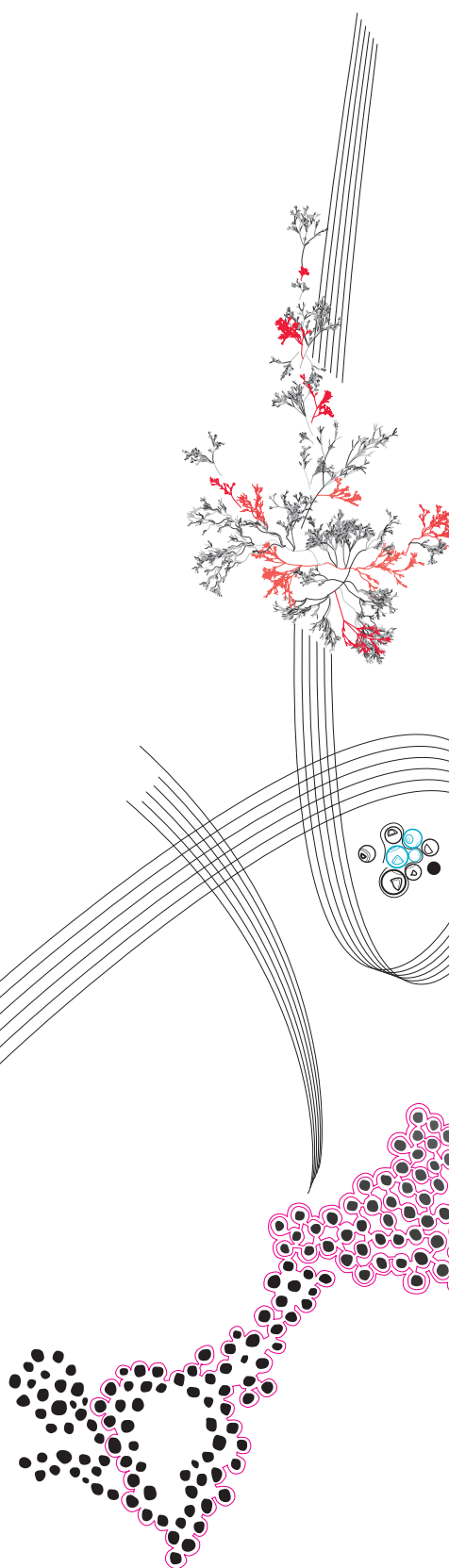


MSc Computer Science



Monitoring training load
and identifying fatigue
in young elite speed skaters
using machine learning methods

Marjolein Bolten

Supervisors:
Dr. E. Talavera Martinez
Dr. J. Reenalda
Dr. D. Reidsma
T. Waanders

March, 2024

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

Preface

With this thesis, I want to conclude my study time at the University of Twente. Which started in 2017 with a bachelor of Applied Mathematics, followed by two board years, a minor in Human Movement Sciences in Amsterdam, and now the Master of Sports Data Science. In this Master, I could combine my interests in Data Science together with my interests in Sports and Movement, and I am grateful that this unique combination was offered at the UT.

During the past year, I did my master's thesis on fatigue towards young elite speed skaters of TalentNED. I really liked working in this environment and with the coaches, the rest of the staff, and of course the twelve speed skaters. Therefore I want to thank everyone at TalentNED for welcoming me the past year.

Next to the people at TalentNED I also want to thank my two daily supervisors, Estefania Talavera Martinez and Jasper Reenalda. Working with two daily supervisors was not always easy in planning meetings, but your insights on both the data science and sports science direction, respectively, were really useful. I also want to thank Dennis Reidsma for taking part in my examination committee and for the last-minute critical insight on my thesis, which improved the quality of my work.

Last but not least, I want to thank all my friends that I met during my years as a student in Enschede. They supported me during the past year when I was working on my thesis. Listened to me when I was complaining about the lack of data collected and helped me by proofreading my thesis.

Abstract

Monitoring training load is crucial for enhancing sports performance, as excessive load can lead to fatigue accumulation and decreased performance. Research has extensively investigated sports monitoring techniques, including measuring external and internal loads. Understanding fatigue states helps coaches prevent nonfunctional overreaching and optimize training for improved performance. Continuous monitoring, such as using non-invasive maximal effort tests, is essential to detect declines in performance and adjust training accordingly.

In this study, a dataset is collected of young elite speed skaters consisting of data from morning questionnaires, daily jump tests, Wingates tests, and information about training sessions. Despite the amount of research already done in the direction of fatigue and sports performance some connections are still not described in full detail. To get a better understanding of these connections, three different methods are examined on this dataset in relation to fatigue. As a first method, the importance of input variables is determined using a classification decision tree method. Using statistical tests the data of Wingate tests is analyzed and lastly, different Long Short-Term Memory (LSTM) models are tested for their ability to predict resting heart rate (HR) data.

Monitoring daily jump height and using a wellness questionnaire did not effectively identify fatigue in young elite speed skaters. Similarly, the Wingate test, conducted three times in five weeks, failed to serve as a reliable measure of fatigue due to inconclusive results influenced by other factors. However, a univariate LSTM model showed promise in predicting daily resting HR data, with an average Root-Mean-Square Error (RMSE) of 1.5 beats per minute. Before this model can be used in a practical situation, further research is needed to improve the performance of the LSTM model. As a practical application, this model can allow for the detection of abnormal HR patterns indicative of fatigue. Consequently, a combination of monitoring internal and external loads, along with predictive resting HR data using LSTM models, offers a possible viable approach to identifying fatigue in young elite speed skaters.

Contents

1	Introduction	6
1.1	Research Gap	7
1.2	Research Objectives	7
1.3	Structure of this report	8
2	Scientific Background	9
2.1	Training Load Management	9
2.1.1	Internal load	9
2.1.2	External load	11
2.2	Applications of machine learning in health and sports research	14
2.2.1	Health care	14
2.2.2	Heart rate prediction	14
2.2.3	Sports performance	15
2.3	Conclusion - literature review	15
3	Dataset collection and curation	16
3.1	Subjects	16
3.2	Available data within TalentNED	16
3.3	Data collection for monitoring fatigue	18
3.4	Collected dataset	19
3.5	Data preprocessing	20
4	Data analysis	22
4.1	Importance of variables	22
4.1.1	k-fold	22
4.1.2	Leave-one-user-out	22
4.1.3	Train on variables with the same distribution	22
4.2	Wingate performance to identify fatigue	23
4.2.1	Standardized t-test	23
4.2.2	Kruskal-Wallis test	24
4.2.3	Wilcoxon Signed Rank	24
4.3	Resting HR prediction	24
4.3.1	Training parameters	25
4.3.2	Univariate Time Series	25
4.3.3	Multivariate Time Series	26
4.3.4	Multivariate Time Series with top 10 relevant features	26
5	Results	27
5.1	‘Fatiguing dataset’	27
5.1.1	Descriptive analysis	27
5.1.2	Data visualisation	27
5.2	Importance of variables	28
5.2.1	k-fold	29
5.2.2	Leave-one-user-out	30
5.2.3	Train on variables with the same distribution	30
5.3	Wingate performance to identify fatigue	31
5.3.1	Visualization	32
5.3.2	Statistical tests	33
5.4	Resting HR prediction	34
5.4.1	Univariate Time Series	34
5.4.2	Multivariate Time Series	35
5.4.3	Multivariate Time Series with top 10 relevant features	35
5.4.4	Comparison univariate vs multivariate LSTM	36
5.4.5	Exponential smoothing	36

5.4.6	Excluding yesterday's resting HR as input	37
6	Discussion	38
6.1	Importance of variables	38
6.2	Wingate performance to identify fatigue	39
6.3	Resting HR prediction	41
7	Conclusion	43
8	Appendix A	48
8.1	Daily questionnaire	48
8.2	Training log	49
8.3	Evening questionnaire	49
8.4	Weekly questionnaire	50
9	Appendix B	51
9.1	Univariate LSTM network	51
9.2	Multivariate LSTM network 1	56
9.3	Multivariate LSTM network 2	61

1 Introduction

Monitoring training load (TL) is incredibly important to increase performance in sports activities (Kreher & Schwartz, 2012). Therefore, a lot of research has been done on monitoring TL over the past years (Halson, 2014). A well-established training scheme can increase sports performance, provided that the training load is not too high. A too high training load, which could for example be caused by too little rest between training sessions can lead to an accumulation of fatigue. An increased state of fatigue can lead to worse sports performance or nonfunctional overreaching (Kreher & Schwartz, 2012).

Training load can be measured as external or internal load. A difference between external and internal load will reveal the level of fatigue of an athlete (Halson, 2014). The external load can be measured as the physical work done by the body in terms of movement, such as power output or pace. On the other hand, the internal load can be measured by internal characteristics, such as heart rate (HR) or blood lactate concentration (McArdle et al., 2015). When keeping the power output and duration constant, a lower HR in the longer term or a higher HR in the short term, can indicate a state of fatigue. This could, for example, be observed during a cycling tour of an athlete.

Overreaching can be divided into three categories: functional overreaching, nonfunctional overreaching, and the overtraining syndrome. While functional overreaching will lead to positive training adaptations after the temporary performance decrement, the other two will keep having negative effects on performance. The difference between functional and nonfunctional overreaching is really small (Kreher & Schwartz, 2012). To stay on the good side of the line and benefit from the long-term positive effects of functional overreaching, coaches must know the status of fatigue of their athletes. An athlete is in functional overreaching if recovery takes days to weeks, while if recovery takes longer, from several weeks to possibly years, one will speak of non-functional overreaching or the overtraining syndrome. Prolonged fatigue is thus a sign of non-functional fatigue and by monitoring fatigue coaches get more knowledge about the type of overreaching of their athletes (Meeusen et al., 2006).

Besides steering towards functional overreaching, monitoring fatigue can also be used as a subjective measure to gain knowledge about how athletes perceive the training load. Perceived training load can be used as an indicator of readiness for competitions. It is necessary to know whether an athlete is prepared for competition to reduce chances of nonfunctional overreaching, injury, and illness (Thorpe et al., 2017). In this paper, ‘recovered’ is defined as the ability to meet a certain performance in a sport-specific activity. For example, completing the next training within the planned intensity zones. The knowledge about the state of fatigue of athletes can also be used to optimize the training load for athletes by individually modifying the training scheme. These modifications can optimize training load and so increase sports performance.

An important indication of nonfunctional overreaching is a decrease in sports performance. To indicate a decrease, sports performance should be monitored on a daily basis. Sports performance should be monitored utilizing a maximal effort test that doesn’t influence the training sessions of the athletes. Following these criteria, the test should be performed daily and must be non-invasive. Jump tests on a weekly basis have been used in previous work as a non-invasive and maximal effort test (Gavanda et al., 2023). Jump tests to monitor fatigue are also performed on an incidental basis, next to other performance measures (Pupo et al., 2021).

1.1 Research Gap

As explained above, in sports, keeping an eye on how fatigued athletes are is really important. Currently, there does not exist one simple way to measure fatigue. If fatigue is monitored, a combination of variables and tools is used which is time consuming for both athletes as coaches and researchers. Ideally, a simple definitive tool to measure fatigue that is accurate and reliable is needed to help coaches and athlete to identify unexplained fatigue (Halsen, 2014). Then, coaches/athletes would be able to know the state of fatigue of an athlete at each moment in time. To find such a tool and take into account the work already done in the field of monitoring fatigue, a logical next step would be to monitor fatigue continuously over a longer time period to contribute to better training load management (TLM).

From literature, it is known that certain variables have a relation with fatigue, for example, jump height and resting heart rate. A lower jump height or a higher resting HR are indicators of a higher state of fatigue (Budgett, 1998; Halsen, 2014). However, other factors could be related to a lower jump height or higher resting heart rate. Therefore the relationship between jump height, resting heart rate, and fatigue is not yet investigated in sufficient detail.

In this study, the goal is to further investigate this connection between fatigue and jump height and/or resting heart. The next goal is to identify fatigue concerning different aspects, such as self reporting questionnaires, Wingate performance, and the training schedule. These two steps will help to reach a bigger goal: a simple-to-use test or tool, which can accurately and reliably detect some of the indicators of fatigue.

1.2 Research Objectives

In this work, the focus will be on the connection between fatigue and some indicators to identify fatigue in young elite speed skaters in the Netherlands. These connections will be further investigated using machine learning models. The research in this work takes place at TalentNED, an organization for talented young speed skaters and mountain bikers. More specifically, this research is focused on the speed skate team of TalentNED. This leads to the following research question:

How can we monitor training load and identify fatigue in young elite speed skaters?

To be able to answer this question, first, a literature review is conducted on overtraining and monitoring training load in different kinds of sports. A second literature review is executed on different existing machine learning models within the health and sports domain. After this, a data collection protocol, to obtain different indicators of fatigue, is designed and executed within the group speed skaters of TalentNED.

This yields the following sub-research questions:

1. *What are relevant variables to describe the training load of athletes?*
2. *How do the existing machine learning methods perform on the prediction of heart rates?*
3. *Can fatigue be identified by monitoring daily jump height and by a wellness questionnaire?*
4. *Can we use Wingate performance to measure fatigue in young elite speed skaters?*
5. *How can we predict resting HR data of young elite speed skaters?*

This work contributes to the field of Sports Data Science by:

- presenting a univariate LSTM model that predicts resting HR.
- showing that Wingate performance data can not be used for identifying fatigue in young elite speed skaters.
- concluding that the variant and format of an unsupervised daily reach and height jump test is not a viable method to identify fatigue in young elite speed skaters.

1.3 Structure of this report

This work is structured as follows. In Section 2 a theoretical background is given with used definitions and related work in TLM. Related works that already used data science for TLM are discussed in Section 2.2. In these two sections, the first two research questions will also be answered. Details about the subjects, the dataset, and data pre-processing steps used in this study are described in Section 3. In Section 4, the methodology used to answer the last three sub-research questions will be explained per sub-research question. The results of those methods are stated in Section 5 in the same structure as the methodology in Section 4. The discussion of these results and recommendations for further research will be given, per sub-research question in Section 6. Finally, in Section 7, conclusions will be drawn and answers to the (sub-) research question(s) will be given.

2 Scientific Background

In this section, first TLM related parameters will be discussed and next some already existing machine learning models in health and sports research will be discussed.

2.1 Training Load Management

In order to identify fatigue, it is important to know what fatigue is, and how fatigue follows from training sessions. Fatigue could be divided into two categories, performance fatigability and perceived fatigability. Performance fatigability describes a decrease in a performance measure, that could be measured objectively. On the other hand, perceived fatigability refers to the subjective feeling of the performer (Behrens et al., 2023). This work, will be mainly focussed on performance fatigability, so from now on, if referred to fatigue it means performance fatigability. Following these definitions, an athlete is called fatigued once he has not recovered from previous exercises (Bishop et al., 2008). Therefore, in this work, ‘fatigued’ is defined as the inability to meet a certain performance in a sport-specific activity. On the other hand, perceived fatigue is defined as an overwhelming sense of tiredness, lack of energy, or feeling of exhaustion (Krupp & Pollina, 1996).

One gets fatigued by training and the training load is monitored to measure someone’s inability to meet a certain performance. The state of fatigue can be seen as the difference between internal load and external load variables. Therefore, one can individually monitor the internal or external load to monitor fatigue (McArdle et al., 2015). In Figure 1, a schematic overview is given of the relevant variables to monitor and measure internal and external load. In the following section, these variables will be described in further detail. Relevant research related to these variables can be found in Table 1 and is also described after introducing the variables.

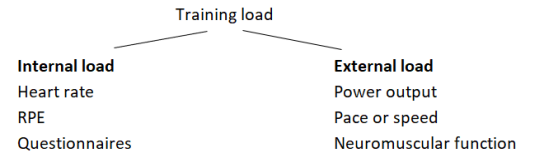


FIGURE 1: Variables describing training load.

2.1.1 Internal load

The internal load can be measured with internal characteristics, such as the heart rate. *Heart rate* (HR) is a variable that can be used as a fatigue measure. One’s oxygen uptake will increase linearly with an increase in heart rate (McArdle et al., 2015). As oxygen uptake is a good indication of exercise intensity, the heart rate can be seen as a viable method to assess internal load (McArdle et al., 2015).

Heart rate recovery (HRR)

Subsequently, looking at heart rate as a plain variable, one can measure the recovery of the heart rate as well. The heart rate will increase during exercise and will return slowly to a resting baseline after exercise. Consequently, the more fatigued the athlete is, the longer it takes before the heart rate is back at the resting baseline (Lambert & Borresen, 2006). Heart rate recovery (HRR) can be calculated by taking the difference between the heart rate at the end of exercise and after 60 seconds of recovery. The higher this value, the more fit the athlete is (Daanen et al., 2012).

Resting heart rate

Instead of looking at HRR, one can also look at the daily resting HR of athletes. An elevated resting HR can be a sign of fatigue or overtraining (Budgett, 1998).

Heart rate variability (HRV)

Another interesting variable related to the heart rate is heart rate variability (HRV), which is the time difference between different heartbeats. With faster heart rates, there is less time between successive heartbeats, and therefore there is a smaller opportunity to have higher variations in these time intervals. Thus, higher heart rates will lead to a decrease in HRV and this is related to stress fatigue (Jiménez Morgan & Molina Mora, 2017). As heart rate varies from person to person, heart rate recovery and variability also have different baseline values for different athletes.

Various works have used heart rate variables to monitor training load and fatigue in athletes. For example, Gavanda et al. (2023) did research on cheerleaders prior to the World Championship. For two weeks, they measured heart rates daily, obtained questionnaire data about training and recovery, and analyzed data from a jump test. During these two weeks of fatiguing pre-competition training, the results of an Analysis of variance (ANOVA) test showed that the heart rate of the cheerleaders increased over time.

Heart rate variables in this work

HRR requires a specific end time of the training, which probably will not be specific enough if athletes should determine this on their own. Small measurement errors will influence the heart rate recovery too much and therefore HRR will not be used in this work.

Not all the smartwatches that are used have HRV as a variable within the smartwatch, so HRV will not be used in this work either. Resting HR and HR during activity will be used in this work as internal load variables, as these are easy-to-collect and reliable variables.

Rate of Perceived Exertion (RPE)

Another popular variable to assess internal training load is the Rate of Perceived Exertion (Inoue et al., 2022). With an RPE, the athletes can indicate how they perceived the training session. One common scale to use RPE is the Borg RPE scale, which goes from 6 to 20. This scale is based on the assumed linear relationship between RPE and heart rate, where an RPE of 6 corresponds with a heart rate of 60 and an RPE of 20 with a heart rate of 200. An RPE of 6 corresponds with no exertion at all, and an RPE of 20, so related to a heart rate of 200, is a maximal effort. Another common RPE scale is the Borg CR10, a Category-Ratio (CR) scale between 0 and 10. In this scale, 0 corresponds with no exertion at all, and exercising at an RPE of 10 is a maximal exercise (Williams, 2017).

RPE in this work

RPE is a reliable variable to use as a measure of internal training load. Training load will be calculated as RPE times duration, also called session RPE, and the obtained training load variable will be used in this work.

Questionnaires

One can also use subjective variables obtained from questionnaires in addition to all these objective variables. For instance, athletes can be asked how ready they are to train, how sore their muscles are, and how they experienced their sleep duration and quality from a scale of one to five. A lot of different questionnaires are available to analyze athletes daily; the Profile of Mood States (POMS) (Douglas M. McNair, n.d.), Daily Analyses of Life Demands for Athletes (DALDA) (Rushall, 1990) and the Recovery-Stress Questionnaire for Athletes (RESTQ-Sport) (González-Boto et al., 2008) are commonly used psychological tools to monitor training load (Nässi et al., 2017). By using subjective questions, perceived fatigue is measured instead of performance fatigability.

Hamlin et al. (2019) researched training load and stress in young elite university athletes participating in all kinds of different sports in America. In their research, they mainly studied subjective variables and with a single logistic regression, they found a negative correlation between the daily questionnaire and the odds of injury. Hence, lower levels of mood or sleep duration and increased levels of energy or stress were able to predict injury. Next to Hamlin et al. (2019) other works used subjective questionnaires next to other variables to conclude training load and fatigue. For example, Mendes et al. (2018) found a relation between the daily wellness of volleyball athletes and the number of competitions in a week. Three days before the second match in a week a lower daily wellness was found than in weeks with only one competition.

Questionnaire in this work

In this work, none of the questionnaires mentioned above will be used. The questionnaire that will be used in this work is a questionnaire that has already been used within TalentNED for the last year. For continuity, this questionnaire will be used. This questionnaire consists of questions about sleep quality and duration, feelings of fatigue, stress, mood, and readiness to train. The questionnaire looks like a simplified version of the DALDA questionnaire and can be found in Appendix A.

2.1.2 External load

The external load can be measured in different ways for particular sports. For example, one can monitor the *power output* obtained during cycling, while one can monitor the *pace* or *speed* obtained for activities such as running or speed skating.

Neuromuscular function

One can inspect the neuromuscular function as well to investigate fatigue in athletes. Neuromuscular fatigue is the reduction in the maximal force a muscle can exert, or the inability to sustain exercise at a required power (Bestwick-Stevenson et al., 2022). One common test type for fatigue is a jump test. Variables such as jump height, flight time, mean power, peak power, and peak velocity could be measured during jump tests to determine fatigue. In previous research, it has been shown that athletes jump lower when one is more fatigued (Halson, 2014).

