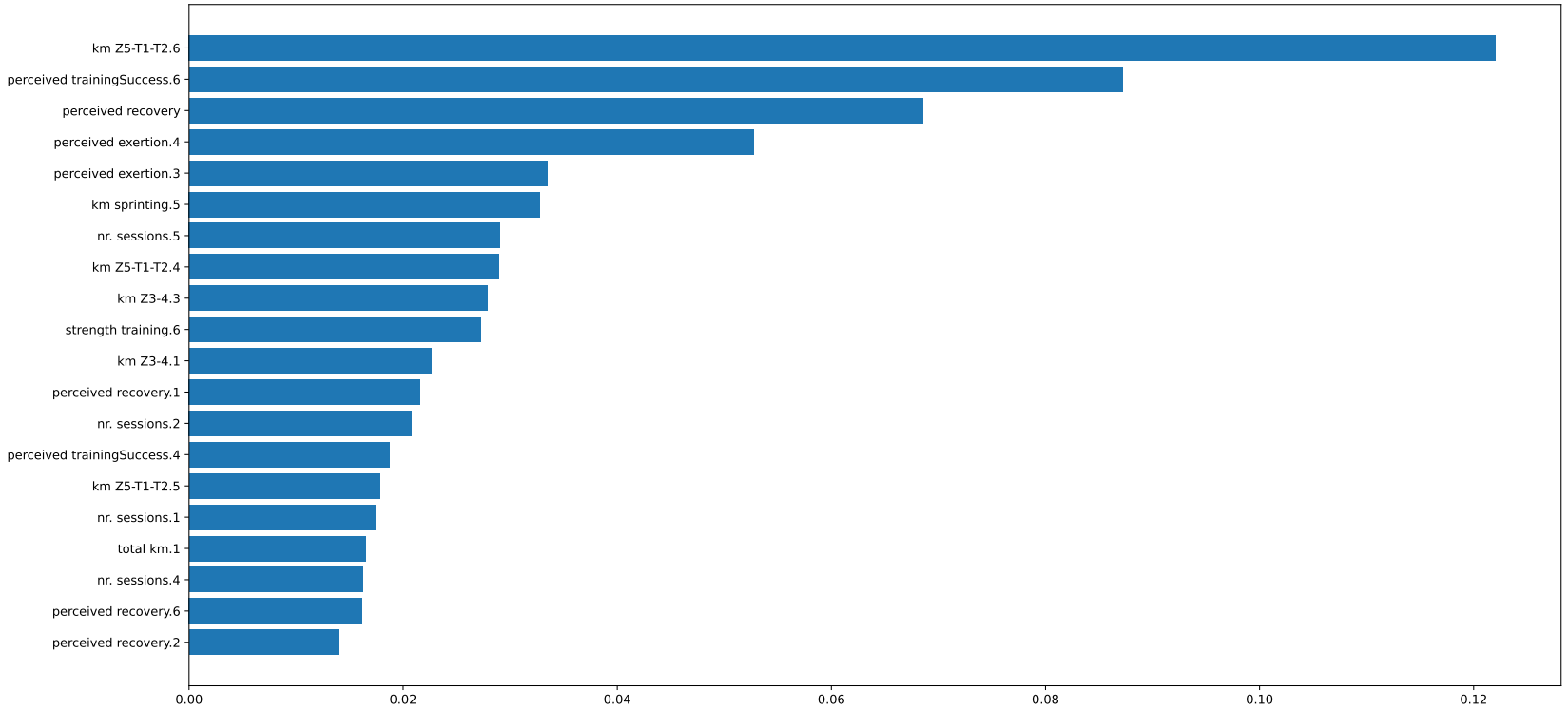


Figure A.1: The feature ranking for the day model

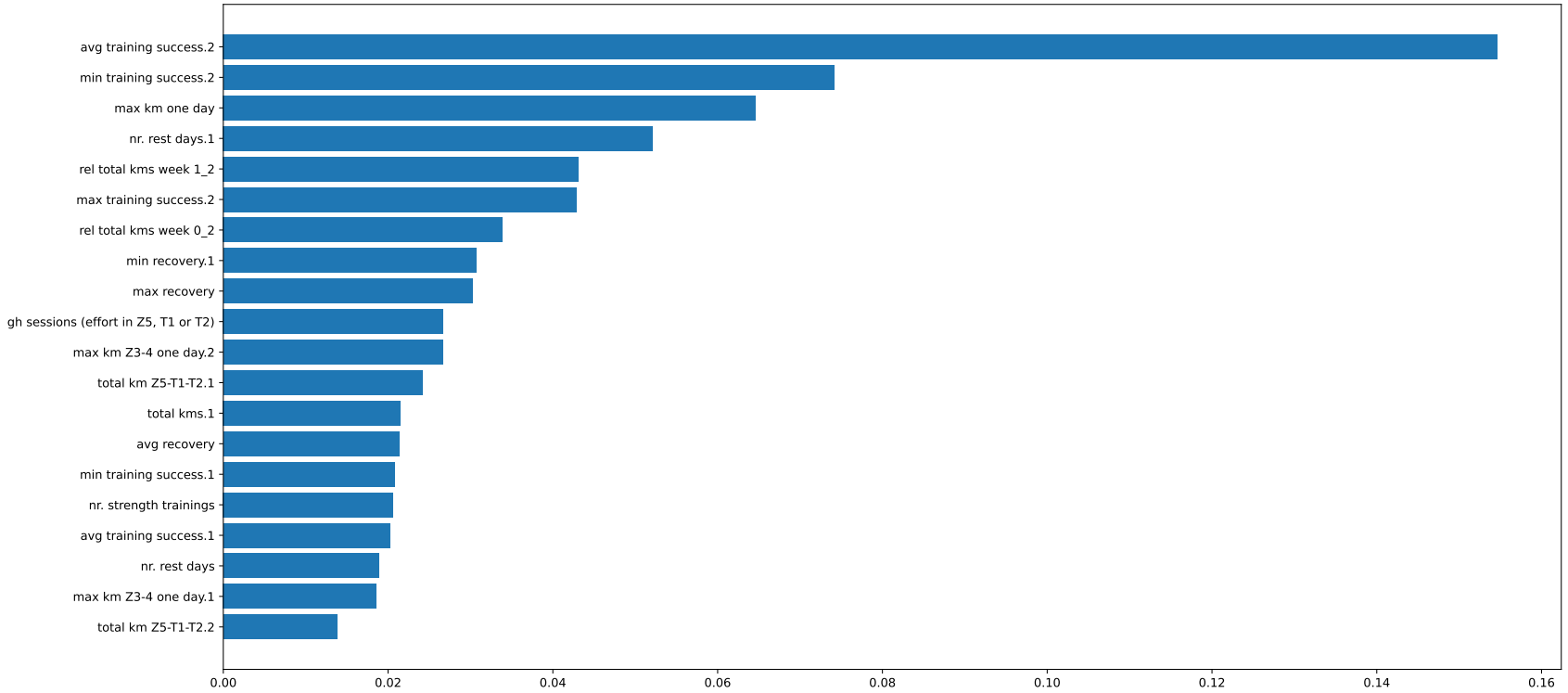


Figure A.2: The feature ranking for the week model

A.5 Ethics Approval Application

The Ethics Approval Application, which was submitted on November 2, 2023, and approval was granted on December 6, 2023, is attached on the following 9 pages.

School of Computer Science and Statistics
Research Ethics Application

Part A

Project Title: **Machine Learning to go *Nyoom*: Using Machine Learning to evaluate rowing training and predict training outcomes or performance**

Name of Lead Researcher (student in case of project work): **Liam Junkermann**

Name of Supervisor: **Dr. Lucy Hederman**

TCD Email: **junkerml@tcd.ie** Contact Tel. No.: **+353 089 484 2123**

Course Name and Code (if applicable): **Integrated Computer Science**

Estimated Start Date of survey/research: **ASAP**

I confirm that I will (where relevant):

- Familiarize myself with the Data Protection Act and the College Good Research Practice guidelines `http://www.tcd.ie/info_compliance/dp/legislation.php`
- Tell participants that any recordings, e.g. audio/video/photographs, will not be identifiable unless prior written permission has been given. I will obtain permission for specific reuse (in papers, talks, etc.)
- Provide participants with an information sheet (or web page for web-based experiments) that describes the main procedures (a copy of the information sheet must be included with this application)
- Obtain informed consent for participation (a copy of the informed consent form must be included with this application)
- Should the research be observational, ask participants for their consent to be observed
- Tell participants that their participation is voluntary
- Tell participants that they may withdraw at any time and for any reason without penalty
- Give participants the option of omitting questions they do not wish to answer if a questionnaire is used
- Tell participants that their data will be treated with full confidentiality and that, if published, it will not be identified as theirs
- On request, debrief participants at the end of their participation (i.e. give them a brief explanation of the study)
- Verify that participants are 18 years or older and competent to supply consent.
- If the study involves participants viewing video displays then I will verify that they understand that if they or anyone in their family has a history of epilepsy then the participant is proceeding at their own risk
- Declare any potential conflict of interest to participants.
- Inform participants that in the extremely unlikely event that illicit activity is reported to me during the study I will be obliged to report it to appropriate authorities.
- Act in accordance with the information provided (i.e. if I tell participants I will not do something, then I will not do it).

Signed: _____ Date: April 3, 2024

Liam Junkermann
Lead Researcher (student in case of project work)

Part B

Please answer the following questions.		***Yes/No***
Has this research application or any application of a similar nature connected to this research project been refused ethical approval by another review committee of the College (or at the institutions of any collaborators)?		No
Will your project involve photographing participants or electronic audio or video recordings?		No
Will your project deliberately involve misleading participants in any way?		No
Does this study contain commercially sensitive material?		No
Is there a risk of participants experiencing either physical or psychological distress or discomfort? If yes, give details on a separate sheet and state what you will tell them to do if they should experience any such problems (e.g. who they can contact for help).		No
Does your study involve any of the following?	Children (under 18 years of age)	No
	People with intellectual or communication difficulties	No
	Patients	No

School of Computer Science and Statistics
Research Ethical Application Form

Details of the Research Project Proposal must be submitted as a separate document to include the following information:

1. Title of project
2. Purpose of project including academic rationale
3. Brief description of methods and measurements to be used
4. Participants - recruitment methods, number, age, gender, exclusion/inclusion criteria, including statistical justification for numbers of participants
5. Debriefing arrangements
6. A clear concise statement of the ethical considerations raised by the project and how you intend to deal with them
7. Cite any relevant legislation relevant to the project with the method of compliance e.g. Data Protection Act etc.

Part C

I confirm that the materials I have submitted/provided are a complete and accurate account of the research I propose to conduct in this context, including my assessment of the ethical ramifications.

Signed: _____ Date: April 3, 2024

Liam Junkermann
Lead Researcher (student in case of project work)

There is an obligation on the lead researcher to bring to the attention of the SCSS Research Ethics Committee any issues with ethical implications not clearly covered above.

Part D

If external or other TCD Ethics Committee approval has been received, please complete below.

External/TCD ethical approval has been received and no further ethical approval is required from the School's Research Ethical Committee. I have attached a copy of the external ethical approval for the School's Research Unit.

Signed: _____ Date: _____

Lead Researcher (student in case of project work)

Part E

If the research is proposed by an undergraduate or postgraduate student, please have the below section completed.

I confirm, as an academic supervisor of this proposed research that the documents at hand are complete (i.e. each item on the submission checklist is accounted for) and are in a form that is suitable for review by the SCSS Research Ethics Committee

Signed: _____ Date: _____

Supervisor

CHECKLIST

Please ensure that you have submitted the following documents with your application:

1.	SCSS Ethical **Application Form**	Yes
2.	**Participant's Information Sheet** must include the following: ▪ Declarations from Part A of the application form; ▪ Details provided to participants about how they were selected to participate; ▪ Declaration of all conflicts of interest.	Yes
3.	**Participant's Consent Form** must include the following: ▪ Declarations from Part A of the application form; ▪ Researchers' contact details provided for counter-signature (your participant will keep one copy of the signed consent form and return a copy to you).	Yes
4.	**Research Project Proposal** must include the following: ▪ You must inform the Ethics Committee **who** your intended participants are i.e. are they your work colleagues, class mates etc. ▪ How will you recruit the participants i.e. **how** do you intend asking people to take part in your research? For example, will you stand on Pearse Street asking passers-by? ▪ If your participants are under the age of 18, you must seek both parental/guardian AND child consent.	Yes
5.	Intended **questionnaire**/survey/interview, protocol/screenshots/representative materials (as appropriate)	No
6.	**URL** to intended on-line survey (as appropriate)	No

Notes on Conflict of Interest

1. If your intended participants are work colleagues, you must declare a potential conflict of interest: you are taking advantage of your existing relationships in order to make progress in your research. It is best to acknowledge this in your invitation to participants.
2. If your research is also intended to direct commercial or other exploitation, this must be declared. For example, *"Please be advised that this research is being conducted by an employee of the company that supplies the product or service which form an object of study within the research."*