As can be seen in Table 1; Gupta et al. (2023), Pupo et al. (2021), Gavanda et al. (2023) and Coutts et al. (2007) used a variant of a jump test to determine fatigue in athletes. Next to the increase in heart rate of the cheerleaders, Gavanda et al. (2023) also found a decrease in jump height during the fatiguing weeks before the World Championships.

Pupo et al. (2021) did research with athletes of different sports to analyze the relationship between the vertical jump and performance in the physical sports judo, futsal, and sprinting. Jump height was measured during a vertical jump, countermovement jump, and squat jump. With Pearson linear correlation coefficients, a relation was found between the performances of sprinting athletes on 20m and 200m and their results on these jump tests was found. After normalizing the obtained results to reduce the effect of individual body mass, Pupo et al. (2021) concluded that jump height and power output of vertical jumping tests are similar and positively correlated with the physical performance tests for all analyzed sports.

Coutts et al. (2007) monitored the recovery of triathletes for a time period of one training block, consisting of four weeks of training and two weeks of tapering. They measured the data of an overload group, following an overload training schedule and a control group, with a normal training schedule. Data from a, twice per week performed, jump and a sub-maximal heart rate test were analyzed using independent t-tests, together with data from the DALDA questionnaire. The athletes performed a 3km time trial once a week as a performance measure for an indication of fatigue. The results of this 3km time trial were compared with the 3km time trial before the fatiguing protocol with a two-factor analysis of the covariance test. The athletes following the overload training were significantly slower in this post-test than in the pre-test. No decrease in performance on the 3 km time trial was found for the control group., which showed that the athletes in the overload group

indeed were fatigued. The distance reached in the five-bound jump test was reduced for the group that followed overload training and was not significantly changed for the group following regular training. As the overload group was fatigued after the overload training, Coutts et al. (2007) showed that a five-bound jump test was able to identify fatigue in the athletes of the overload group.

Neuromuscular function in this work

In this work, a countermovement jump (CMJ) test is used as a measure of neuromuscular function and where maximal jump height is the only measured variable. The CMJ is easy to perform without restrictions and has high repeatability. The CMJ is a vertical jump where the movement starts with a rapid downward motion (the countermovement) followed immediately by an explosive upward jump to achieve maximum height. The reason to only measure the jump height of the athletes is that they are capable of collecting this variable on their own without large measuring errors. In this work, the approach of Coutts et al. (2007) is used as inspiration for the data protocol explained in Section 3.3.

TABLE 1: Selection of relevant related work in the field of this work.

Authors	Sport	Interesting parameters	Results
Only used variables related to internal load			
Hamlin et al. (2019)		Subjective variables from questionnaires	Negative correlation between the daily questionnaire and the odds of injury
Mendes et al. (2018)	Volleybal	Training load and wellness	Wellness of athletes is best on match days and worst in weeks with two or more competitions
Otter et al. (2022)	Speed ice skating	Intended, perceived RPE and RESTQ-questionnaire	When athletes approach or exceed the intended duration per session, their perception of physical recovery and self-regulation is improved
Knobbe et al. (2017)	Speed ice skating	Training, competition and test data	Training load has a positive correlation with long-term fitness. Findings are athletic-specific and patterns should not be interpreted as general rules of exercise
Only used variables related to external load			
Gupta et al. (2023)	Ice hockey	Jumping tests, acceleration tests and repeated sprint analyses	A relation between the on-ice and off-ice performance variables was found
Daigle et al. (2022)	Ice hockey	Upper body strength test, a pull-up test, lower body muscular power test, on-ice tests and performance	They concluded that the broad jump test results have a positive correlation with on-ice skating speed, in the forward and backward skating test.
Pupo et al. (2021)	Judo, futsal and sprinting	Jump height and performance	Jump height and power output of vertical jumping tests are similar and positively correlated with the physical performance tests
Used both variables related to internal and external load			
Gavanda et al. (2023)	Cheerleading	Daily heart rates, jump test questionnaires	During fatiguing training weeks the heart rate of the cheerleaders increased over time and their jump height did decrease as well
Coutts et al. (2007)	Triathlon	Jump test, sub-maximal heart rate test, DALDA questionnaire and time trials	A five-bound jump test was able to identify fatigue in the athletes of the overload group.

2.2 Applications of machine learning in health and sports research

In this section, some works will be described that already use machine learning techniques within health and sports situations to predict and classify. First, articles in the health domain will be discussed, following some predictive heart rate models and lastly some articles within the sports and performance domain will be considered.

2.2.1 Health care

Recently, research has been done on data from COVID-19 patients. Pasic et al. (2022) and Giotta et al. (2022) both used machine learning techniques, such as Neural Networks and Decision Trees respectively, to predict the outcome of these COVID-19 patients. Within health care, with the outcome of a patient, the results from care and treatments those patients have received whilst in hospital are meant. In the following two studies, the possible outcomes are: discharged alive or death. Pasic et al. (2022) looked at whether a combination of Neural Nets, hypothesis testing, and confidence intervals could help physicians in their work of nursing COVID-19 patients. As input variables, initial laboratory findings, demographics, and comorbidities were used, and a precision of almost 97% in predicting a patient's survival or death.

Giotta et al. (2022) did research into the application of a Decision Tree to predict the outcome of COVID-19 patients. With this Decision Tree method, predictive variables were found, and on a validation set the decision tree model reached a sensitivity of 99% in predicting survival or death. A high sensitivity is preferable for both Pasic et al. (2022) and Giotta et al. (2022), because as few false positives as possible are desired for the outcome of COVID-19 patients.

Antwi-Afari et al. (2023) and Bustos et al. (2022) have used Machine Learning techniques to model physical fatigue in construction workers and firefighters respectively. Antwi-Afari et al. (2023) has tested Artificial Neural Network (ANN), Decision Tree, Random Forest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) on the input of wearable insole devices to model physical fatigue within construction workers. They found that the best results were achieved by the Random Forest model with an accuracy of 86%.

In another field of work, Bustos et al. (2022) also tested different models, such as KNN, Boosted Trees, Bagged Trees, Random Forests, SVM, and ANN to predict fatigue. Bustos et al. (2022) used sensory data from heart rate, breathing rate, and core temperature from 24 firefighters during an incremental running protocol to model physical fatigue. The best model they obtained was the XGBoost classifier, a variant of Boosted Trees, with an accuracy of 82%.

2.2.2 Heart rate prediction

Oyeleye et al. (2022) and Luo and Wu (2020) looked at predicting heart rate using different Machine Learning techniques. Oyeleye et al. (2022) compared autoregressive integrated moving average (ARIMA) model, linear regression, support vector regression (SVR), KNN regressor, decision tree regressor, random forest regressor and Long Short-Term Memory (LSTM) with as input variable different lengths of heart rate recordings. They found that the ARIMA and linear regression models were the best to predict future HR with any given HR recording length. For HR recordings of less than one minute, especially the KNN, LSTM and random forest models were not good.

Luo and Wu (2020) were able to find a multivariate LSTM model that with input variables heart rate, gender, age, physical activities, and mental state was able to predict the heart rate. They used an Adam optimizer, which resulted in high validity and a root mean square error of less than 0.5 beats per minute (bpm).

2.2.3 Sports performance

Within monitoring in the sports domain, Knobbe et al. (2017), De Leeuw et al. (2023) and Wang et al. (2023) all used Machine Learning techniques to monitor the fitness and/or health of an athlete. Wang et al. (2023) used data from wearable devices during training and or competition with a model that combines Convolutional Neural Network (CNN), with LSTM and self-attention mechanisms to predict the health status of an athlete. Their model achieves an accuracy of 0.93, and also the specificity, precision, and F1 score are above 0.9.

To improve speed ice skating training programs Knobbe et al. (2017) analyzed fifteen years of historical training, competition, and test data of a Dutch professional ice skating team. Using aggregation techniques together with Linear Regression and Subgroup Discovery, they were able to extract actionable and meaningful patterns that can be used to improve training schedules for this ice skating team.

De Leeuw et al. (2023) presented a model for monitoring fitness in road cycling only dependent on sensor data collected during bike rides. This model, unlike previous approaches, does take into account the effect of earlier training days on the fitness of the athlete after a training or competition. With this addition, the model of De Leeuw et al. (2023) has an explained variance of 0.86.

Instead of using machine learning techniques to monitor the fitness and/or health of an athlete, Campbell et al. (2021) aimed to predict training load using wellness questionnaires in different sports. With input variables such as sleep quality, readiness to train, general muscular soreness, fatigue, stress, and mood machine learning models were trained on classifying the training load of athletes. Different approaches such as regression, classification, and random forest models were used, but all gave a low accuracy on the classification problem. Campbell et al. (2021) conclude that their results suggest that wellness items have no predictive capacity towards training load.

2.3 Conclusion - literature review

The first two sub research questions are already answered in this literature section. Therefore, in this subsection, a first conclusion and some hypotheses will be given related to the first two sub research questions.

1) What are relevant variables to describe the training load of athletes?

According to the literature as described above, important variables to describe training load of athletes can be divided into internal training load and external training load variables. An overview of these variables is shown in Figure 1.

2) How do the existing machine learning methods perform on the prediction of heart rates?

The current machine learning and deep learning methods that are already used for the prediction of heart rates do show some promising results. The best methods found in the literature are; ARIMA, linear regression, and a multivariate LSTM model (Luo & Wu, 2020; Oyeleye et al., 2022).

Given all the works above, the connection between fatigue and jump height and/or resting heart needs to be examined further in detail. Next to these two variables, Hofman et al. (2017) showed that specific for speed skaters the performance on a Wingate test is a predictive variable for performance on ice. Therefore in this work the connection between power output on the Wingate test and fatigue will also be investigated. The classification decision tree model as described by Campbell et al. (2021) will be used for a better insight of the daily morning questionnaire. Finally, the multivariate LSTM model of Luo and Wu (2020) shows promising results for predicting heart rates. No model tried to predict resting heart rates yet, therefore this work tries to see if the LSTM method of Luo and Wu (2020) has potential to predict daily resting HR.

3 Dataset collection and curation

This work is based on a dataset of a junior sport development team 'TalentNED' with a speed skating and mountain bike team. The participants in this study are all members of the speed skating team and on weekdays live and train together. In this section, the subject information of this research is given and the dataset already present within TalentNED is described. Next, a fatigue data collection protocol is presented to collect some extra data related to fatigue and finally data processing will be discussed.

3.1 Subjects

In total twelve junior elite speed skaters participated in this research. The characteristics of the subjects are shown in Table 2. All of the athletes are top-level athletes and perform on the highest (inter)national ice skate level in their age category. In this work, they will be classified as elite junior speed skaters.

TABLE 2: Characteristics of the subjects.

Variable	Male athletes	Female athletes
Total (n=12)	(n=8)	(n=4)
Age (years)	17.63 ± 0.92	17.50 ± 1.0
Height (cm)	183.38 ± 3.29	169.50 ± 5.8
Weight (kg)	78.48 ± 5.47	64.58 ± 6.36
Sprinter	(n=2)	(n=0)
All round	(n=6)	(n=4)

3.2 Available data within TalentNED

TalentNED

Starting April 2023, data has been collected for the twelve speed skaters. All this data is stored and analyzed in an online platform named Sport Data Valley (SDV). Some analyzing methods are already provided within this platform. Athletes and coaches can visual see the results of the questionnaires in graphs. Next to the questionnaire data, they can see the power output, speed and height profile of a bike ride once a smartwatch is connected to the platform. These graphs and analysing tools are all fixed within the platform, and can thus not be changed per user. However, a Jupyter environment is available where one can write their own analyses using all the available data with for example Python (Van Rossum & Drake Jr, 1995). SDV aims to provide a data-privacy-ensured platform that can be used by a broad audience.

Questionnaires

A total of four different questionnaires were taken by the athletes.

- Daily questionnaire: A daily questionnaire is taken in the morning, with questions regarding sleep duration/quality, soreness, and readiness to train (Appendix A 8.1).
- Training log: After each training session, the athletes are asked to fill in their training log. In the training log, they can specify the kind of training, the duration of the session, and an RPE for the training, indicating how hard the training was (Appendix A 8.2).
- Evening questionnaire: Each evening, the athletes get a questionnaire asking how stressful their day was on three different levels; physical, mental, and total (Appendix A 8.3).
- Weekly questionnaire: Every week the athletes are asked whether they are injured/ill and whether they missed any training sessions because of this (Appendix A 8.4).

Training load

The intended training load is known for all training sessions. This training load is calculated by the training duration multiplied by the intended RPE, determined by the coach. The perceived training load is a known variable, which is calculated by the variables, training duration, and RPE, obtained from the training log. Next to the training load, it is known which kind of training took place on which day, and therefore on which days multiple training sessions took place.

Wearable devices: Smartwatches

Next to the questionnaires, all the athletes are wearing a smartwatch from the brand ‘Polar’ or ‘Garmin’. From this device, it is possible to obtain their daily resting heart rate. Resting HR is therefore one of the variables stored on a daily basis too. For the Garmin smartwatches, the resting HR is the average heart rate of the 30-minute lowest heart rate over the day.

Wearable devices: Heart rate band

The athletes wear a heart rate band of either ‘Garmin’ or ‘Polar’ during all training sessions.

Wearable devices: Power output pedals

Next to heart rate data, power data is collected with Garmin Rally (Garmin Inc., Wichita KS, USA) for training sessions on the bike. With a Garmin Edge (Garmin Inc., Wichita KS, USA), these power data are collected and stored in SDV. For the training sessions on the bike, the heart rate and power data can be combined for further analysis.

Wearable devices: Transponder

The athletes are also wearing a transponder during the ice skating training, which is connected to MyLaps (MYLAPS Sports Technology, Haarlem, The Netherlands). With this timing system, lap times are measured and saved for all athletes for the whole training session.

Tests: Powerpeak

Next to the data collected during training sessions, four times a year a six-second Powerpeak test on a Watt bike (Wattbike Ltd, Nottingham, UK) is performed. During this test, the athletes have to do an all-out sprint for six seconds (Herbert et al., 2015). The following variables are collected for this test:

- Peak power: Highest power output obtained during the six seconds
- Mean power: Average power output obtained during the six seconds
- Body Mass: Before the Powerpeak, body mass is measured, such that Peak power and Mean power could be normalized for analysis.

Tests: Wingate

Somewhat less frequent, around three times a year, a Wingate test, also on a Watt bike (Wattbike Ltd, Nottingham, UK), is performed. During a Wingate test, the athletes have to perform a 30-second all-out performance at maximal speed (Bar-Or, 1987). The following variables are collected in the Wingate protocol:

- Peak power: Highest power output obtained during the 30 seconds
- Mean power: Average power output obtained during the 30 seconds
- Fatigue index: Drop off of mean power output, calculated by dividing the mean power output of the first five seconds by the last five seconds, represented in a percentage.
- Per 5 seconds: The average power output for each interval of five seconds during the Wingate is measured.
- Body Mass: Before the Wingate, body mass is measured, such that Peak power and Mean power can be normalized for analysis.

Both the Wingate test and the six-second Powerpeak tests provide valid measures of peak power output (Herbert et al., 2015). However, the six-second Powerpeak test is less invasive as it involves a shorter duration of intense exercise than the Wingate test.

3.3 Data collection for monitoring fatigue

To answer the sub-research question: ‘Can fatigue be identified by monitoring daily jump height and by wellness questionnaires?’ the variable *daily jump height* should be added to the existing dataset. To obtain this new dataset, there will be a data collection protocol of five weeks. These five weeks consist of two training blocks of two weeks and the first week of a third training block. These training blocks are centered around the first competitions of the ice skating season. A schematic overview of the training sessions in the training blocks is shown in Figure 2.

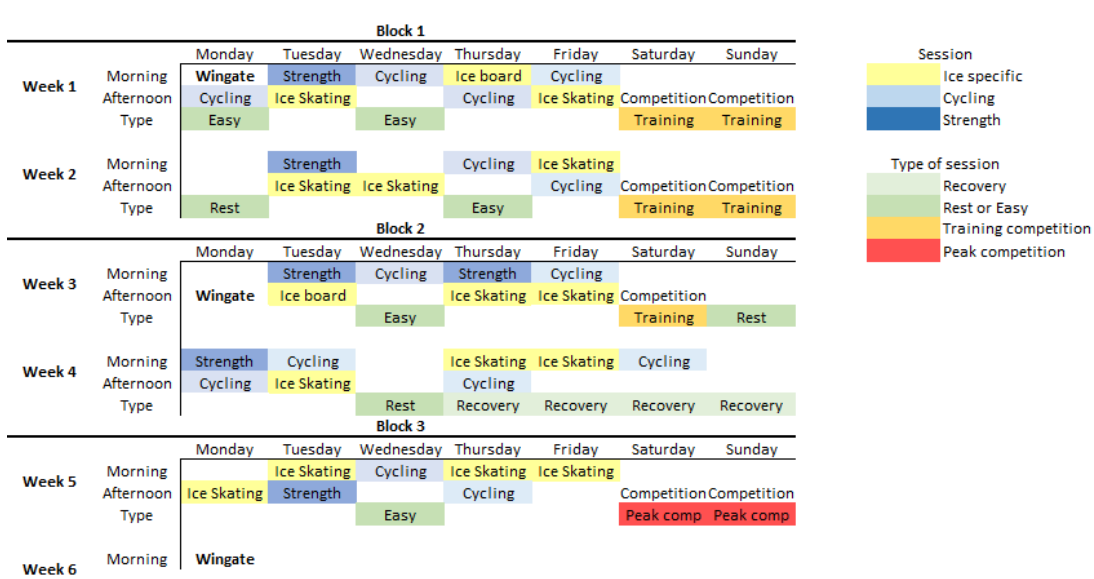


FIGURE 2: Schematic overview of the different training sessions during the data collection protocol.

The collected variables in this protocol can visually be found in Figure 3 and consist of *daily questionnaire*, *resting HR*, *a maximal jump test* and *a Wingate test*.

Questionnaire

The questionnaire taken is the daily questionnaire described in Section 3.2, which has already been used for some months. Interesting variables collected from this questionnaire are:

- readiness to train
- fatigue
- soreness
- mood
- stress
- sleep duration
- sleep quality

Resting HR

Resting HR is a variable that is collected via a wearable device, the smartwatch. All the athletes wear their smartwatch day and night and resting HR is calculated over the current day.

Maximal jump test

To measure the jump height, a vertec measuring tool is used, and the jump height is measured by the highest bar the person is capable of reaching. The vertec measuring tool consists of 60 bars, with between every bar 1 cm, such that in total the difference between the lowest and highest bar is 60 cm. Daily jump height is then measured as the difference between the height of the highest bar reached minus the length of the person with a stretched arm. The athletes will get three tries for a maximal jump and only the highest jump height will be used as a variable.

Week 1,3,5	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
	Questionnaire	Questionnaire	Questionnaire	Questionnaire	Questionnaire	Questionnaire	Questionnaire
	Resting HR	Resting HR	Resting HR	Resting HR	Resting HR	Resting HR	Resting HR
	Jump test	Jump test	Jump test	Jump test	Jump test		
	Wingate						
Week 2,4	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
	Questionnaire	Questionnaire	Questionnaire	Questionnaire	Questionnaire	Questionnaire	Questionnaire
	Resting HR	Resting HR	Resting HR	Resting HR	Resting HR	Resting HR	Resting HR
	Jump test	Jump test	Jump test	Jump test	Jump test		

FIGURE 3: Protocol of fatigue data collection.

Wingate test

At the start of weeks 1, 3, and 6, a Wingate test is performed to obtain data about the anaerobic capacity of the athletes. From the Wingate test, the interesting variables are mean power output, peak power output, the fatigue index, and the progression of power output per five seconds. In consideration with the coaches, it is decided to perform the Wingate test on Monday morning, at the start of the new training week.

The athletes are familiar with the data collected as described in Section 3.2; *daily questionnaire*, *resting HR*, and *Wingate test*. Because the athletes are not yet familiarised with the reach and jump test used in this protocol, the athletes will have a two-week familiarisation period for the *jump test* before the start of the protocol.

3.4 Collected dataset

In Table 3 the number of data points per athlete per variable could be found. The data collection protocol eventually took place between September 11 and October 16. The data collection is therefore extended by one day, as during the protocol the decision was made, in collaboration with the coaches, to perform the last Wingate at the start of week 6. This deviates from the data collection protocol, which was to perform the last Wingate at the start of week 5, as can be seen in Figure 3.3

TABLE 3: Compliance of the collected variables during the data collection protocol.

Owner ID	Daily Questionnaire	Jump Height	Resting HR	Wingate
1819	33	14	35	3
2577	35	15	0	2
2583	33	12	15	3
1817	34	10	29	2
511	30	10	31	3
1818	32	17	27	3
1629	34	9	24	2
2573	35	15	24	2
2572	26	17	30	3
1319	35	18	35	3
2575	34	19	34	3
1821	32	12	33	2
Total	35	35	35	3

As can be seen in Table 3 the athlete who performed the most jump tests, still only performed a jump test on 19 out of the 35 days. Excluding weekend days, the athletes were expected to jump 26 times and only four athletes were able to miss less than 10 days of the jump test. To obtain a dataset with more data points for the jump test, *daily questionnaire*, *resting heart rate* and *daily jump height* of the two weeks which were meant as familiarisation for the jump height are also used. This dataset consists of the amount of data points as can be seen in Table 4.