Notes for questionnaires and interviews

1. If your questionnaire is paper based, you must have the following opt-out clause on the top of each page of the questionnaire: *"Each question is optional. Feel free to omit a response to any question; however the researcher would be grateful if all questions are responded to."*
2. If you questionnaire is **on-line**, the first page of your questionnaire must repeat the content of the information sheet. This must be followed by the consent form. If the participant does not agree to the consent, they must automatically be exited from the questionnaire.
3. Each question must be **optional**.
4. The participant must have the option to '**not submit, exit without submitting**' at the final submission point on your questionnaire.
5. If you have open-ended questions on your questionnaire you must warn the participant against naming **third parties**: *"Please do not name third parties in any open text field of the questionnaire. Any such replies will be anonymised."*
6. You must inform your participants regarding illicit activity: *"In the extremely unlikely event that illicit activity is reported I will be obliged to report it to appropriate authorities."*

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Information Sheet for Participants

Title of Research Project:

Machine Learning to go *Nyoom*:
Using Machine Learning to evaluate rowing training and predict training outcomes or performances

Name of Lead Researcher: Liam Junkermann

Email: junkerml@tcd.ie

Phone Number: +353 089 484 2123

Name of Supervisor: Dr. Lucy Hederman

You are invited to participate in a research project as part of a final year project for the partial fulfillment of an Integrated Computer Science undergraduate Degree at Trinity College Dublin. This project aims to determine if, and how, machine learning can be used to quantify rowing training, and its outcomes and performances. This can be used to improve training effectiveness for individual athletes allowing them to adjust their training according to how their bodies respond to different training loads.

Please ensure that you read all the information below before deciding to participate in this project. Please feel free to ask any questions that you may have, as it is important that you clearly understand what participating in this project involves.

You are under no obligation to participate in this project. It is voluntary, and your decision to take part or otherwise will not result in any penalty against you.

The sharing of data is optional. You can withdraw from the project at any time before the project submission in April 2024, even after it has started. After the submission, it will no longer be possible to exclude your data.

All information gathered as part of this project will be pseudonymised when presented as results. Participant names will be pseudonymised, with code names being used instead of the participants' own names. The researcher will securely maintain a translation key to facilitate exclusion of participant's data from the project if requested or to provide feedback which may be the result of a model that has been trained by the researcher. All information about this project will be removed from the researcher's personal devices and cloud storage once the results of the report have been ratified by the exam board. At this point, the researcher will delete all data gathered for this project unless otherwise requested by participants.

What is the purpose of this project?

Quantifying the effect of training on the body is something that trained athletes try to do daily to make their training as efficient as possible to hit their fitness and performance goals. As technology has evolved, different systems and approaches to quantify metrics like load and recovery have been developed. Original models and systems used a linear approach to model these biological processes and adaptations. The formative model for modelling human performance, developed by Eric Bannister, has been supported and built upon many times since its publication in 1975. This model is, unfortunately, incomplete due to its linear nature. Machine Learning has been introduced to the field of study in an attempt to increase both the accuracy of performance modelling and the number of variables that can be used in the model. This project aims to develop a model that can be used to quantify training load and recovery and predict performance based on the available training and recovery data. The goal is to make participants' individual training more effective to drive stronger performances. By participating in this project, you are contributing to the completion of the researcher's undergraduate degree.

Who is organising the project?

Liam Junkermann is the principle researcher and is receiving academic supervision from Dr. Lucy Hederman of the School of Computer Science and Statistics in Trinity College Dublin.

Why am I being asked to partake in this project

You are being asked to partake in this project as you are a rower with intentions to compete, at a minimum, at a national level this season. As a result, you will be committing to 8-12+ training sessions per week where consistent data may be collected, paired with semi-regular benchmark tests (eg. 6k r20, 30min r20, 6k Open, 2k Open), which can provide the researcher with different options to measure model success and provide feedback to you, the athlete.

What will my role in the project entail?

To participate in this project, you will be asked to provide data through various pipelines built by the researcher. The researcher will be collecting any and all training and recovery data you might have available from the time of joining until the time of submission in April 2024. Most of the data collection will happen through automatic API collection pipelines, but may also include manual entry through a training log. Data will be collected with initial signup where you will log in to any accounts, such as Polar, Concept2 Logbook, Strava, or Whoop – anywhere you might log your training. The goal is to collect as much relevant data as possible. Once the signup is completed, your data will be collected without any further engagement from you. You may also asked to engage in monitoring to collect relevant information about changes in oxygen utilisation zones (as a result of a lactate test), information about illness or injury, and any other information which could have an impact on the effectiveness of training. Providing any or all data is voluntary and you may at any time, before the submission date in April 2024, ask to have the data you have shared be deleted and leave the project data set.