TABLE 4: Compliance of the collected variables during the familiarisation weeks of the jump test protocol.

Owner ID	Daily Questionnaire	Jump Height	Resting HR
1819	15	13	15
2577	15	13	5
2583	14	13	15
1817	15	13	15
511	15	12	13
1818	15	13	11
1629	15	13	15
2573	15	13	10
2572	13	10	13
1319	15	0	14
2575	15	13	15
1821	15	13	15
Total	15	13	15

The training sessions as shown in Figure 2 are converted to values. Each day has been split up into two variables: the number of training sessions, and the intended RPE. The same is done for the training sessions which took place during the familiarisation period of the jump test.

For the Wingate tests, the following variables are stored in a dataset:

- Peak power
- Mean power
- Fatigue Index
- 1"-5"
- 6"-10"
- 11"-15"
- 16"-20"
- 21"-25"
- 26"-30"
- Peak power /kg
- Mean power /kg

3.5 Data preprocessing

The four main groups of variables, *jump height*, *answers on the questionnaires*, *Wingate performance* and *resting HR* are all collected in different ways and therefore are stored as different data streams.

Ideally, one data point per day is available for the resting HR and the answers from the questionnaire. This stream of data can be seen as a continuous time series. Next to resting HR and questionnaire data, the daily height jump test is a time series, but this is not a continued time series as the weekend results are missing. As the last collected variable, peak power output, mean power output, and the fatiguing index are collected by a Wingate test. These variables can be seen as three separate measurement points.

After collecting the data, some preprocessing steps took place before analyses could be done on the dataset. First of all, data cleaning was performed where duplicate or missing data was handled (Fan et al., 2021). Duplicate data was sometimes present for the resting HR data and then the data with the lowest resting HR was kept. Missing data could occur in the variables: *daily questionnaire*, *jump height*, and *resting heart rate*. If one of the first two variables is missing it is chosen to do nothing and thus not replace values for these Not A Numbers (NaNs). If the resting HR is missing this value is replaced by the average value of the resting HR of that athlete. Now on successive time periods, a resting HR value is present and therefore this data stream can be seen as a time series.

The second step is data integration. Here the different datasets are combined into one dataset. The first two datasets containing data from the questionnaire, resting heart rate, jump test, and Wingate tests are all variables collected, primarily, in the morning or during the day. The information about the training sessions on the day self is a variable collected in the evening. This variable will have an influence on the fatigue of the athlete the day after the variable is collected and therefore a time shift, of one day, is applied when combining the datasets.

After data integration, the next step is data reduction. Nothing related to data reduction is done for this dataset.

The final step in the preprocessing is data transformation. Within data transformation, mainly feature construction took place. With feature construction, new features (variables) are created from the existing ones in a dataset with the goal of improving the performance of machine learning models (Zheng & Casari, 2018). For all the variables from the questionnaire, the following extra variables are created:

- *Variable $t - 1$* , the value of one day before
- *Variable $t - 2$* , the value of two days before
- *Variable $t - 3$* , the value of three days before
- *Variable avg_{t3}* , the average value of the past three days
- *Variable avg_{t5}* , the average value of the past five days

After the data preprocessing steps the final dataset is obtained: ‘Fatiguing athletes’, consisting of 9 variables for which feature construction took place. For the variables *Readiness to train* and the variables for the jump test and Wingate tests no feature construction happened. In total, the dataset now consists of 67 features.

4 Data analysis

In this section, the different tests applied to the dataset will be explained. First, the techniques used to determine the importance of different variables will be introduced. Next, the statistical analysis of the data for the Wingate tests will be discussed. Lastly, in the third subsection, a model is presented to predict daily resting heart rates.

4.1 Importance of variables

Next to the literature review of which variables are important in monitoring and predicting fatigue, the importance of features is also ranked for the 'fatiguing' dataset. For four different dependent variables; daily jump height, daily resting HR, daily feeling of fatigue, and daily readiness to train, a classification decision tree is trained to classify the dependent variable. The 'fatiguing' dataset consists of wellness questionnaires and a subjective variable; daily resting HR. This dataset matches the dataset that Campbell et al. (2021) used for predicting an objective variable, namely training load. Therefore, for our classification problem, a classification decision tree method can be used too. In this subsection, a classification decision tree method with k-fold cross-validation is described with an adjustment towards only using variables with the same probability distribution.

4.1.1 k-fold

The first step in training the decision tree model is with k-fold cross-validation, where k equals three. Better performance can be obtained using k-fold cross-validation and it is a model that is more robust for new data. The model is trained k times, each time using $k - 1$ folds for training and the remaining fold for testing. This is done for each athlete individually, meaning that the decision tree model differs for each athlete. Moreover, knowledge about the athlete is necessary before the decision tree method can be applied.

4.1.2 Leave-one-user-out

The next step is to create a more generalized model, which is even more robust against newly seen data. Therefore, k-fold cross-validation is still used, but now on the data of all twelve athletes together. In this model k equals twelve, meaning the decision tree model is trained on the data of the other eleven athletes and the test set is the resulting athlete. The advantage of k-fold cross-validation in comparison with the leave-one-user-out approach is that a model is trained on data of only that athlete. The goal of this is to let the model be more specific to the characteristics of the athlete. For a new athlete, it is better to use the leave-one-user-out approach to make predictions, as there is no data yet of the new athlete to train the model on. A disadvantage is that the model is trained on the general group of athletes and no personal characteristics are taken into account. Another advantage of the leave-one-user-out approach is that there is more data for the model to train on, which will improve the performance of the model.

4.1.3 Train on variables with the same distribution

For the model to perform better, the distribution of the data of the training and test set should match. Otherwise, the model will not be able to give a good classification of the dependent variable of the test set. To ensure that these distributions match, an extra preprocessing step is added to the process. A statistical test, the two-sample Anderson-Darling test, is used after the train and test set are constructed. This test, which assesses if two samples come from the same distribution, is added to the variables of the training and test set. In this test, the null hypothesis is be that the two samples come from the same distribution, and the alternative hypothesis is be that the samples come from a different distribution.

For the Anderson-Darling test, the test statistic is given by:

$$A^2 = -\frac{N}{N_1 N_2} \sum_{i=1}^N [(2i-1) \cdot (\ln F_1(X_i) + \ln 1 - F_2(X_i))],$$

with:

- N denotes the total sample size
- N_1 and N_2 are the sample sizes for the first and second samples
- $F_1(X_i)$ is the empirical cumulative distribution function of the first sample at the i -th ordered observation
- $F_2(Y_i)$ is the empirical cumulative distribution function of the second sample at the i -th ordered observation
- \ln is the natural logarithm

The test statistic A^2 is then compared to the critical values from the Anderson-Darling distribution to decide whether to reject the null hypothesis. If the null hypothesis can be rejected, this means that the two samples are not coming from the same distribution (Razali & Wah, 2011). For each feature, this test is done and only the features for which the null hypothesis could not be rejected are used for the classifying decision tree method. Not being able to reject the null hypothesis does not mean that the samples do or do not come from the same distribution. Still, it does give a better indication of the underlying distribution of the samples.

4.2 Wingate performance to identify fatigue

To answer the sub-research question if the performance of a Wingate test could predict fatigue, statistical analyses are done on the results of the three Wingate tests. With performance of a Wingate test, the variables obtained during the Wingate test are meant. In this subsection the connection between power output obtained of a Wingate test and fatigue will be investigated further.

Only data is used from the athletes who did all three of the Wingate tests, so eventually, a dataset of seven athletes was used. Before applying statistical tests, it is important to visualize the dataset. Therefore, in the results section, a visualization of the dataset will be given first. Furthermore, in this section, the following tests will be described and in the results section they will be evaluated on the Wingate dataset; standardized t-test, Kruskal-Wallis test, and the Wilcoxon Signed Rank test.

4.2.1 Standardized t-test

After visualizing, a standardized t-test (Kim, 2015) was performed to verify if the results of the Wingate test differed at the start of the data collection protocol compared to the results of the Wingate test at the end of the data collection protocol. The formula for the standardized t-test is expressed as:

$$Z = \frac{X - \mu}{\frac{s}{\sqrt{n}}},$$

with the following parameters.

- Z is the standardized test statistic
- X is the sample mean
- μ is the population mean under the null hypothesis
- s is the standard deviation of the sample
- n is the sample size

The resulting Z value will be compared to critical values from the standard normal distribution to determine statistical significance. A large Z value indicates a significant deviation from the population mean, suggesting that the sample mean is unlikely to have occurred by chance. The p-value will be calculated using the Z value and the distribution under the null hypothesis and statistical significance will be found if the p-value is less than a pre-determined significance level (α) (Kim, 2015).

4.2.2 Kruskal-Wallis test

It is not known if the data of the Wingate tests satisfies the assumptions of a parametric test. Therefore, next to the standardized t-test also a non-parametric test, the Kruskal Wallis test is executed. This test is used to test if there are statistically significant differences among three or more independent groups. The test involves ranking all observations across groups, calculating the sum of ranks for each group, and then computing a test statistic H. The formula for H is given by:

$$H = \frac{12}{N(N+1)} \sum_i \frac{R_i^2}{n_i} - 3(N+1),$$

with the following parameters.

- N is the total number of observations
- R_i is the sum of ranks for the i -th group
- n_i is the number of observations in the i -th group

Just like the standardized t-test, a p-value is calculated and statistical significance will be found if the p-value is less than a pre-determined significance level (α). The null hypothesis for this test is that the population medians are equal, so for a p-value less than α we can assume that the population means are different.

4.2.3 Wilcoxon Signed Rank

The Kruskal-Wallis test is a test for independent groups. However, the three particular groups consist of the performance data on the three Wingates of the same group of athletes. Therefore, it may be better to use a dependent test to check for differences in the means. Still, a non-parametric test was used, but the decision was made to use the Wilcoxon Signed Rank test (Durango & Refugio, 2018). This test evaluates whether there is a significant difference between paired observations. The test statistic W is calculated as the minimum of the sum of the ranks of positive and negative differences. Statistical significance is then determined by comparing the test statistic to the critical values from the Wilcoxon Signed Rank distribution.

$$W = \min(\sum \text{Ranks of positive differences}, \sum \text{Ranks of negative differences})$$

All these tests will only be performed on the single input features, like the power output of the Wingate test or the jump height of the CMJ test. The same statistical tests, with small adaptations, could also be performed on a combination of variables, while this is not done in this study it will be discussed further in Section 6, the Discussion.

4.3 Resting HR prediction

One of the research questions as defined in the Introduction, Section 1.2, is specific on predicting daily resting HR data. The goal of predicting daily resting HR is to check if the predicted resting HR deviates from the measured resting HR. If a large deviation is occurring, this could be a sign of the start of an elevated resting HR and thus a sign of non-functional overreaching. To predict the resting HR of an athlete for the next day, a LSTM network is used. In this subsection, an LSTM network is explained, and two variants of this network are described in detail. In the results section, these two networks will be performed and evaluated.

LSTM is a type of Recurrent Neural Network (RNN) architecture employed for time series forecasting. An LSTM model works like a smart tool that remembers important details and decides what to focus on from the past. It uses special memory cells and gates to store and recall information selectively from previous time points. The model figures out how much importance to give to each piece of historical data, allowing it to understand complex patterns in the time series (Anani, 2018).

Daily resting HR can be seen as a time series, so an LSTM model would be a logical tool for predicting resting heart rate. As a basis, the model proposed by Luo and Wu (2020) is used such that the LSTM network consists of an input layer, a hidden layer, and lastly a connected layer. The structure of the LSTM network can be seen in Figure 4 and is described below. The input layer has a dimension of the input data of 50 and an activation function of tangent hyperbolic. The hidden layer is a combination of a fully connected dense layer with a dropout rate of 0.2 to prevent overfitting (Srivastava et al., 2014). At last, a fully connected layer is added, with an activation function of ReLU to connect the hidden layer and the output.

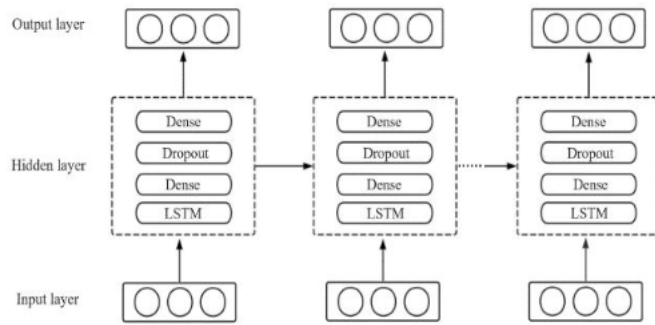


FIGURE 4: Visualisation of the LSTM network as used by Luo and Wu (2020).

The LSTM network can be used for an univariate time series problem as well as for a multivariate time series problem. For the univariate time series problem, as an input variable only resting HR data is used to predict a new resting heart rate, while for the multivariate time series problem also other variables could be taken into account as input variables. In this work, both methods have been used and are respectively described in the following two subsections.

4.3.1 Training parameters

A training set of 80% and a test set of 20% were used as training parameters for all three methods. The random state was set at an integer, such that the same randomization was used each time the model was run. Fitting the model is done in 50 epochs with batch size 16. The look-back parameter of the model determines the number of previous resting HR used as input to predict the next resting HR. This variable is set on five days in all the models.

4.3.2 Univariate Time Series

The first LSTM model that was trained used previous heart rate samples of an athlete to predict the next resting heart rate. To generalize the model and provide more training data a model was created that predicts the resting HR of one athlete trained on the resting HR data of the other eleven athletes. The model weights are set by the training set, consisting of resting HR data points of eleven athletes and the model is tested on the test dataset. The test dataset consists of the raw dataset of HR points as well as a dataset where smoothing has been applied. As a smoothing technique, exponential smoothing is used with different values for α , the smoothing factor.

With exponential smoothing, data points are smoothed by assigning exponentially decreasing weights to past observations. The new smoothed value will be calculated as follows:

$$\hat{y} = \alpha \cdot y_t + (1 - \alpha) \cdot \hat{y}_{t-1},$$

with the following parameters:

- y is the actual observation at time t
- α is the smoothing parameter, where a smaller α gives more weight to past observations, while a larger α gives more weight to the most recent observation (Shan et al., 2023)

4.3.3 Multivariate Time Series

Instead of only using previous resting HR data to predict resting heart rate, a multivariate time series model takes into account more input variables. A first multivariate time series model is trained on the following input variables; the daily readiness to train and feeling of fatigue together with the RPE and number of trainings yesterday. The same training and testing distribution as in the second univariate time series model is chosen, such that the model is generalized and can predict resting HR data of unseen athletes. Here as well the test set consists of the raw dataset and a smoothed dataset.

4.3.4 Multivariate Time Series with top 10 relevant features

In the previously described multivariate time series model the input variables are manually chosen. To improve the prediction of the resting heart rate, a new model is trained on variables that have the biggest correlation with resting HR. Therefore the features that come back in the top ten for predicting resting HR with the highest importance are chosen from the previously described decision tree method.

5 Results

In this section, all the results of the tests and models described in Section 4 are stated. The interpretation of these results will be given in Section 7. The structure of this section is the same as the structure in the method section. It starts with an interpretation of the dataset, following by the results about the importance of the variables, the statistical tests on the Wingate performance, and ends with the model presented to predict daily resting heart rates.

5.1 ‘Fatiguing dataset’

Before giving the results to the subresearch questions, first in this subsection, an overview of the complete dataset is given, with descriptive analysis and data visualization.

5.1.1 Descriptive analysis

As stated in Section 3.3 the collected dataset now consists of 67 features. In Table 3 and Table 4 the number of completed questionnaires, performed jump and Wingate tests, and the collected daily resting heart rates can be found. This adds up to a total of 570 completed questionnaires, 307 jump tests, 473 daily resting heart rates, and 31 results of a Wingate test from 12 different athletes.

5.1.2 Data visualisation

In Figure 5 four histograms could be seen of the variables; *readiness to train*, *daily fatigue*, *normalized max jump*, and *resting HR*. For *readiness to train* and *daily fatigue* on the x-axis the different answers are given and on the y-axis the frequency of how many times that answer is given. For *daily resting HR* and *daily max jump* the x-axis shows the range of all the possible values these variables could take and the y-axis shows the frequency of how many times that value could be found in the dataset.

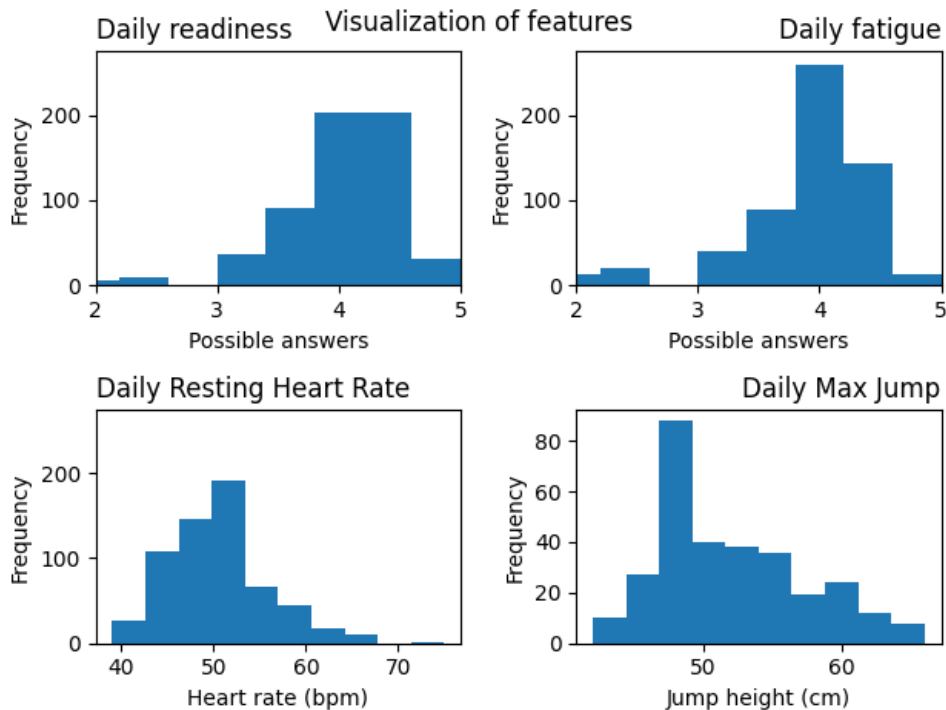


FIGURE 5: Visualization of the four variables; readiness to train, daily fatigue, normalized max jump, and resting HR over the whole dataset.

Inspecting the data in this way shows some interesting details. For the two variables of the daily morning questionnaire *daily fatigue* and *daily readiness*, the most common answers given are 4.0 and 4.5. This means a low feeling of fatigue and a high feeling of readiness to train. Something else worth noting is that the tail of the distribution is light-distributed, answers given below 3.0 or higher than 4.5 don't occur that often. The distribution of the data is different for all four variables. The distribution of the *daily max jump* seems to follow a normal distribution, if the peak at a jump height of 48 or 49 cm is neglected.

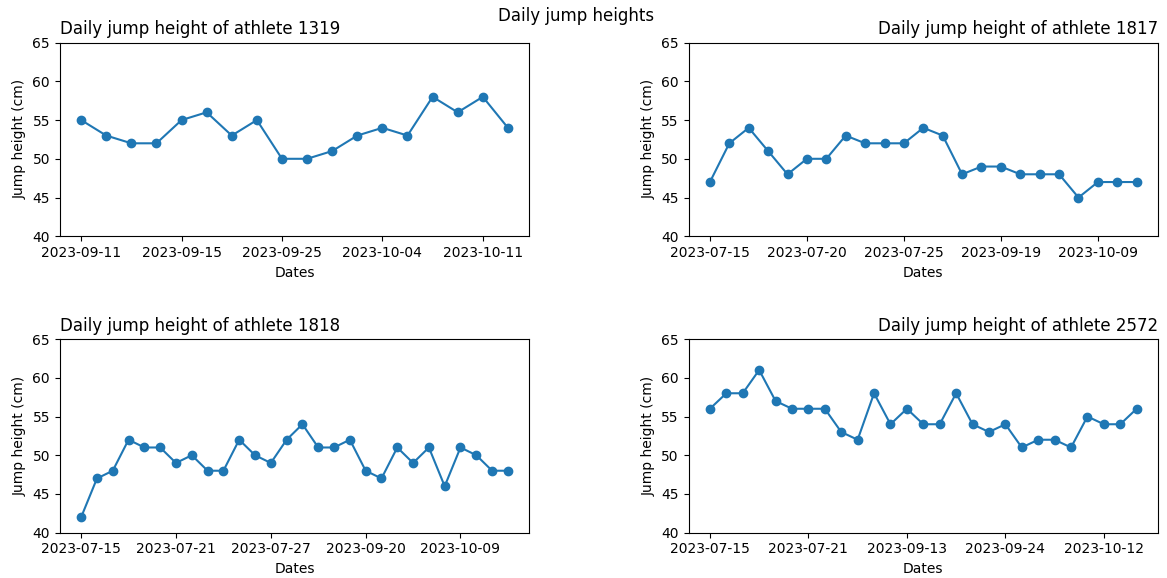


FIGURE 6: Jump height of four athletes, where athlete 1319 only participated in the second data collection protocol.