What are the benefits of my taking part in this project?

By participating in this project, you are providing invaluable data to the researcher, which ultimately could benefit you in providing feedback on your training, and understanding which trends in your training produce stronger results.

What are the risks in my taking part in this project?

There are no expected risks associated with taking part in this project. Agreeing or declining to participate in this project will not impact you in any way. Collected data will remain pseudonymised and will be disposed of once the report has been ratified by the exam board unless otherwise requested by participants.

Will it cost me to take part in this project?

There are no monetary costs associated with participation in this project. There may be a small time commitment to set up data collection, and periodic check-ins for training-related data, such as heart rate zones and the result of HP testing such as lactate or VO2 max testing. Initial setup for data collection should take no more than 10-15 minutes. Periodic check-ins may take up to 5 minutes per week.

Is this project confidential?

All information collected for the purposes of this project will be treated with the strictest of confidence. Any data collected will be stored securely, in an unidentifiable form, and deleted from the researcher's personal devices and cloud storage unless otherwise requested by participants when the project is complete and has been ratified by the exam board.

Participants' names will be masked with code names being used instead of participants' own names before being presented as results. The researcher will make every attempt to obfuscate any training plans which may become apparent in raw data. The researcher will securely maintain a translation key to facilitate the exclusion of a participant's data if requested or to provide feedback which may be the result of a model which has been trained by the researcher. No information will be shared with anyone other than the principal researcher in any manner that may be easily identifiable. All data will be stored securely for the duration of the project and will be deleted from the researcher's personal devices and cloud storage, unless otherwise requested by the participant, once the study has concluded in April 2024 and the project has been ratified by the exam board.

The researcher is aware that some squads or athletes may have policies regarding the sharing of training data, all data provided to the researcher for the purposes of this project will only be accessed by the researcher and only de-pseudonymised on the request of the participant to provide training feedback or to delete a participant's data.

Are there any conflicts of interest in this project?

This project forms part of the principle researcher's undergraduate degree in Integrated Computer Science at Trinity College Dublin. By participating in this project, you are contributing to the completion of this degree.

Where can I get further information regarding this project?

If you have any queries or concerns regarding this project now, or at any point in the future, please contact the researcher via email at junkerml@tcd.ie.

You can reach the Data Protection Officer using the contact details below:
Email: dataprotection@tcd.ie
Post: Data Protection Officer, Secretary's Office, Trinity College Dublin, Dublin 2, Ireland

Thank you for taking the time to read this information sheet.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Project Informed Consent Form

Lead Researcher: Liam Junkermann

Background of Research: The development of a machine learning model to evaluate rowing training to predict training outcomes and performances. This research is being carried out as part of a final-year project for the fulfillment of an Integrated Computer Science degree at Trinity College Dublin.

Procedures of this Project: During this project, you will be asked to share your wearable and training data through an online portal. Additionally, you may be asked to complete some monitoring tasks.

Publication: There will be a final report submitted to the School of Computer Science and Statistics at Trinity College Dublin. Individual data may be included in the final report pseudonymised.

Declaration:

- I am 18 years or older and am competent to provide consent.
- I have read or had read to me, a document providing information about this research and this consent form.
- I have had the opportunity to ask questions and all my questions have been answered to my satisfaction I understand the description of the research that is being provided to me.
- I agree that my data is to be used for scientific purposes and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- I understand that if I make illicit activities known, these will be reported to the appropriate authorities.
- I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
- I have received a copy of this agreement.

Participant's Name In Bold: _____

Participant's Signature: _____

Statement of Researcher's Responsibility: I have explained the nature and purpose of this research project, the procedures to be undertaken, and any risks that may be involved. I have offered to answer any questions and thoroughly answered any such questions. I believe that the participant understands my explanation and has freely given their informed consent.

Signed: _____ Date: _____
Liam Junkermann

Researcher's Contact Information:

Email: junkerml@tcd.ie
Telephone: +353 089 484 2123

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Outline of Research Proposal

Title of Project
Machine Learning to go *Nyoom*: Using Machine Learning to evaluate rowing training and predict training outcomes or performance

Dates and Duration
This project will run from ASAP until April 2024.