In Figure 6 the results of the daily jump test are plotted for four athletes. Due to an injury athlete 1319 only started his jump test in the second data collection protocol and not during the familiarisation period. Fluctuations in jump height can be found for all the athletes, but by visual inspection, no general trend during the training blocks can be found.

5.2 Importance of variables

In this subsection the results of the classification decision tree as considered in Section 4.1 will be presented. Therefore this section is based on the subresearch question:

Can fatigue be identified by monitoring daily jump height and by a wellness questionnaire?

Before looking at the results of the classification, the features will be ranked on importance first. The first results that will be discussed, are the ones of a standard k-fold cross-validation technique, followed by the more generalized leave-one-user-out cross-validation technique. Lastly, the approach where only samples with a comparable probability distribution are used as features is described. As a measure to analyse the performance of a decision tree method, accuracy is used. Accuracy is calculated as follows:

$$\text{Accuracy} = \frac{\text{Correct}}{\text{Correct} + \text{Incorrect}},$$

with correct the total amount of correct classified values and incorrect the amount of incorrectly classified values.

Before predicting the outcome of one of the four target variables, the decision tree model was used to rank the input features on importance in classifying the target variable. Where a higher importance means a higher contribution in predicting the target variable. For a random train set, which is 80% of the total set, a decision tree model is trained and the ten features with the highest importance are presented in Table 5 together with their importance. This is repeated for four different target variables; *Resting HR*, *Jump Height*, *Daily fatigue* and *Daily readiness*.

TABLE 5: Ranked importance of the input features for different target variables, where a higher importance means a higher contribution in classifying the target variable.

Rank	Target variable			
	<i>Resting HR</i>	<i>Jump Height</i>	<i>Daily fatigue</i>	<i>Daily Readiness</i>
1	daily sleep duration t-3 (0.052)	daily soreness (0.043)	resting hr (0.051)	daily sleep quality (0.103)
2	daily-fatigue (0.041)	daily soreness t-3 (0.035)	daily sleep quality (0.050)	daily readiness (0.101)
3	daily sleep duration (0.032)	daily sleep duration t-1 (0.033)	avg daily stress t5 (0.049)	resting hr t-2 (0.034)
4	daily sleep quality (0.031)	avg daily fatigue t3 (0.031)	daily sleep duration t-2 (0.045)	daily sleep duration t-3 (0.029)
5	daily readiness (0.031)	daily stress (0.027)	daily sleep quality t-2 (0.044)	avg resting hr t3 (0.027)
6	daily soreness t-1(0.024)	avg daily sleep duration t5 (0.026)	daily stress t-2 (0.044)	daily soreness t-1 (0.027)
7	daily fatigue t-3 (0.021)	avg daily fatigue t5 (0.020)	avg resting hr t5 (0.042)	daily sleep duration (0.024)
8	nr training (0.020)	resting hr t-3 (0.0195)	avg daily sleep duration t3 (0.042)	daily stress t-2 (0.023)
9	daily soreness (0.019)	daily fatigue t-2 (0.019)	avg daily sleep duration t5 (0.041)	avg nr training t5 (0.022)
10	avg nr training t3 (0.017)	daily sleep quality t-1 (0.019)	daily sleep duration (0.039)	daily stress t-1 (0.021)

Two variables that appear in all the four top ten lists are daily sleep duration and quality. With both the single value of the previous days as well as the average over the last three and/or five days. When looking at the four target variables, the feeling of fatigue influences all the other three variables. Readiness to train and resting HR are less commonly found in the top ten important input variables than the feeling of fatigue, but jump height never comes back as an important variable. Next to these four variables and next to the sleep duration and quality, soreness and stress do influence the results given in the wellness questionnaire.

5.2.1 k-fold

The results that can be found in Table 6 are from the classification decision tree model with k-fold cross-validation where k equals three. The accuracies are shown per fold and an average is shown.

TABLE 6: Accuracies of classification of the decision tree with the k-fold cross-validation model for different target variables.

Fold	Target variable			
	<i>Resting HR</i>	<i>Jump Height</i>	<i>Daily fatigue</i>	<i>Daily Readiness</i>
Fold 1	0.09	0.25	0.49	0.43
Fold 2	0.10	0.19	0.29	0.11
Fold 3	0.13	0.31	0.39	0.40
Average	0.11	0.25	0.39	0.31

5.2.2 Leave-one-user-out

The next classification decision tree model is trained on data of eleven athletes and as test set data of a new, unseen, athlete is used. The accuracies per athlete can be seen in Table 7 and at the end the average accuracy per target variable is given.

TABLE 7: Accuracies of classification of the decision tree with the leave-one-user-out cross-validation model for different target variables.

Athlete	Target variable			
	<i>Resting HR</i>	<i>Jump Height</i>	<i>Daily fatigue</i>	<i>Daily Readiness</i>
511	0.02	0.35	0.41	0.35
1319	0.02	0.29	0.37	0.27
1629	0.02	0.45	0.45	0.39
1817	0.08	0.29	0.35	0.39
1818	0.09	0.29	0.29	0.14
1819	0.02	0.25	0.55	0.47
1821	0.02	0.35	0.41	0.25
2572	0.02	0.33	0.49	0.47
2573	0.02	0.27	0.41	0.35
2575	0.12	0.18	0.53	0.61
2577	0.16	0.29	0.67	0.59
2583	0.02	0.29	0.59	0.45
Average	0.05	0.30	0.46	0.39

5.2.3 Train on variables with the same distribution

As a next step only features with the same probability distribution as variables in the test dataset are used, to see if this improves the accuracies of the classification. Still, the leave-one-user-out approach was used and accuracies can be found in Table 8.

TABLE 8: Accuracies of classification of the decision tree with extra preprocessing step to compare distributions of the samples for different target variables.

Athlete	Target variable			
	<i>Resting HR</i>	<i>Jump Height</i>	<i>Daily fatigue</i>	<i>Daily Readiness</i>
511	0.10	0.05	0.47	0.49
1319	0.08	0	0.21	0.22
1629	0.06	0	0.39	0.31
1817	0.04	0.04	0.27	0.37
1818	0.12	0.04	0.41	0.31
1819	0.08	0.15	0.63	0.47
1821	0	0.04	0.59	0.51
2572	0	0.07	0.49	0.45
2573	0.08	0.19	0.45	0.33
2575	0	0	0.18	0.27
2577	0.06	0.04	0.29	0.24
2583	0.04	0.04	0.43	0.20
Average	0.06	0.07	0.41	0.35

In Table 9, the comparison between the leave-one-user-out and train-on variables with the same distribution is given. Because the assumption was that training on variables with the same distribution would increase the accuracy in classifying, a positive value means an increase in performance and a negative value will correlate with a decrease in performance.

TABLE 9: Comparison of accuracies of the two different decision tree models where a positive value indicates that the last decision tree model has a better accuracy.

Athlete	Target variable			
	<i>Resting HR</i>	<i>Jump Height</i>	<i>Daily fatigue</i>	<i>Daily Readiness</i>
511	0.08	-0.30	0.06	0.14
1319	0.06	-0.29	-0.16	-0.05
1629	0.04	-0.45	-0.06	0.31
1817	-0.04	-0.25	-0.12	-0.02
1818	-0.03	-0.25	0.12	0.17
1819	0.06	-0.10	0.08	0
1821	-0.02	-0.31	0.18	0.26
2572	-0.02	-0.26	0.49	-0.02
2573	0.06	0.08	0.04	-0.02
2575	-0.12	-0.18	-0.35	-0.34
2577	-0.10	-0.25	-0.38	-0.35
2583	0.02	-0.25	-0.16	-0.25
Average	0.01	-0.23	-0.05	-0.04

In general, the accuracy of the classification of *daily Resting HR* is low in both models and in the comparison, the difference does alternate in sign. Also, the performance of the models in classifying the *daily jump height* is not high. Comparing both models on the variables *Daily fatigue* and *Daily Readiness* does give a slightly better result for the model where all the input variables are used as features.

5.3 Wingate performance to identify fatigue

In this subsection results related to the following sub-research question will be presented:
Can we use Wingate performance as a measure for fatigue in young elite speed skaters?

In this subsection, the dataset related to the three Wingate tests will be used, first for visual inspection, and followed by three statistical tests to determine if a statistical difference can be found between the results of the three tests.

5.3.1 Visualization

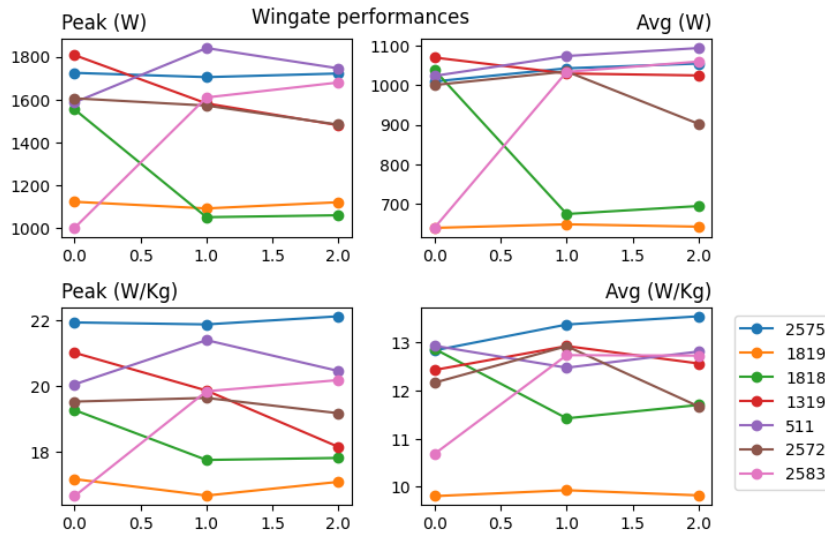


FIGURE 7: Visualization of the results of the three different Wingates for the athletes who performed three Wingates.

In Figure 7 the variables; peak and average power output and peak and average power output per kilogram body weight are plotted in four separate figures. Looking at these four subplots, no general trends between the three Wingate tests could be found. Some athletes seem to improve their power output in the third Wingate test. But other athletes do show a decrease in the second or third Wingate test.

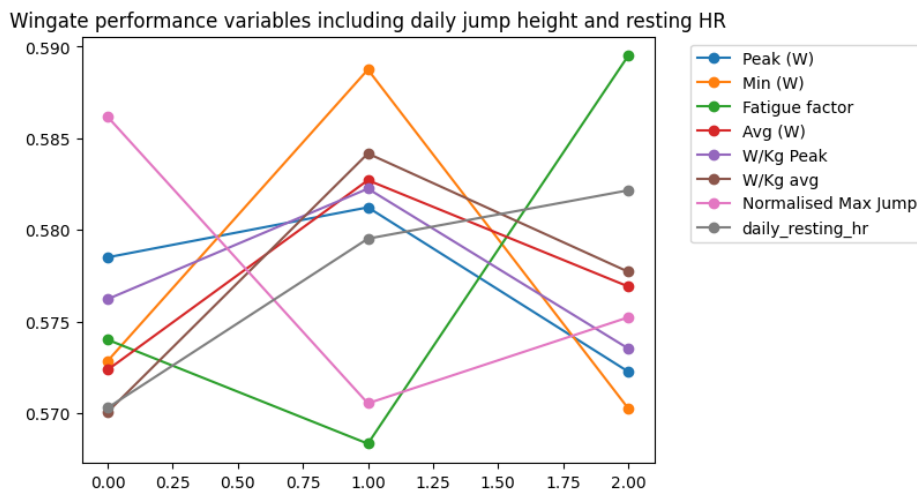


FIGURE 8: Normalised Wingate performance variables including a normalized result of the daily jump test and resting HR on the day of the performed Wingate test.

A trend can be found by visually inspecting the Wingate performance variables, jump height, and resting HR in Figure 8. It seems that there is an increase in performance on the second Wingate test, while there is a decrease in the height the athletes jumped on this day. When looking at the resting HR, the resting HR increases from test one to test two and from test two to test three. These observations stated above are confirmed when looking at the means in Table 10.

TABLE 10: Mean of the samples of the three different moments of the Wingate test.

Variable	Mean of the sample		
	Wingate 1	Wingate 2	Wingate 3
Peak (W)	1485.0	1492.0	1469.0
Average (W)	917.0	933.57	924.29
Peak (W/Kg)	19.37	19.57	19.28
Average (W/Kg)	11.96	12.26	12.12
Fatigue index	0.51	0.50	0.52
Jump height (cm)	53.57	52.14	52.57

5.3.2 Statistical tests

The results of the three different statistical tests performed on the dataset can be found in this subsection.

TABLE 11: Results of the statistical tests on the three different Wingate tests during the data collection protocol.

Variable	T-test			Wilcoxon signed rank			Kruskal Wallis
	p-value			p-value			p-value
	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	
Peak (W)	0.9664	0.9205	0.8855	0.8125	0.6875	0.5781	0.9890
Avg (W)	0.8719	0.9433	0.9270	0.5781	0.9375	0.5781	0.9863
Peak (W/Kg)	0.8439	0.9281	0.7665	1	0.9375	0.5781	0.9547
Avg (W/Kg)	0.6563	0.8097	0.8367	0.3750	0.8125	0.8125	0.9012
Fatigue index	0.8361	0.5917	0.3081	0.9375	0.5781	0.3750	0.6907
Jump height (cm)	0.6145	0.7470	0.8858	0.5781	0.8125	0.5781	0.8882

All the p-values in Table 11 are high, with the lowest one being 0.3750. With an α of 0.05, this means that none of the tests was able to reject the null hypothesis and thus no statistical difference between the three Wingate tests was found.

5.4 Resting HR prediction

The last sub-research question that will be treated in the results section is:

How can we predict resting HR data of young elite speed skaters?

As a performance measure for the LSTM networks, the Root Mean Square Error (RMSE) is used. RMSE is a commonly used metric to assess the quality of a predictive model and can be calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

with the following parameters:

- n the number of observations
- y_i the actual value
- \hat{y}_i the predicted value (Chai & Draxler, 2014)

5.4.1 Univariate Time Series

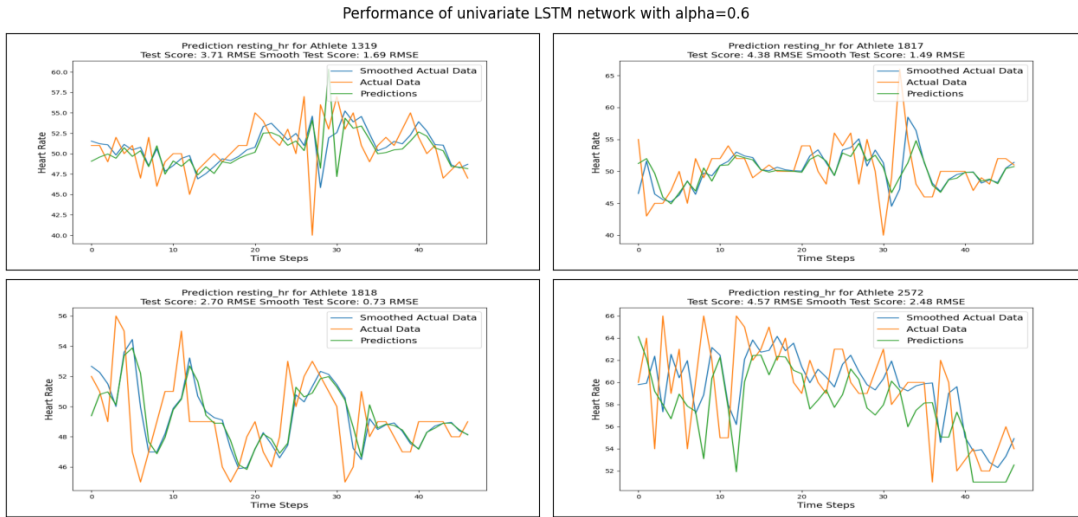


FIGURE 9: Prediction of the Resting HR of four athletes where the univariate LSTM model is trained on the Resting HR data of the other athletes.

The predictions of the daily resting HR of the four athletes can be seen in Figure 9 and the first thing to notice is that the green line of predictions does seem to follow the orange line of HR. Sometimes it seems the green line does follow the orange line exactly and it seems to copy the resting HR of yesterday. This will be further discussed in the Discussion, Section 6.3. In the raw HR dataset sometimes large deviations from the average are found, and the model is not able to capture those peaks. Comparing the green line of predictions with the blue line of the smoothed HR dataset the trend is even better captured.

5.4.2 Multivariate Time Series

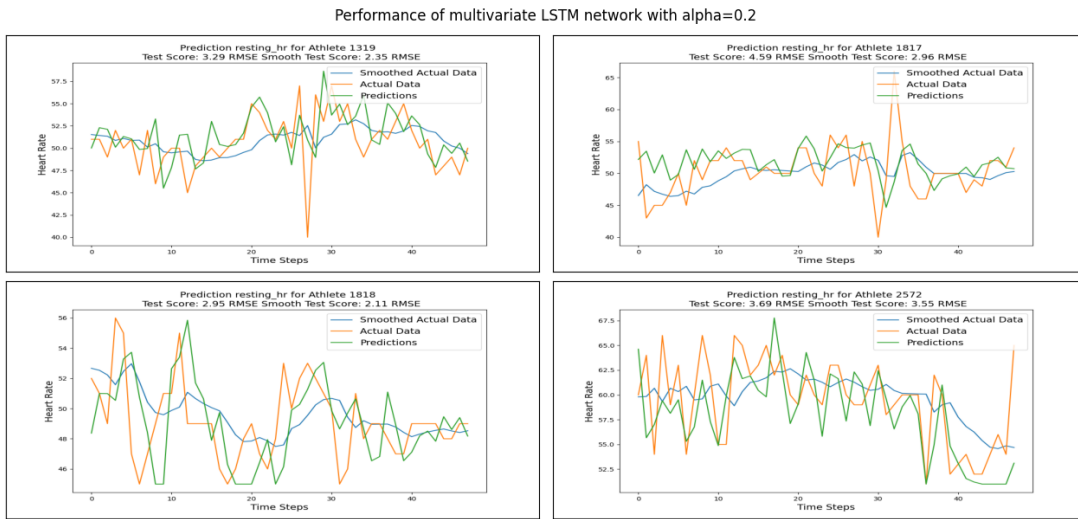


FIGURE 10: Prediction of the Resting HR of four athletes where the multivariate LSTM model is trained on the Resting HR data and features of the other athletes.

The same results can be seen for the predictions using the multivariate LSTM network. In general, the predictions do follow the pattern of the resting HR data points but outliers are not predicted. The green line still seems to follow the one-day shifted orange line, meaning the predicted resting HR of today is approximately the resting HR of yesterday. In the prediction of the HR of athlete 1818 a strange peak in HR can be found, which can not be seen in the original data.

5.4.3 Multivariate Time Series with top 10 relevant features

For the improved multivariate LSTM model, the LSTM model is trained on the following variables; daily sleep duration and quality, daily fatigue, daily readiness, daily soreness, and the number of training sessions yesterday. These variables were used as these are the variables that came back in the 5.2 section as the most important variables to predict resting HR. The results of this improved multivariate LSTM model can be found in Figure 11.

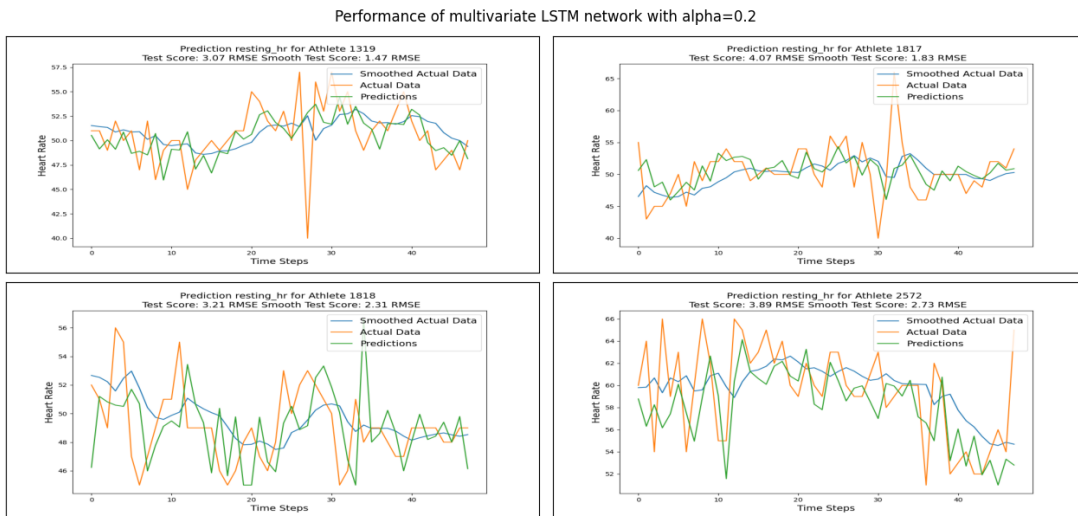


FIGURE 11: Prediction of the Resting HR of four athletes where the multivariate LSTM model is trained on the Resting HR data and relevant features of the other athletes.

This multivariate LSTM network does not seem to improve the prediction of the daily resting HR. On average the RMSE is the same as in the previous LSTM network, with randomly chosen input variables.

5.4.4 Comparison univariate vs multivariate LSTM

In Table 12 the RMSE per athlete can be found in the predictions with the univariate and multivariate models. For all three the models the RMSE per athlete is given for prediction on the raw dataset and on the smoothed dataset. For athletes 2577 and 2583 not enough HR data points were present to evaluate the prediction. Thus no predictions are done.

TABLE 12: Comparison of the univariate and two multivariate LSTM models.

Athlete	Root mean square error (RMSE)					
	Univariate LSTM	Smooth Univariate LSTM	Multivariate LSTM 1	Smooth 1 Multivariate LSTM	Multivariate LSTM 2	Smooth 2 Multivariate LSTM
511	3.52	1.04	3.14	2.01	2.94	2.50
1319	3.83	1.67	2.31	1.42	3.56	2.53
1629	5.22	1.69	4.27	1.96	4.97	2.52
1817	4.47	1.50	3.65	1.75	4.59	3.38
1818	2.47	0.67	3.12	2.10	3.23	2.42
1819	6.42	1.28	6.25	1.83	6.10	1.80
1821	2.57	1.12	4.29	3.15	3.80	3.26
2572	4.52	2.74	3.94	3.55	3.42	3.59
2573	4.23	1.38	4.62	3.37	4.39	2.87
2575	3.60	0.82	3.67	1.26	3.58	1.38
Average	4.09	1.51	3.93	2.40	4.06	2.63

When looking at these results, no increase in performance can be found for the predictions made on the raw HR data points. However, the predictions on the smoothed resting HR data points do differ between the univariate and multivariate LSTM models. Especially for athlete 1819 the RMSE for the raw dataset is really high. The individual predictions of all athletes for the three different models can be found in Appendix B.

5.4.5 Exponential smoothing

Different values of smoothing factor α are tested and the average RMSE for the three models described above is calculated per α . The results are plotted in Figure 12 below.

The performance of the two multivariate models decreases a lot if α becomes larger, while the result of the univariate model has a small decrease between $\alpha = 0.2$ and $\alpha = 0.6$ and an increase between $\alpha = 0.6$ and $\alpha = 0.8$. In general, the average RMSE of the univariate model is better than those of the multivariate models. For the models described above results are given for the α for which the average RMSE is the lowest, so $\alpha = 0.2$ for the multivariate models and $\alpha = 0.6$ for the univariate model.

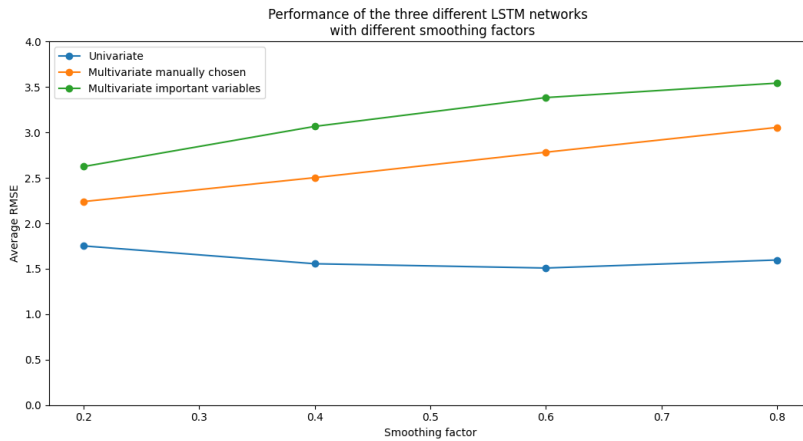


FIGURE 12: Average RMSE values of the three different LSTM models and different α values.

5.4.6 Excluding yesterday's resting HR as input

It seems as if the LSTM models attach great weight to yesterday's resting HR as an input variable in predicting today's resting HR. To reduce this behavior the univariate LSTM model is now trained on a slightly different look-back variable. Still, five days are used, but yesterday's resting HR is excluded from the input data. The model knows the resting HR data of the days $t - 6$ till $t - 2$ to predict the resting HR of day $t = 0$. Hence, $t - 1$ is excluded as input data to prevent the model from just copying the resting HR of yesterday as the predicted value of today.

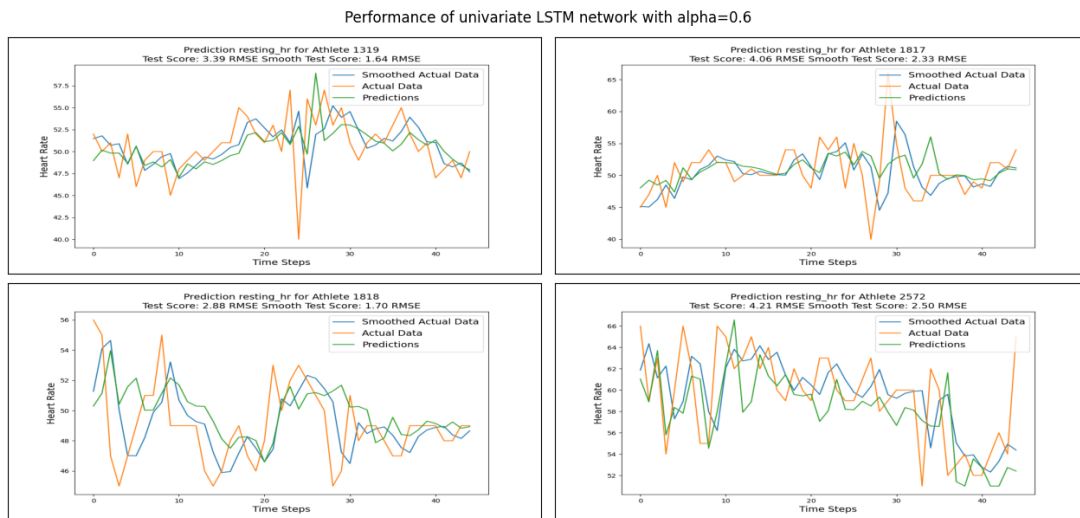


FIGURE 13: Prediction of the Resting HR of four athletes where the univariate LSTM model is trained on the Resting HR data of the other athletes and the model does not know yesterday's resting HR.

The results can be found in Figure 13. The performance metric, the RMSE, is approximately the same as in Figure 9 for the four athletes. The copying behavior seems to be still present, however shifted in days. This will be further discussed in the next section, the Discussion.

6 Discussion

In this section, the results given in Section 5 will be analyzed and discussed. The results will be interpreted in the same structure as the method and results section, starting with a subsection about the variables, followed by the Wingate subsection, and ending with heart rate predictions. At the end of each subsection, after the discussion of the results, also some recommendations for further research are indicated.

6.1 Importance of variables

In the literature study, the most important variables to describe training load according to previous studies are noted. Daily jump height is one of the variables used in previous research to monitor and predict neuromuscular fatigue. In this study, a daily countermovement jump test was performed in the morning, where jump height was measured by a vertec measuring tool. Unfortunately, this jump test was not performed daily by the athletes participating in this study. The jump and reach method was chosen because of its high reproducibility and because the test was easy to perform by the athletes without supervision. But because it was to be performed without supervision, the athletes forgot to jump in the morning. Hence, a considerable number of data points were absent during the data collection process, leading to limited insights from the analyses conducted on this dataset. A recommendation for other research is to have a data collection protocol where supervision is present when working with young athletes. A classifying decision tree model is run on the *daily jump height*, and three other variables; *resting HR*, *daily fatigue*, and *daily readiness*. From this model, it was found that the most important input variables are:

- readiness to train
- stress
- fatigue
- sleep duration
- soreness
- sleep quality

All these variables are collected in the daily morning questionnaire and from the variables in this questionnaire, only the daily mood is not present in the list of important input variables. The importance of these variables is low, the highest importance is 0.103. Meaning that no single variable has a dominant influence on the prediction of the target variable. Therefore no definite conclusions can be made of these results. However, it is good to observe that many of the variables currently used in the daily questionnaire at TalentNED are high in the ranked importance list. Campbell et al. (2021) found that wellness questionnaires are not good at predicting training load variables. The results of the classifying decision tree in this study are in line with that conclusion. The accuracies for classifying *resting HR* and *jump height* are lower than 10%. The model is better at classifying *daily fatigue* and *daily readiness* as those accuracies are higher. These variables come from the same wellness questionnaire as a lot of input variables for this model, so higher values for the importance are also expected. This can also explain why the model is better at classifying *daily fatigue* and *daily readiness*.

The folds of the k-fold cross-validation are made in chronological order. So the first one-third of the days are in fold 1 and the last one-third of the days are in fold 3. The results of the different folds do vary a lot in their accuracy of predicting, this distribution of the folds could explain this. If one fold consists of a specific sort of training in the periodization and the other two on another type, it can explain why the testing on the first fold is worse once the model is trained on the other two folds. Maybe this is happening by comparing fold 2 with fold 1 and 3 of the results, the average accuracy in classifying *daily fatigue* and *daily readiness* is worse than in fold 1 and 3. This way of making the folds would be interesting to use for analysis if the different types of training weeks match the number of folds made and this should be kept in mind when interpreting the results. There are more interesting distributions in how the folds are made for the k-fold cross-validation method. It could be argued that randomizing the folds would give an average better accuracy, which improves the model but has less practical applications. Another interesting option to take into account for further research is to make seven folds, where fold 1 consists of the data on Mondays, fold 2 on Tuesdays, etc.

The model which uses the leave one user out approach is a more generalized model than only testing on the data of one athlete. This more generalized approach shows higher accuracies in predicting the target variables, so the results are getting better. However, the results are still lower than 50%. In total, there are nine answer options in the questionnaire, but none of the athletes use the whole range of answer options. In the entire dataset, no athlete responded with a score lower than 2, and also not all athletes ever rated a question in their questionnaire with a 5. This influences the accuracies in classifying the target variables, as now only seven answer options are left in the dataset. Therefore random guessing for these classifications is $\frac{1}{7} = 14\%$, and at least both the 3-fold cross-validation as the leave one user out perform better than random guessing.

With the leave one user out method, it is possible that in the training set the answer option 5 is present, but in the test set the athlete never answered with a 5. Another issue in grouping the answers of all athletes in one dataset is that it is not known if a 4 answer on the feeling of fatigue for the athletes in the training set has the same meaning as a 4 on the feeling of fatigue for the athlete in the test set. This will influence the accuracy in classifying but also has an influence on the practical conclusions taken from this classification. It is also possible that some athletes are worse at self-reflection in those wellness questionnaires and are therefore less consistent in their answer options. Other possibilities are that they just always fill in an average or socially accepted answer. This is something that always can happen when using daily wellness questionnaires and looking for subjective answers but is important to take into account when interpreting these questionnaires.

In this study, there is only looked at the leave-one-user-out approach where the total dataset consists of all athletes. For further research, it would also be interesting to perform leave-one-user-out approaches on different groups of athletes. For example on a dataset of only female athletes, male athletes, all-round or sprint athletes or combinations of those characteristics. These analyses would be beneficial for the practical applications of the coaches and staff of the sports team. In this study, this is not done, because the dataset didn't consist of enough athletes to make this distinction between different groups of athletes.

In general, based on this sub-research question it would be interesting in future research to look at:

- Another range of answer options, for example, one to ten, such that, hopefully, more fluctuations in the answers are present.
- The interpretation of answers of different athletes/subjects. Looking at the answers of athletes over a longer time period to get a better idea of what someone's interpretation of an answer is.
- How good is a subject in reflecting on their wellness in a subjective questionnaire?

A first step would be to better instruct the subjects that they use the whole range of answer options and that the questionnaire is not meant for a socially accepted, quick answer. Next to this, an option could be to normalize the answers to the questionnaire between zero and one before analyzing the data.

6.2 Wingate performance to identify fatigue

The next objective in this study was related to the three Wingate tests that were taken during the data collection protocol. The dataset used in this analysis only consists of the period of the data collection in September and October. The familiarisation period was added to the total dataset to have more data for the analyses on daily jump height. Unfortunately, no Wingates are performed before or after this familiarisation period and thus this dataset is not useful for the analysis regarding the Wingate tests. Hofman et al. (2017) found that the Wingate test is a good performance indicator for speed skaters on the 1500m, so according to their research, this test can be seen as an ice-specific test. The 1500m in speed ice skating can be seen as a fatiguing event, and because the Wingate test is used as a predictor for the performance on the 1500m it is hypothesized that the Wingate test can also be used as an indicator for fatigue. As the Wingate test is a test on

a Wattbike, it only can measure fatigue in the lower extremities. For this research, it is not a problem, as ice skating is a sport where mainly the legs are used, but it can give a discrepancy in the results.

The results of the performance on the Wingate test of the athletes are compared with the training schedule and therefore intended state of fatigue of the coaches. If the power output obtained during a Wingate test is an indicator of fatigue it is expected that on the second Wingate, the power output is the highest, as fatigue is less. Consequently, when also looking at the jump height on the same day as the Wingate, it is expected that the jump height is also higher, as fatigue is less.

By visual inspection, there seems to be an improvement in performance in the second Wingate test. The peak and average power output are higher than in tests one and three and the fatigue index is lower. If looking at the graph and training scheme combined, the first interpretation is that the Wingate test can detect fatigue in young elite speed skaters. However, when inspecting the absolute values, only a small increase, between 10 and 30 Watt, in peak power output can be seen. According to the coaches, this increase is negligible. For example, a change in air friction and humidity could already lead to a change of 25 Watt in power output, when cycling at 1000 Watt (Wainwright et al., 2017). Next to this error in power output, also the time of the day at which the Wingate took place can influence the results. Souissi et al. (2007) found that the performance on Wingate tests in the afternoon is better than the performance of the same group performed in the morning. The order in which the athletes took their Wingate was different each time, resulting in the possibility that one athlete took their first Wingate at 9 in the morning and their third Wingate at 12, or the other way around. This can thus also influence the results of the performance on the Wingate tests. All athletes did their second Wingate later in the morning/beginning of the afternoon, which thus could lead to a better performance of this Wingate, without concluding anything about the fatigue state of the athletes. These findings of no significant difference between the results of the three Wingate tests are supported by the statistical tests performed on this data. No significant results were found, meaning that it can't be said that there is a difference between the means of the three groups.

Due to time limitations, in this study, only single features are used as variables for the different statistical tests. While on these single tests, no statistical significance could be found, it is possible that when testing on the combination of multiple features, there would be statistical significance between the three groups. The next step, for further research, would be, to also test on the combination of certain features. For example, the average and peak power output combined. A paired t-test could be used for these analyses, and different combinations of features should be tested.

In the visualization of the three Wingate tests in Figure 8 also the results of the daily jump height were shown. By comparing the results of the three jump tests, the athletes seem to jump less high on the day of the second Wingate test. Using the same assumption that the athletes should be less fatigued on the day of the second Wingate test compared with the first and third, contradicts the expected results. These results contradict the findings of Gupta et al. (2023), Pupo et al. (2021), Gavanda et al. (2023) and Coutts et al. (2007), who all found that a jump test could be used to identify fatigue. However, they all used different variants of jump tests, and no one used the reach and height jump test as used in this research. There are some limitations with this jump test, which could explain the contradicting results. First of all, the athletes were personally responsible for jumping and writing down the results of their tests. Because the test was unsupervised the athletes did not perform the test at the same moment in the morning. The agreement with the athletes was that the test was performed around breakfast time, but some athletes sometimes performed the test an hour after breakfast. Because the amount of jumps was already really low it was decided to use all the jump tests that were performed before noon. But just like the Wingate, the height jumped by a counter movement jump is higher in the afternoon than in the morning (Heishman et al., 2017). Also on a group level, the time at which the jump test was performed was not fixed. If on the schedule ice training was planned, the athletes jumped around seven in the morning, while on a rest day on average, the athletes jumped at 10:00 AM.

Both results contradict the expectation that both power output and daily jump height are higher in the second Wingate test, compared to the first and third Wingate test. Next to the limitations of the used protocols as mentioned above, there is another discussion point. The assumption of the coaches, that the athletes should be less fatigued on the day of the second Wingate could be wrong. If this is the case and the state of fatigue during the three days is approximately the same, it is expected that the tests will not show a significant difference. However, it is not expected that the assumption of the coaches about the state of fatigue of the athletes is wrong, as the coaches of the speed skating team are professional coaches.

For further research towards using the Wingate test as a measure for fatigue, it is important to make sure the subjects perform the Wingate at the same time on different days. According to Hofman et al. (2017), the Wingate test is a 1500m specific test for speed skaters, but in other research, this isn't confirmed nor contradicted. Therefore in further research, a more speed skaters-specific test should be developed and the validity of the Wingate test for speed skaters could be tested. To be able to say something about this form of the jump test, further research should be done on this format where the athletes jump on a fixed time and with supervision, such that there are enough data points to analyze.

6.3 Resting HR prediction

In this last subsection, the results of the models used for predicting daily resting HR are discussed. The univariate LSTM method was already able to capture the trend but missed the peaks. The average RMSE of the exponential smoothed model is around 1.5 beats per minute, which is decent. Without smoothing the average RMSE is a lot higher, so the exponential smoothing makes the predictions more accurate.

Two multivariate LSTM models were tested on the dataset, one where input variables are chosen manually and one where the input variables come from the decision tree model. Both models had comparable results in predicting without smoothing the data points, but the performance decreases with the exponential smoothing of the data in comparison with the univariate model. The input of extra variables does not increase the performance of the prediction of daily resting HR data. From this, we can conclude that the most important input variable in predicting resting HR is previous resting HR samples. And that the input variables are not that important in predicting resting HR. This is in line with the results of the decision tree as discussed in Section 5.2 where the importance of the variables is below the 0.052.

The performance of the models was better if tested on the smoothed data. With smoothing the data, outliers, which are possible measurement errors, do have less influence on the performance of the model. This of course benefits the performance. Yet, if the outliers signify potential non-functional overreaching or illness rather than measurement errors, it poses a risk to smooth the data, potentially erasing valuable insights encoded in these outliers.

As already mentioned in the Results section, the performance of the LSTM models could be debated. The model does seem to copy the last input and with a small modification predicts this value as the predicted heart rate of today. Additionally, when removing yesterday's resting heart rate from the input for the LSTM model, it still seems to copy patterns. However, now this copying happens with a value a few days before. If comparing the predictions from both models, it is noticed that the forecasts are shifted by a few days.

Copying is not a desired outcome of the LSTM network. LSTM networks are good models for capturing trends and seasonality over a longer period of a time series. However, the training set used in this study consists of, nine series of less than 60 input heart rates. No seasonality or trend may be present (yet) in this, relatively small, dataset. It is expected that seasonality would be present in daily resting HR, when measured over a longer period. During heavier training weeks daily resting HR is expected to be higher than during rest or recovery weeks. Multiple training blocks with the same structure of heavy and rest weeks are present in the year schedule and could show seasonality in daily resting HR. To verify in further research whether this assumption is correct, an LSTM model to predict daily resting HR with a longer data collection protocol should be tested. Oyeleye et al. (2022) did find that LSTM models with a recording duration of 1 min or shorter had a worse performance. This corresponds with the assumption that LSTM is not a good machine learning model for predicting daily resting HR with a small training set. According to Oyeleye et al. (2022) the ARIMA and linear regression methods showed promising results in predicting heart rates also for input data with a small duration. For further research, it could be interesting to check if an ARIMA or linear regression model could improve performance.

This research worked with daily resting HR data. In further research, it would be interesting to use HRV in combination or instead of resting HR. Ni et al. (2022) used different HRV variables to classify fatigue with a high accuracy. In this research, HRV is not used, due to not all the athletes had a smartwatch that was able to monitor HRV. But as HRV as input variable does give good results for classifying results it would be interesting to look at in another research. Another interesting direction for further research is to investigate if deviations of predicted vs actual resting HR data match certain moments of the athletes. For example, does a discrepancy happen between the actual resting HR and the predicted resting HR the days before an athlete was feeling ill?

7 Conclusion

In this section the conclusions of the results as given in Section 5 and 6 will be drawn and answers to the research questions will be given. First, the answers to the first two sub-research questions are repeated and the conclusions of the last three sub-research questions as stated in Section 1.2 will be treated one by one. Finally, an answer will be given to the main research question and an overall conclusion of this research will be given.

1) What are relevant variables to describe the training load of athletes?

According to the literature as described above, important variables to describe the training load of athletes can be divided into internal training load and external training load variables. A nice overview of these variables is shown in Figure 1.

2) How do the existing machine learning methods perform on the prediction of heart rates?

The current machine learning and deep learning methods that are already used for the prediction of heart rates do show some promising results. The best methods found in the literature are; ARIMA, linear regression, and a multivariate LSTM model.

3) Can fatigue be identified by monitoring daily jump height and by a wellness questionnaire?

During this research, the daily jump test was not executed correctly to use the daily jump height to measure or predict fatigue. There should have been more jumps and more at the same time moment for viable results. The classifying decision tree method used on, among other variables, the answers on the wellness questionnaire doesn't give promising results. Therefore, in this study and with this group of subjects, the reach and height daily jump test is not a viable method to identify fatigue. Also, the daily morning questionnaire as is currently used within TalentNED is not capable of classifying fatigue with a decision tree model.