Purpose and Academic Rationale

The purpose of this project is to develop a machine learning model that can quantify rowing training load more effectively to predict the resulting outcomes and performances from a certain training program. The formative Human Performance Model from 1975 relied on a linear model, only using Fitness and Fatigue as inputs. The linear model for human performance is inherently flawed as adaptations to training load are far more complex. Since then, technology has evolved to allow researchers to build more complex and effective models of human performance. The development in machine learning allows models to be trained on sports-specific data to extrapolate what training plans work and provide individualised feedback using the data available.

At present, no tool like this exists for Rowing. Given the cardiovascular intensity of training, many strain models built for sports generally overestimate the strain, or fatigue, of a given training session. Additionally, there is some disagreement about which approach to training is best. The most popular approaches for endurance training are pyramidal or periodization. These approaches guide what percentage of time spent training should be spent in various heart rate zones.

Rowing training is uniquely placed to be used to train a machine learning model. With countless hours spent on the rowing machine (erg) each week, most often with a heart rate monitor (HRM) connected, most pre-elite and elite athletes have granular, stroke to stroke, data about their training. Given the prevalence of using heart rate to train many athletes also wear heart rate monitors on the water, with some also using Stroke Coaches (GPS computers which various metrics on the water such as speed, stroke rate, in some cases power, and the capability to connect to HRMs) to record distance, time, and speed. Given the time and energy commitment required for rowing many athletes also use fitness-tracking wearables (eg. Polar Watches, Whoop Straps, etc.) to collect data about recovery and sleep. Considering this volume of data, a machine learning approach to analyse and provide feedback, and predictions, on a given training block or to help work to a given goal.

Procedures of the Project

There will be two steps to the project

1. **Data Collection**
   Participants will be asked to do a once-off signup process which will enable automatic collection of their wearable and training data through the use of APIs. Participants may be asked to do weekly monitoring each week to provide supplemental data. The goal of the data collection step is to be as unobtrusive to the athletes as possible. Their data will be collected and analyzed on the researcher's personal computer, before being transferred to cloud storage, in a pseudonymised format, to allow for processing in the machine learning step.
2. **Data Processing and Model Generation**
   Once a sufficient amount of data has been collected, the researcher will engage in processing the data and developing a model. Data will continue to be collected to continue adding to the data available to the model. Any findings from the model will be released to the participants. The researcher will maintain a translation key securely on their local machine to provide this feedback, and if a participant asks for their data to be deleted.

Participants

The participants for this project will be recruited through word of mouth, text messages, and email. The researcher is an active member of a Dublin senior rowing squad with friends and former teammates in other squads in Ireland and the United States of America who can be recruited for the project. The researcher will be reaching out to people whom he already has contact details for, and will be emailing captains and coaches of squads for additional participants. These email addresses will be obtained from publicly posted locations such as club Instagram accounts and websites.

The criteria that a participant must meet in order to take part in this project:

- Be at least 18 years old.
- Be training for Rowing at least 8 times per week (>45 minutes per session)

- Have competed at a minimum, national level in the last 6 months or intend to compete at this level in the current season.

Health conditions and injuries which may have occurred before or during the course of the project will not exclude a participant but will be noted as part of the initial intake or the weekly monitoring.

Debriefing Arrangements

All participants in this project will be debriefed through a written document explaining the process to them. If requested, participants will receive the results of the project once the project is completed.

Ethical Considerations

Some ethical considerations arise from the project with regard to participation, data collection, and protection.

Participation Participation in this project is entirely voluntary. If a participant no longer feels comfortable sharing their data, they may withdraw at any time, even after the project has commenced, with all of their data being removed from the researcher's personal device and cloud storage.

Data Collection and Protection All data collected is done through authorised APIs, secured with the participants' login and an API key provided to the researcher. The data is then processed to pseudonymised, by removing any names or user IDs and replacing them with randomly generated keys for each participant. The data is then stored in an encrypted cloud storage provider. A translation key list will be stored securely and encrypted on the researcher's personal device, which is password-secured and has an encrypted disk. The researcher is aware that some squads may have policies regarding the sharing of training data, all data provided to the researcher for the purposes of this project will only be accessed by the researcher and only de-pseudonymised at the request of the participant for the purposes of providing training feedback or to delete a participant's data.

Given that some elements of the data collected may include GPS data. This data cannot be effectively used to discover who a given person is as this will only be for on-the-water rowing sessions and as such would be one of many potential participants rowing on that body of water. The location data available does not put any of the participants at any risk of personal information being revealed.

Legislation

All data gathered as part of this project will be held and maintained in accordance with the General Data Protection Regulation (GDPR). All participants will be anonymised before being included in the results. Information gathered throughout the project will be stored securely, which only the researcher will have access to.