To conclude, fatigue can not be identified by monitoring daily jump height and the wellness questionnaire as used in this study.

With using classifying decision trees unfortunately the connection between daily jump height and fatigue could not be explained further. Another technique that might be interesting to try is the subgroup discovery as used by Knobbe et al. (2017) in their research towards speed skaters.

4) Can we use Wingate performance as a measure for fatigue in young elite speed skaters?

The data collection period for the Wingate tests of this study consisted of four weeks, with a Wingate test at the start, middle, and end of this period. The results visually showed some changes in mean between the three tests but these discrepancies could be explained by other factors than fatigue. On the days of the Wingate test also the daily jump test was performed, and these results differed from the Wingate test and the expectations. As the reliability of the jump test in this research was already shown to be low, no conclusions can be drawn from this data.

In conclusion, the Wingate test, executed as in this study, can not be used as a measure of fatigue in young elite speed skaters.

5) How can we predict resting HR data of young elite speed skaters?

Three LSTM models are compared in their results of predicting daily resting HR data. The model with the best performance, defined as the lowest average RMSE, is the univariate LSTM model, where the predictions were tested on exponential smoothed data points. The predictions had some fluctuations compared to the actual data, but the trend was captured by the model. However, a copying behavior seems to be present which makes the current results, despite a low RMSE, less reliable. Before this model can be used for predicting resting HR, the copying of yesterday's resting HR behavior should be further investigated. If the problem related to this behavior is solved, the univariate LSTM model has the potential to be used for predicting the resting HR data of young elite speed skaters.

The answers to the three sub-research questions above lead to an answer to the main research question:

How can we monitor training load and identify fatigue in young elite speed skaters?

As found in the literature review, resting HR is a viable method to monitor fatigue in athletes. A longer period of an elevated resting HR indicates fatigue in athletes. If resting HR can be predicted, fatigue can also be identified.

To finalize, the training load can be monitored by separate monitoring internal and external loads. The use of a univariate LSTM model has the potential to predict resting HR, but more research is needed for this. More research is also needed to test the hypothesis that a discrepancy in predicted vs measured resting HR can identify fatigue in young elite speed skaters.

References

- Anani, W. (2018). Recurrent Neural Network Architectures Toward Intrusion Detection. *Electronic Thesis and Dissertation Repository*. <https://ir.lib.uwo.ca/etd/5625>
- Antwi-Afari, M. F., Anwer, S., Umer, W., Mi, H.-Y., Yu, Y., Moon, S., & Hossain, M. U. (2023). Machine learning-based identification and classification of physical fatigue levels: A novel method based on a wearable insole device. *International Journal of Industrial Ergonomics*, *93*, 103404. <https://doi.org/10.1016/j.ergon.2022.103404>
- Bar-Or, O. (1987). The Wingate Anaerobic Test: An Update on Methodology, Reliability and Validity. *Sports Medicine*, *4*(6), 381–394. <https://doi.org/10.2165/00007256-198704060-00001>
- Behrens, M., Gube, M., Chaabene, H., Prieske, O., Zenon, A., Broscheid, K.-C., Schega, L., Husmann, F., & Weippert, M. (2023). Fatigue and Human Performance: An Updated Framework. *Sports Medicine*, *53*(1), 7–31. <https://doi.org/10.1007/s40279-022-01748-2>
- Bestwick-Stevenson, T., Toone, R., Neupert, E., Edwards, K., & Kluzek, S. (2022). Assessment of Fatigue and Recovery in Sport: Narrative Review. *International Journal of Sports Medicine*, *43*(14), 1151–1162. <https://doi.org/10.1055/a-1834-7177>
- Bishop, P. A., Jones, E., & Woods, A. K. (2008). Recovery From Training: A Brief Review: Brief Review. *Journal of Strength and Conditioning Research*, *22*(3), 1015–1024. <https://doi.org/10.1519/JSC.0b013e31816eb518>
- Budgett, R. (1998). Fatigue and underperformance in athletes: The overtraining syndrome. *British Journal of Sports Medicine*, *32*(2), 107–110. <https://doi.org/10.1136/bjism.32.2.107>
- Bustos, D., Cardoso, F., Rios, M., Vaz, M., Guedes, J., Torres Costa, J., Santos Baptista, J., & Fernandes, R. J. (2022). Machine Learning Approach to Model Physical Fatigue during Incremental Exercise among Firefighters. *Sensors*, *23*(1), 194. <https://doi.org/10.3390/s23010194>
- Campbell, P. G., Stewart, I. B., Sirotic, A. C., Drovandi, C., Foy, B. H., & Minett, G. M. (2021). Analysing the predictive capacity and dose-response of wellness in load monitoring. *Journal of Sports Sciences*, *39*(12), 1339–1347. <https://doi.org/10.1080/02640414.2020.1870303>
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Coutts, A. J., Slattery, K. M., & Wallace, L. K. (2007). Practical tests for monitoring performance, fatigue and recovery in triathletes. *Journal of Science and Medicine in Sport*, *10*(6), 372–381. <https://doi.org/10.1016/j.jsams.2007.02.007>
- Daanen, H. A., Lamberts, R. P., Kallen, V. L., Jin, A., & Van Meeteren, N. L. (2012). A Systematic Review on Heart-Rate Recovery to Monitor Changes in Training Status in Athletes. *International Journal of Sports Physiology and Performance*, *7*(3), 251–260. <https://doi.org/10.1123/ijsp.7.3.251>
- Daigle, A.-P., Bélanger, S., Brunelle, J.-F., & Lemoyne, J. (2022). Functional Performance Tests, On-Ice Testing and Game Performance in Elite Junior Ice Hockey Players. *Journal of Human Kinetics*, *83*, 245–256. <https://doi.org/10.2478/hukin-2022-000076>
- De Leeuw, A.-W., Heijboer, M., Verdonck, T., Knobbe, A., & Latré, S. (2023). Exploiting sensor data in professional road cycling: Personalized data-driven approach for frequent fitness monitoring. *Data Mining and Knowledge Discovery*, *37*(3), 1125–1153. <https://doi.org/10.1007/s10618-022-00905-5>
- Douglas M. McNair, L. F. D., Maurice Lorr. (n.d.). *EdITS manual for the Profile of Mood States (POMS)*. Educational; Industrial Testing Service, San Diego, 1992.
- Durango, A. M., & Refugio, C. N. (2018). An Empirical Study on Wilcoxon Signed Rank Test. <https://doi.org/10.13140/RG.2.2.13996.51840>
- Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research*, *9*. <https://doi.org/10.3389/fenrg.2021.652801>
- Gavanda, S., Von Andrian-Werburg, C., & Wiewelhove, T. (2023). Assessment of fatigue and recovery in elite cheerleaders prior to and during the ICU World Championships. *Frontiers in Sports and Active Living*, *5*, 1105510. <https://doi.org/10.3389/fspor.2023.1105510>

- Giotta, M., Trerotoli, P., Palmieri, V. O., Passerini, F., Portincasa, P., Dargenio, I., Mokhtari, J., Montagna, M. T., & De Vito, D. (2022). Application of a Decision Tree Model to Predict the Outcome of Non-Intensive Inpatients Hospitalized for COVID-19. *International Journal of Environmental Research and Public Health*, *19*(20), 13016. <https://doi.org/10.3390/ijerph192013016>
- González-Boto, R., Salguero, A., Tuero, C., Márquez, S., & Kellmann, M. (2008). Spanish adaptation and analysis by structural equation modeling of an instrument for monitoring overtraining: The recovery-stress questionnaire (RESTQ-Sport). *Social Behavior and Personality: an international journal*, *36*(5), 635–650. <https://doi.org/10.2224/sbp.2008.36.5.635>
- Gupta, S., Baron, J., Bieniec, A., Swinarew, A., & Stanula, A. (2023). Relationship between vertical jump tests and ice-skating performance in junior Polish ice hockey players. *Biology of Sport*, *40*(1), 225–232. <https://doi.org/10.5114/biolsport.2023.112972>
- Halson, S. L. (2014). Monitoring Training Load to Understand Fatigue in Athletes. *Sports Medicine*, *44*(S2), 139–147. <https://doi.org/10.1007/s40279-014-0253-z>
- Hamlin, M. J., Wilkes, D., Elliot, C. A., Lizamore, C. A., & Kathiravel, Y. (2019). Monitoring Training Loads and Perceived Stress in Young Elite University Athletes. *Frontiers in Physiology*, *10*, 34. <https://doi.org/10.3389/fphys.2019.00034>
- Heishman, A. D., Curtis, M. A., Saliba, E. N., Hornett, R. J., Malin, S. K., & Weltman, A. L. (2017). Comparing Performance During Morning vs. Afternoon Training Sessions in Intercollegiate Basketball Players. *Journal of Strength and Conditioning Research*, *31*(6), 1557–1562. <https://doi.org/10.1519/JSC.0000000000001882>
- Herbert, P., Sculthorpe, N., Baker, J. S., & Grace, F. M. (2015). Validation of a Six Second Cycle Test for the Determination of Peak Power Output. *Research in Sports Medicine*, *23*(2), 115–125. <https://doi.org/10.1080/15438627.2015.1005294>
- Hofman, N., Orié, J., Hoozemans, M. J., Foster, C., & De Koning, J. J. (2017). Wingate Test as a Strong Predictor of 1500-m Performance in Elite Speed Skaters. *International Journal of Sports Physiology and Performance*, *12*(10), 1288–1292. <https://doi.org/10.1123/ijssp.2016-0427>
- Inoue, A., Dos Santos Bunn, P., Do Carmo, E. C., Lattari, E., & Da Silva, E. B. (2022). Internal Training Load Perceived by Athletes and Planned by Coaches: A Systematic Review and Meta-Analysis. *Sports Medicine - Open*, *8*(1), 35. <https://doi.org/10.1186/s40798-022-00420-3>
- Jiménez Morgan, S., & Molina Mora, J. A. (2017). Effect of Heart Rate Variability Biofeedback on Sport Performance, a Systematic Review. *Applied Psychophysiology and Biofeedback*, *42*(3), 235–245. <https://doi.org/10.1007/s10484-017-9364-2>
- Kim, T. K. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, *68*(6), 540. <https://doi.org/10.4097/kjae.2015.68.6.540>
- Knobbe, A., Orié, J., Hofman, N., Van Der Burgh, B., & Cachucho, R. (2017). Sports analytics for professional speed skating. *Data Mining and Knowledge Discovery*, *31*(6), 1872–1902. <https://doi.org/10.1007/s10618-017-0512-3>
- Kreher, J. B., & Schwartz, J. B. (2012). Overtraining Syndrome: A Practical Guide. *Sports Health: A Multidisciplinary Approach*, *4*(2), 128–138. <https://doi.org/10.1177/1941738111434406>
- Krupp, L. B., & Pollina, D. A. (1996). Mechanisms and management of fatigue in progressive neurological disorders. *Current Opinion in Neurology*, *9*(6), 456–460. <https://doi.org/10.1097/00019052-199612000-00011>
- Lambert, M., & Borresen, J. (2006). A Theoretical Basis of Monitoring Fatigue: A Practical Approach for Coaches. *International Journal of Sports Science & Coaching*, *1*(4), 371–388. <https://doi.org/10.1260/174795406779367684>
- Luo, M., & Wu, K. (2020). Heart rate prediction model based on neural network. *IOP Conference Series: Materials Science and Engineering*, *715*(1), 012060. <https://doi.org/10.1088/1757-899X/715/1/012060>
- McArdle, W. D., Katch, F. I., & Katch, V. L. (2015). *Exercise physiology: Nutrition, energy, and human performance* (Eighth edition). Wolters Kluwer Health/Lippincott Williams & Wilkins.
- Meeusen, R., Duclos, M., Gleeson, M., Rietjens, G., Steinacker, J., & Urhausen, A. (2006). Prevention, diagnosis and treatment of the Overtraining Syndrome: ECSS Position Statement

- ‘Task Force’. *European Journal of Sport Science*, 6(1), 1–14. <https://doi.org/10.1080/17461390600617717>
- Mendes, B., Palao, J. M., Silvério, A., Owen, A., Carriço, S., Calvete, F., & Clemente, F. M. (2018). Daily and weekly training load and wellness status in preparatory, regular and congested weeks: A season-long study in elite volleyball players. *Research in Sports Medicine*, 26(4), 462–473. <https://doi.org/10.1080/15438627.2018.1492393>
- Nässi, A., Ferrauti, A., Meyer, T., Pfeiffer, M., & Kellmann, M. (2017). Psychological tools used for monitoring training responses of athletes. *Performance Enhancement & Health*, 5(4), 125–133. <https://doi.org/10.1016/j.peh.2017.05.001>
- Ni, Z., Sun, F., & Li, Y. (2022). Heart Rate Variability-Based Subjective Physical Fatigue Assessment. *Sensors*, 22(9), 3199. <https://doi.org/10.3390/s22093199>
- Otter, R. T. A., Bakker, A. C., Van Der Zwaard, S., Toering, T., Goudsmit, J. F. A., Stoter, I. K., & De Jong, J. (2022). Perceived Training of Junior Speed Skaters versus the Coach’s Intention: Does a Mismatch Relate to Perceived Stress and Recovery? *International Journal of Environmental Research and Public Health*, 19(18), 11221. <https://doi.org/10.3390/ijerph191811221>
- Oyeleye, M., Chen, T., Titarenko, S., & Antoniou, G. (2022). A Predictive Analysis of Heart Rates Using Machine Learning Techniques. *International Journal of Environmental Research and Public Health*, 19(4), 2417. <https://doi.org/10.3390/ijerph19042417>
- Pasic, M., Begic, E., Kadic, F., Gavrankapetanovic, A., & Pasic, M. (2022). Development of neural network models for prediction of the outcome of COVID-19 hospitalized patients based on initial laboratory findings, demographics, and comorbidities. *Journal of Family Medicine and Primary Care*, 11(8), 4488. https://doi.org/10.4103/jfmpe.jfmpe_113_22
- Pupo, J., Ache-Dias, J., Kons, R. L., & Detanico, D. (2021). Are vertical jump height and power output correlated to physical performance in different sports? An allometric approach. *Human Movement*, 22(2), 60–67. <https://doi.org/10.5114/hm.2021.100014>
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk , kolmogorov-smirnov , lilliefors and anderson-darling tests. <https://api.semanticscholar.org/CorpusID:18639594>
- Rushall, B. S. (1990). A tool for measuring stress tolerance in elite athletes. *Journal of Applied Sport Psychology*, 2(1), 51–66. <https://doi.org/10.1080/10413209008406420>
- Shan, F., He, X., Armaghani, D. J., & Sheng, D. (2023). Effects of data smoothing and recurrent neural network (RNN) algorithms for real-time forecasting of tunnel boring machine (TBM) performance. *Journal of Rock Mechanics and Geotechnical Engineering*, S1674775523002202. <https://doi.org/10.1016/j.jrmge.2023.06.015>
- Souissi, N., Bessot, N., Chamari, K., Gauthier, A., Sesboüé, B., & Davenne, D. (2007). Effect of Time of Day on Aerobic Contribution to the 30-s Wingate Test Performance. *Chronobiology International*, 24(4), 739–748. <https://doi.org/10.1080/07420520701535811>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56), 1929–1958. <http://jmlr.org/papers/v15/srivastava14a.html>
- Thorpe, R. T., Atkinson, G., Drust, B., & Gregson, W. (2017). Monitoring Fatigue Status in Elite Team-Sport Athletes: Implications for Practice. *International Journal of Sports Physiology and Performance*, 12(s2), S2–27–S2–34. <https://doi.org/10.1123/ijsp.2016-0434>
- Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.
- Wainwright, B., Cooke, C. B., & O’Hara, J. P. (2017). The validity and reliability of a sample of 10 Wattbike cycle ergometers. *Journal of Sports Sciences*, 35(14), 1451–1458. <https://doi.org/10.1080/02640414.2016.1215495>
- Wang, T. Y., Cui, J., & Fan, Y. (2023). A wearable-based sports health monitoring system using CNN and LSTM with self-attentions (S. Veerappampalayam Easwaramoorthy, Ed.). *PLOS ONE*, 18(10), e0292012. <https://doi.org/10.1371/journal.pone.0292012>
- Williams, N. (2017). The Borg Rating of Perceived Exertion (RPE) scale. *Occupational Medicine*, 67(5), 404–405. <https://doi.org/10.1093/occmed/kqx063>
- Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists* (First edition) [OCLC: ocn957747646]. O’Reilly.

8 Appendix A

8.1 Daily questionnaire

Welzijn

Slaapkwaliteit

1 1.5 2 2.5 3 3.5 4 4.5 5
Slapeloosheid Onrustige slaap Normaal Goed geslapen Erg goed geslapen

Slaapduur

0 0
Uren Minuten

Vermoeidheid

1 1.5 2 2.5 3 3.5 4 4.5 5
Heel vermoeid Meer vermoeid dan normaal Normaal Energiek Heel energiek

Stress

1 1.5 2 2.5 3 3.5 4 4.5 5
Erg gestresst Gestresst Normaal Relaxed Erg relaxed

Algemene spierpijn

1 1.5 2 2.5 3 3.5 4 4.5 5
Erg veel spierpijn Verhoogde spierspanning / spierpijn Normaal Voelt goed Voelt geweldig

Gemoedstoestand

1 1.5 2 2.5 3 3.5 4 4.5 5
Erg prikkelbaar / Kortaf tegen teamgenoten, familie en down collega's Normaal Een overwegend goede gemoedstoestand Een erg positieve gemoedstoestand

Gezondheid

Readiness-to-train

1 1.5 2 2.5 3 3.5 4 4.5 5
Onmogelijk te trainen Niet ready om te trainen Normaal Ready om te trainen Helemaal ready om te trainen

Rusthartslag (slagen per minuut)

Vul een getal in _____

Ziek?

ja
 nee

Wat is uw lichaamsgewicht op dit moment?

Vul een getal in (kg) _____

Geblesseerd?

ja
 nee

Overig

Opmerkingen

Vul iets in (optioneel) _____

FIGURE 14: Example of the daily questionnaire

8.2 Training log

The form is titled 'Training log' and contains several input fields and a scale. On the left, there are three dropdown menus: 'Trainingtype', 'Sessietype', and 'Tijdsduur (min)'. Below these is a text input field for 'Tevredenheid over training' and a section for 'Opmerkingen'. On the right, there is a 'Begonnen op' field with a date and time '19-09-2023 13:00'. Below this is an 'RPE score' vertical color scale ranging from 1 (Heel licht) to 10 (Maximaal). At the bottom, there is a horizontal scale from 1 to 5 with labels: 'Totaal ontevreden', '1.5', '2', '2.5', '3', '3.5', '4', '4.5', '5', 'Totaal tevreden'. The labels 'Ontevreden', 'Normaal', and 'Tevreden' are positioned under their respective values on the scale.

FIGURE 15: Example of the training log questionnaire

8.3 Evening questionnaire

The form is titled 'Vragenlijst belasting' and contains four questions. Question 1: 'Hoe mentaal belastend heb je vandaag ervaren? *' with a 10-point scale from 'Totaal niet belastend' to 'Extreem belastend'. Question 2: 'Hoe fysiek belastend heb je vandaag ervaren? *' with a 10-point scale from 'Totaal niet belastend' to 'Extreem belastend'. Question 3: 'Hoe heb je de totale belasting van vandaag ervaren? *' with a 10-point scale from 'Totaal niet belastend' to 'Extreem belastend'. Question 4: 'Waren er bijzonderheden op deze dag die je wilt delen?' with a text input field labeled 'Enter your answer' and an 'Add new' button.

FIGURE 16: Example of the evening questionnaire

8.4 Weekly questionnaire

Fysieke- & gezondheidsklachten vragenlijst

Heb je in de afgelopen 7 dagen hinder ondervonden bij het sporten ten gevolge van een blessure, ziekte of andere gezondheidsklachten?

- Nee, ik heb volledig deelgenomen zonder klachten
- Ja, ik heb hinder ondervonden door een **blessure / fysieke klachten**
- Ja, ik heb hinder ondervonden door een **ziekte / gezondheidsklachten**
- Ja, ik heb hinder ondervonden door **beide**, zowel een blessure / fysieke klachten als een ziekte / gezondheidsklachten

OPSLAAN

Vraag 1 - Deelname

Heb je enige moeite met deelname aan het sporten gehad door ziekte/gezondheidsklachten in de afgelopen 7 dagen?

- Volledige deelname, maar met klachten
- Verminderde deelname door klachten
- Kan niet deelnemen door klachten

Hoeveel dagen in de week heb je niet volledig kunnen deelnemen aan het sporten ten gevolge van je gezondheidsklachten?

- 0
- 1
- 2
- 3
- 4
- 5
- 6
- 7

Welke symptomen heb je de afgelopen 7 dagen ervaren?

- Angstig / Onrustig
- Buikpijn
- Diarree
- Flauwvallen
- Gevoel van slape / Vermoeidheid
- Hoesten
- Hoofdpijn
- Jek / Uitslag
- Koorts
- Kortademig / Benauwd
- Misselijkheid
- Onregelmatige poos / Palpaties
- Oogklachten
- Oorklachten
- Opperste gewelsten
- Overgeven
- Rijnvloed
- Rijn op de hant
- Pissebaer
- Slepekoehed
- Stijf / Gevoelloos
- Teneergelagen / Depressief
- Unwaaagelosten of problemen met de gelastheden
- Verstopping
- Verstope neus / Looppeus / Niesen
- Wil ik niet zeggen
- Zere taal
- Anders namelijk:

Vraag 1 - Deelname

Heb je enige moeite met deelname aan het sporten gehad door ziekte/gezondheidsklachten in de afgelopen 7 dagen?

- Volledige deelname, maar met klachten
- Verminderde deelname door klachten
- Kan niet deelnemen door klachten

Vraag 2 - Aangepaste training/competitie

In welke mate heb je het sporten aangepast door ziekte/gezondheidsklachten in de afgelopen 7 dagen?

- Niet aangepast
- In geringe mate aangepast
- In redelijke mate aangepast
- In grote mate aangepast

Vraag 3 - Prestatie

In welke mate heeft de ziekte/gezondheidsklachten een negatieve invloed gehad op je prestatie in de afgelopen 7 dagen?

- Geen invloed
- In geringe mate beïnvloed
- In redelijke mate beïnvloed
- In grote mate beïnvloed

Vraag 4 - Symptomen

In welke mate heb je symptomen of gezondheidsklachten ervaren in de afgelopen 7 dagen?

- Geen symptomen of gezondheidsklachten
- Geringe symptomen of gezondheidsklachten
- Redelijke symptomen of gezondheidsklachten
- Ernstige symptomen of gezondheidsklachten

Ben je voor je gezondheidsklachten de afgelopen 7 dagen behandeld door een therapeut of arts?

- Ja
- Nee

Hoeveel uur heb je de afgelopen week gesport?

Vul een getal in

OPSLAAN

FIGURE 17: Example of the weekly questionnaire

9 Appendix B

9.1 Univariate LSTM network

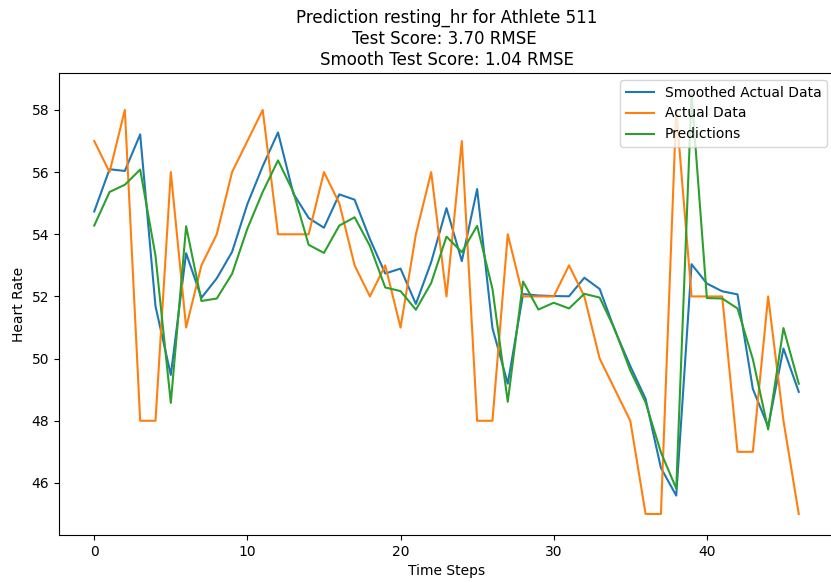


FIGURE 18: Prediction of resting HR for athlete 511

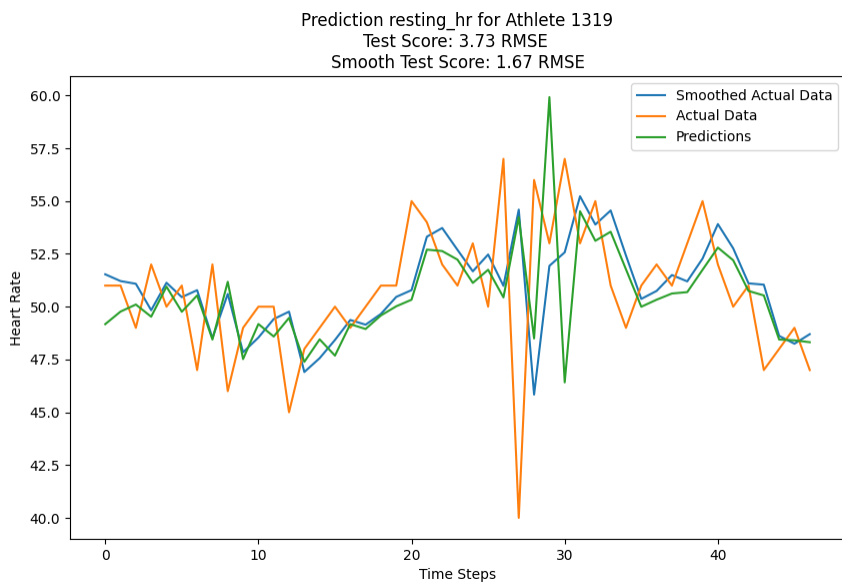


FIGURE 19: Prediction of resting HR for athlete 1319

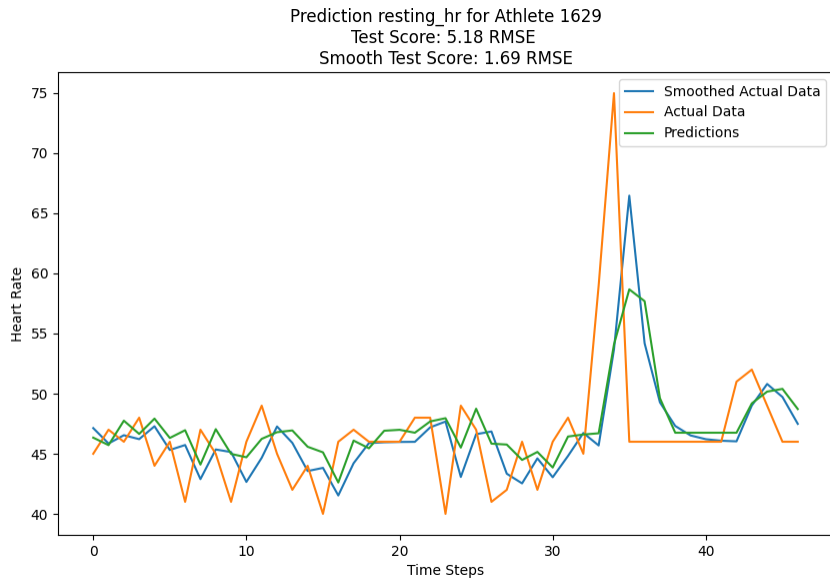


FIGURE 20: Prediction of resting HR for athlete 1629

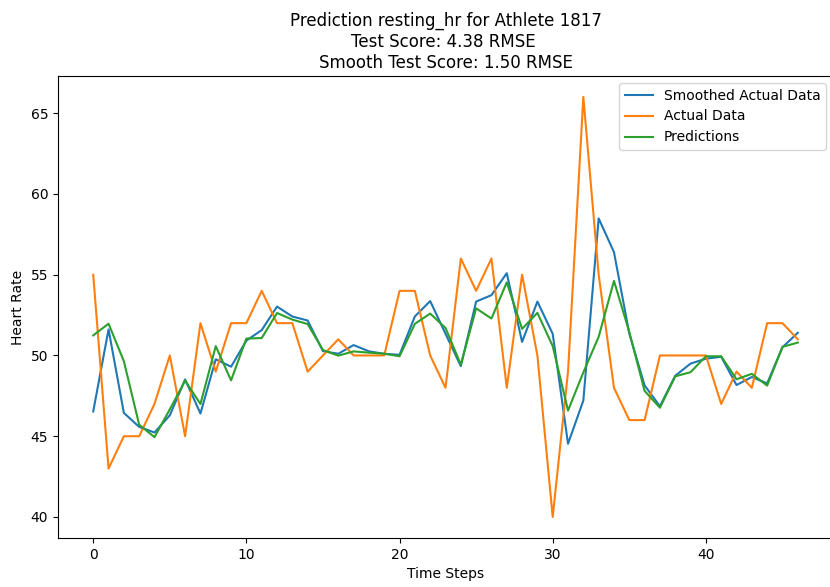


FIGURE 21: Prediction of resting HR for athlete 1817

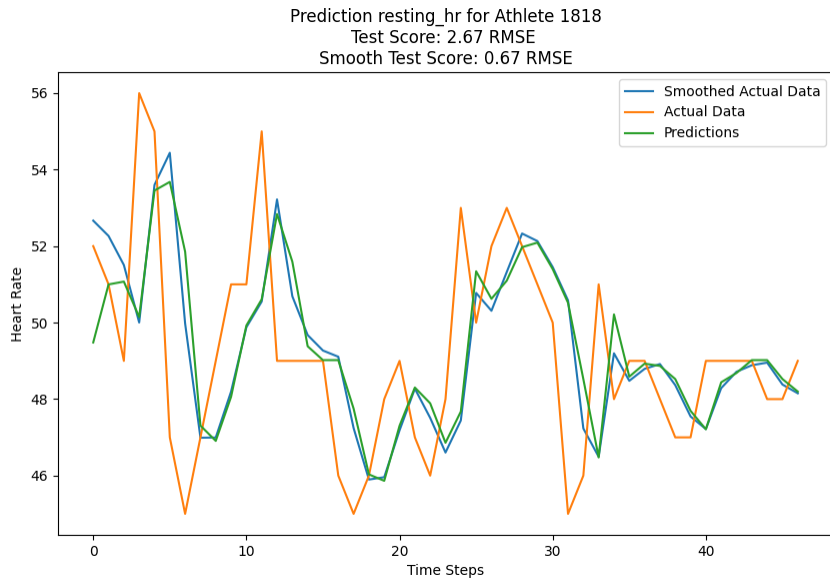


FIGURE 22: Prediction of resting HR for athlete 1818

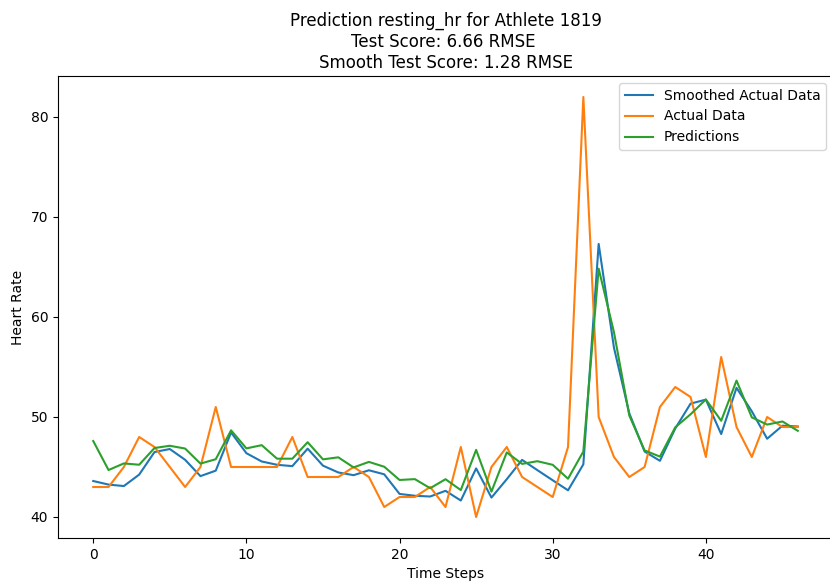


FIGURE 23: Prediction of resting HR for athlete 1819

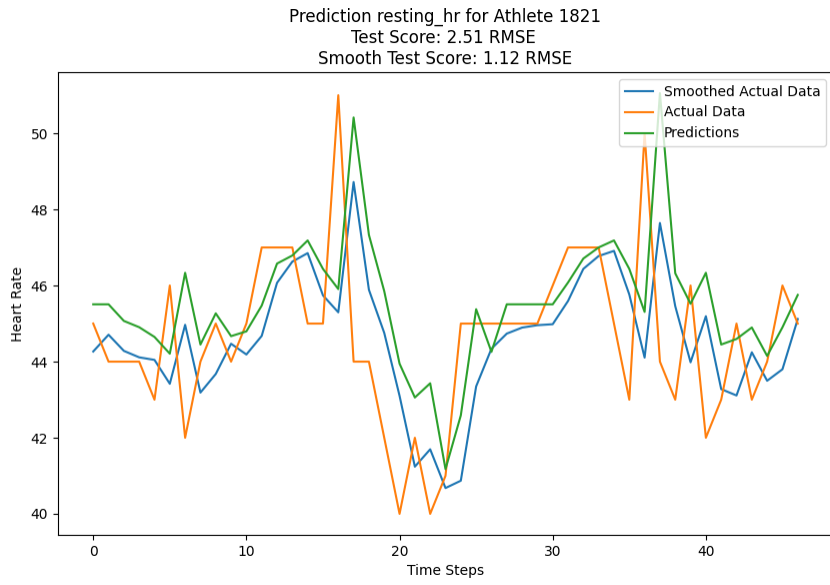


FIGURE 24: Prediction of resting HR for athlete 1821

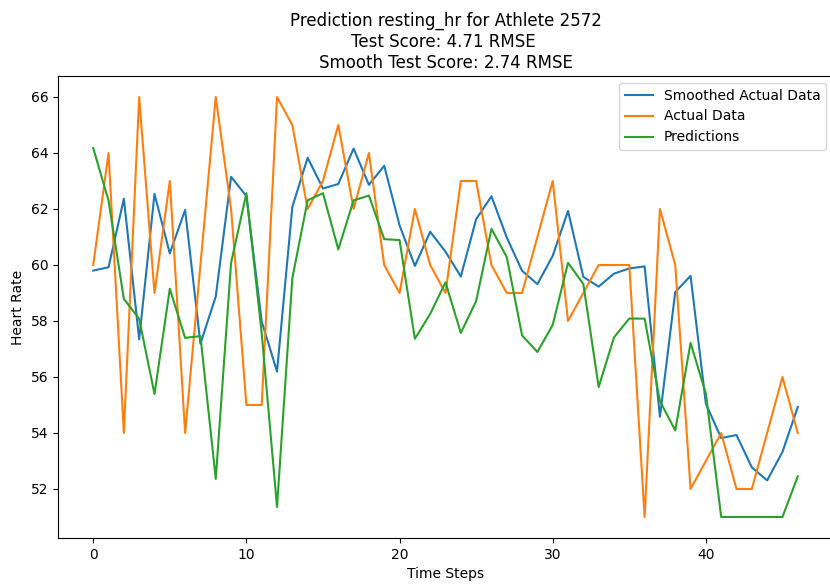


FIGURE 25: Prediction of resting HR for athlete 2572

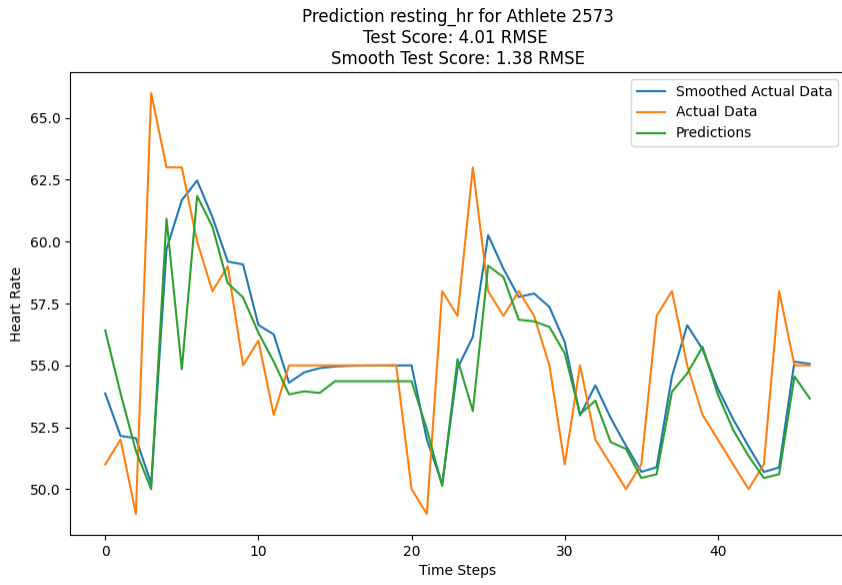


FIGURE 26: Prediction of resting HR for athlete 2573

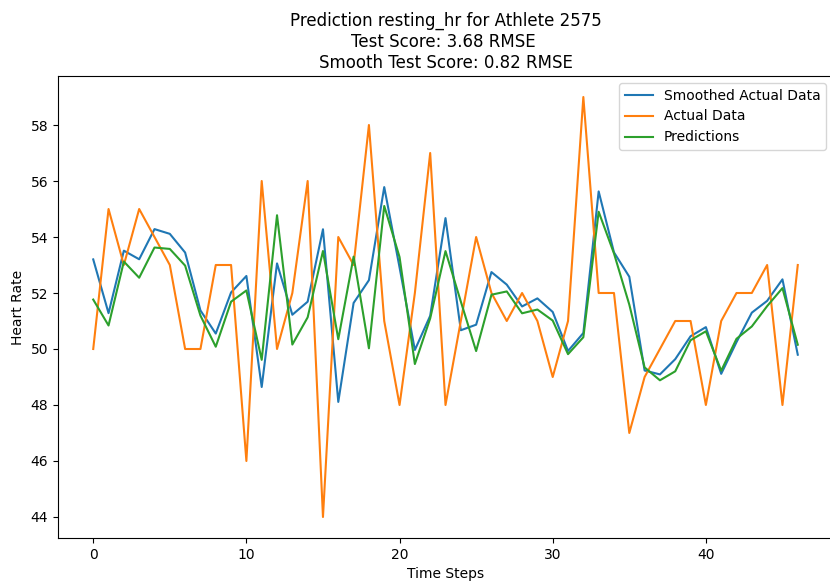


FIGURE 27: Prediction of resting HR for athlete 2575

9.2 Multivariate LSTM network 1

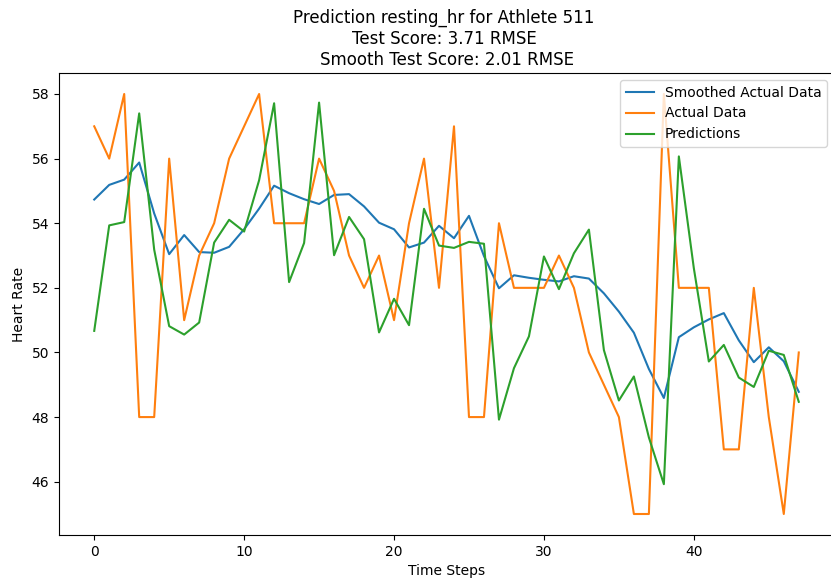


FIGURE 28: Prediction of resting HR for athlete 511

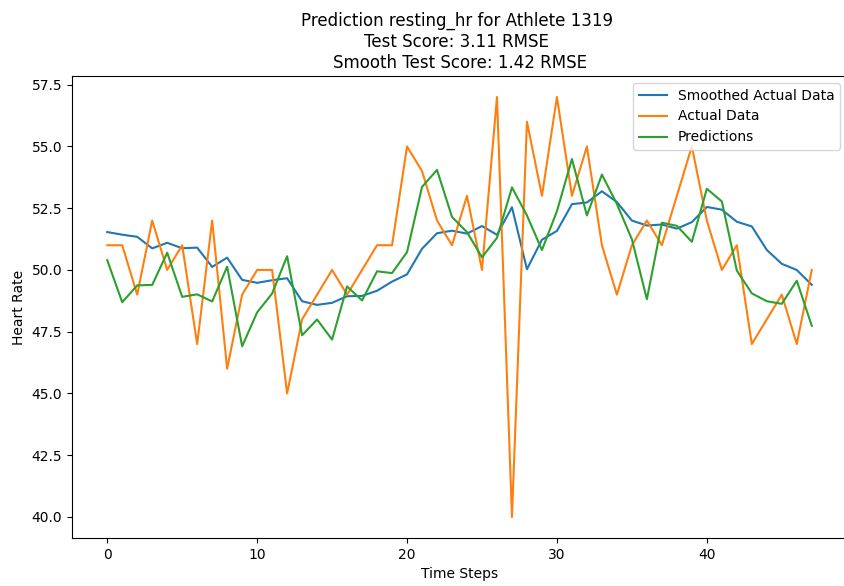


FIGURE 29: Prediction of resting HR for athlete 1319

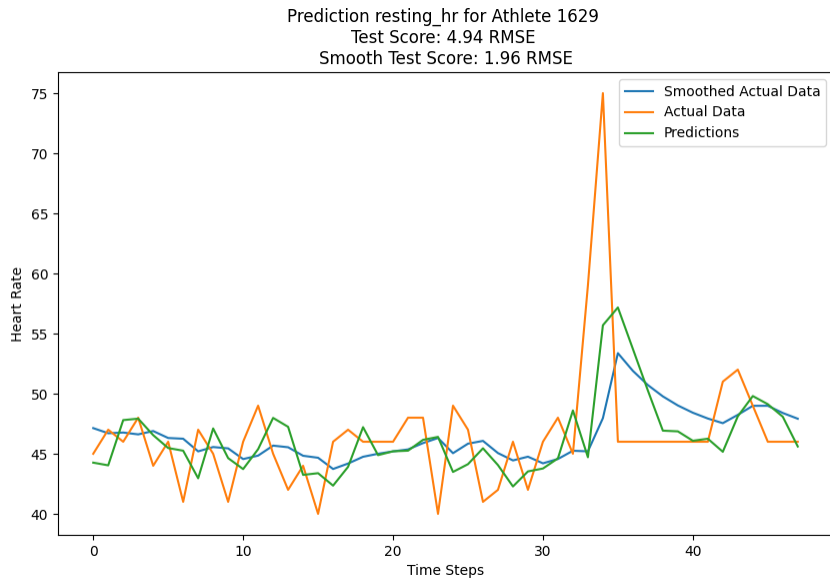


FIGURE 30: Prediction of resting HR for athlete 1629

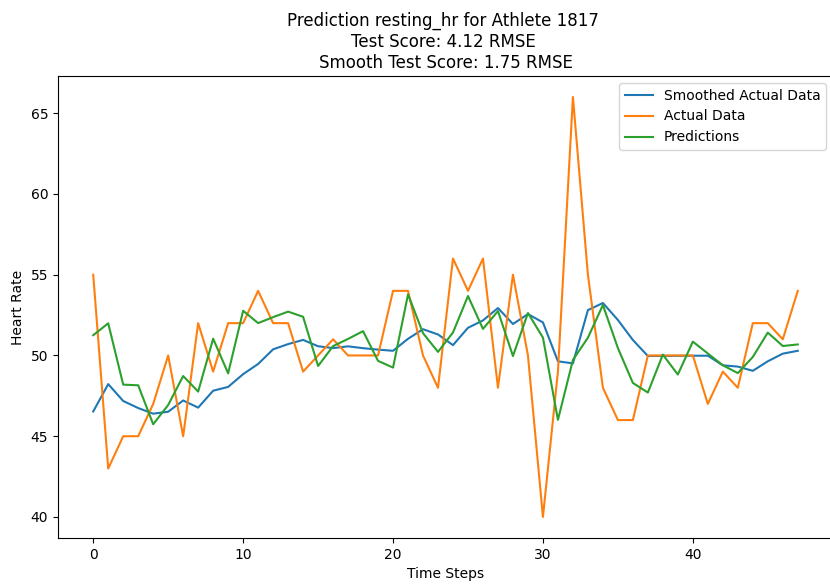


FIGURE 31: Prediction of resting HR for athlete 1817

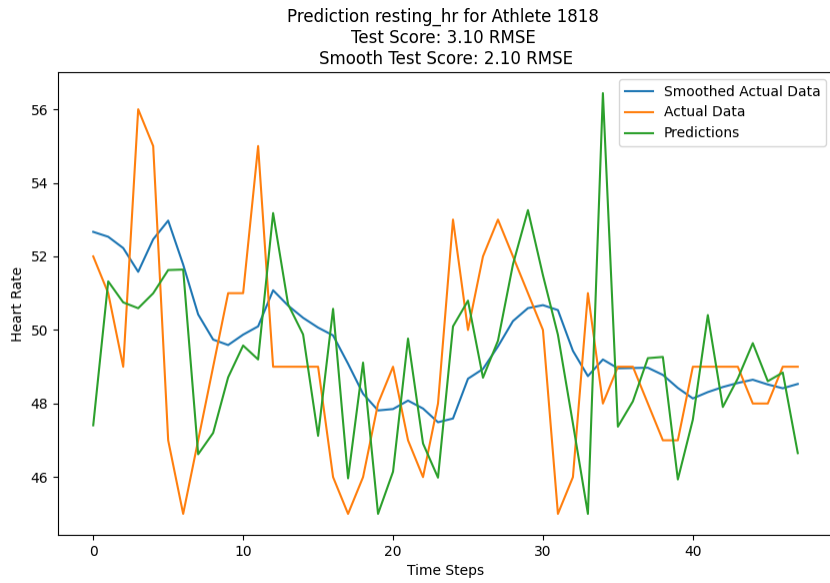


FIGURE 32: Prediction of resting HR for athlete 1818

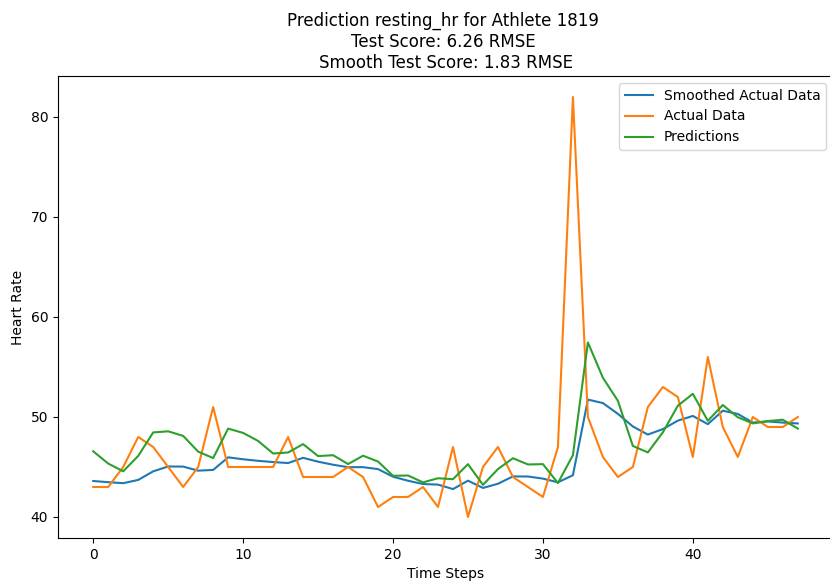


FIGURE 33: Prediction of resting HR for athlete 1819

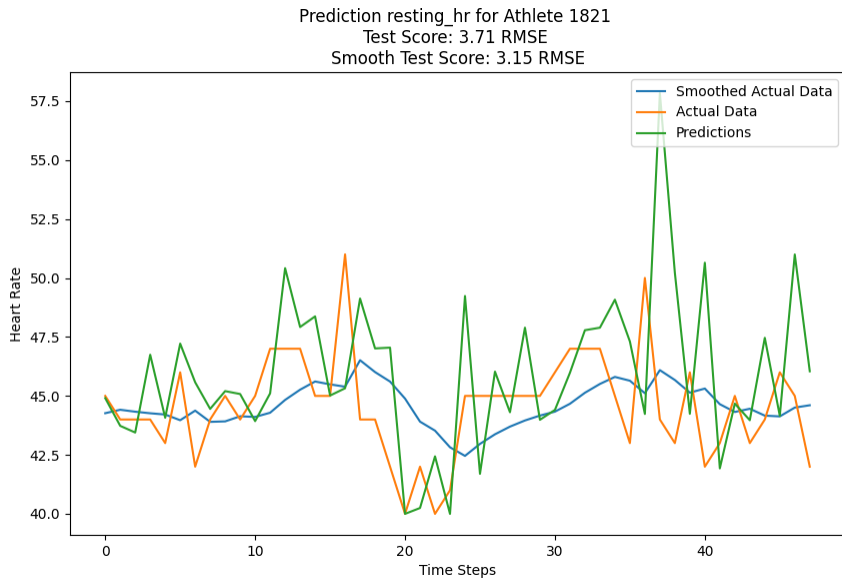


FIGURE 34: Prediction of resting HR for athlete 1821

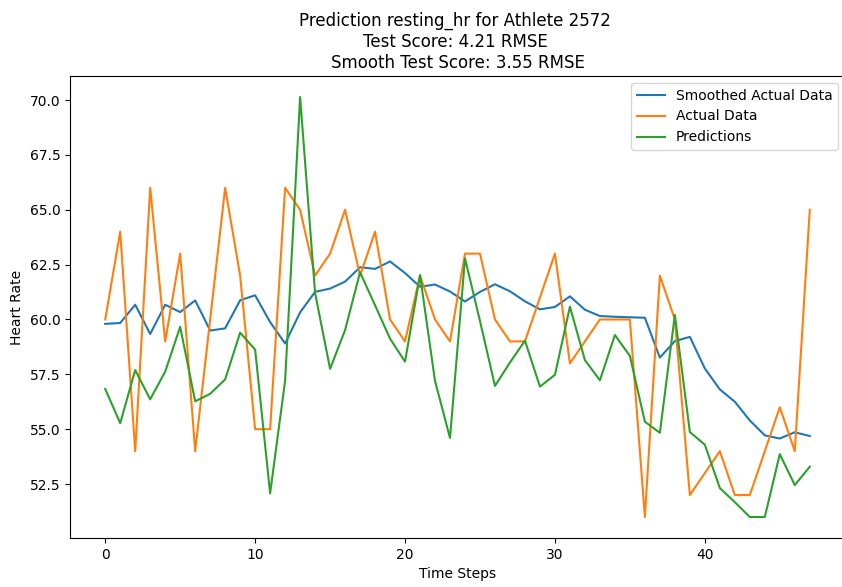


FIGURE 35: Prediction of resting HR for athlete 2572

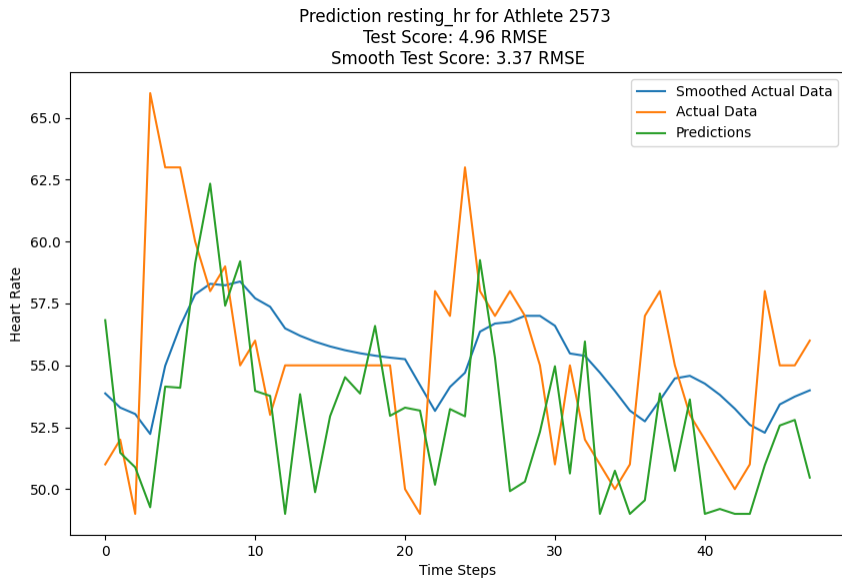


FIGURE 36: Prediction of resting HR for athlete 2573

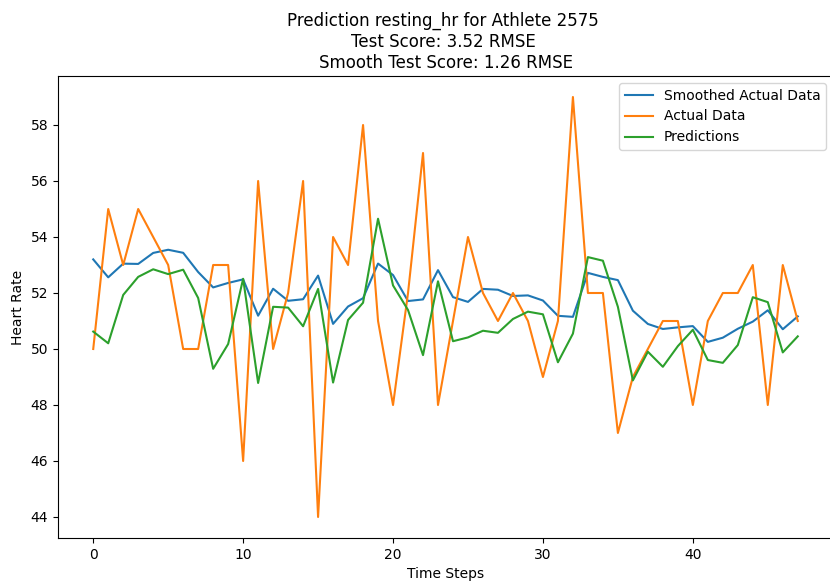


FIGURE 37: Prediction of resting HR for athlete 2575

9.3 Multivariate LSTM network 2

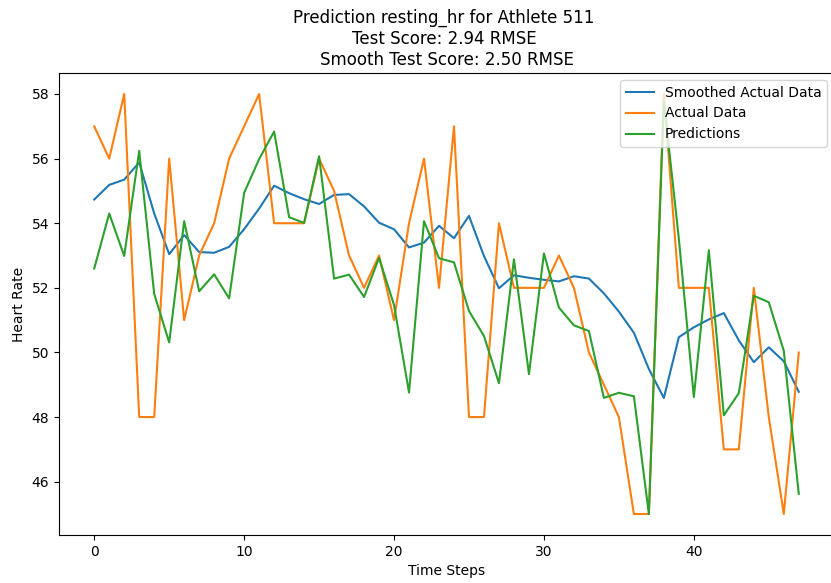


FIGURE 38: Prediction of resting HR for athlete 511

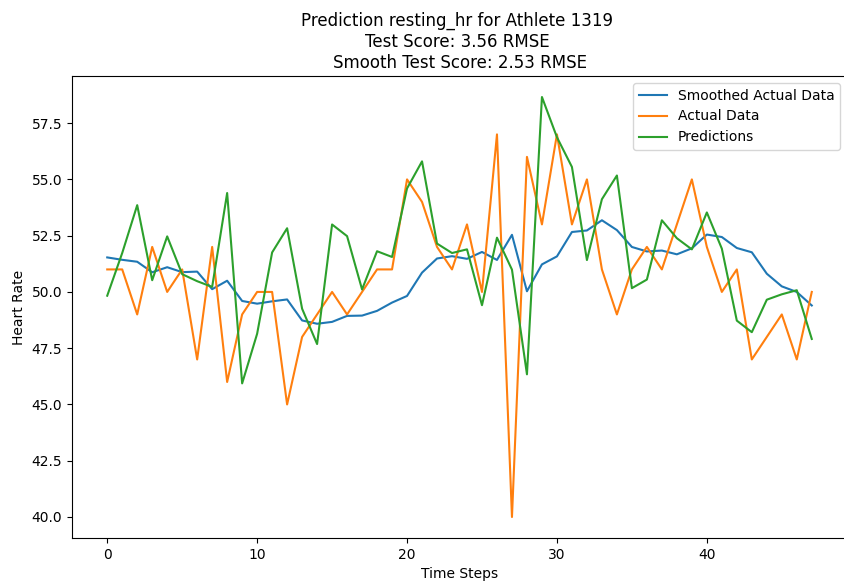


FIGURE 39: Prediction of resting HR for athlete 1319

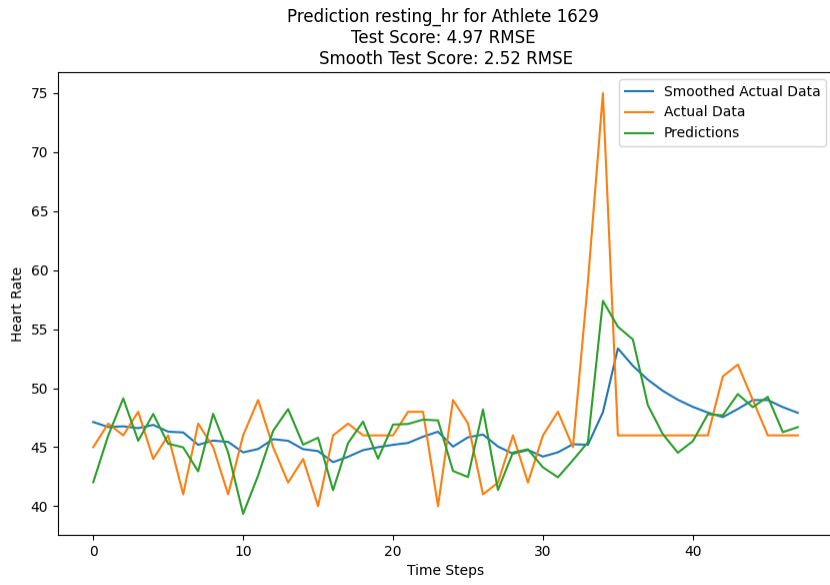


FIGURE 40: Prediction of resting HR for athlete 1629

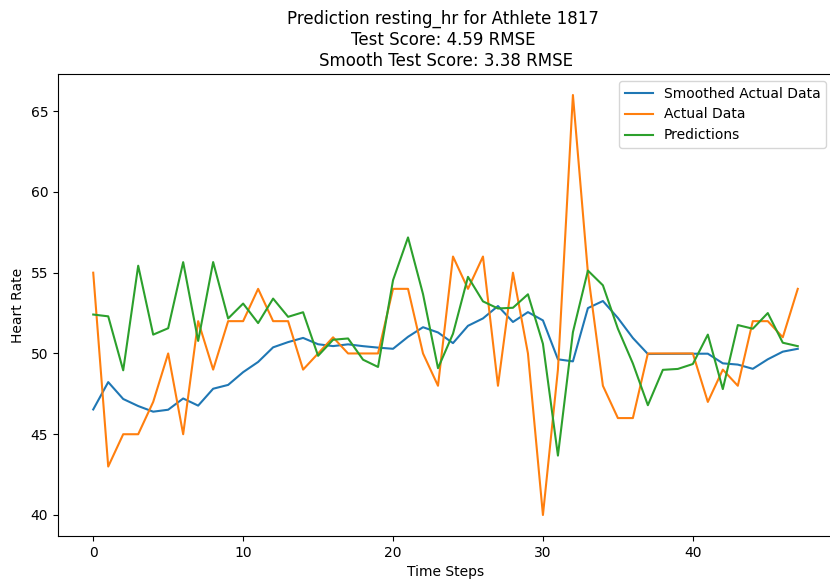


FIGURE 41: Prediction of resting HR for athlete 1817

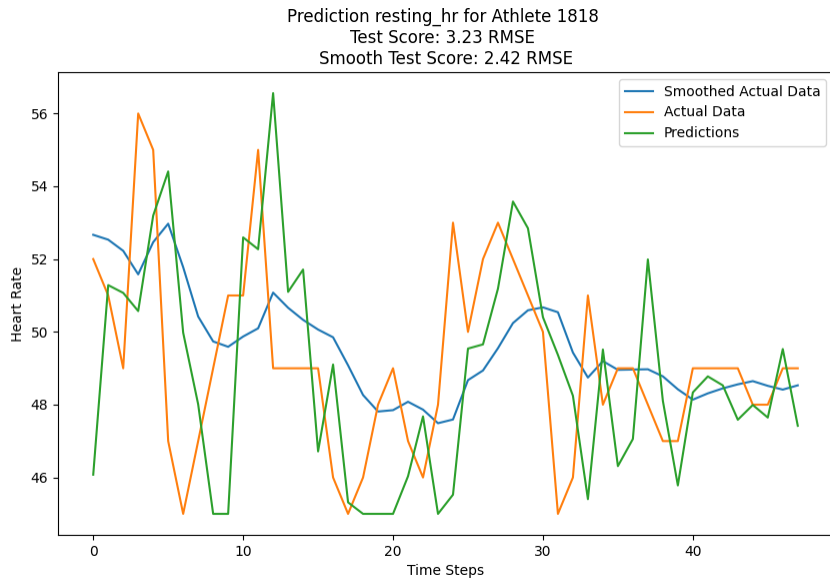


FIGURE 42: Prediction of resting HR for athlete 1818

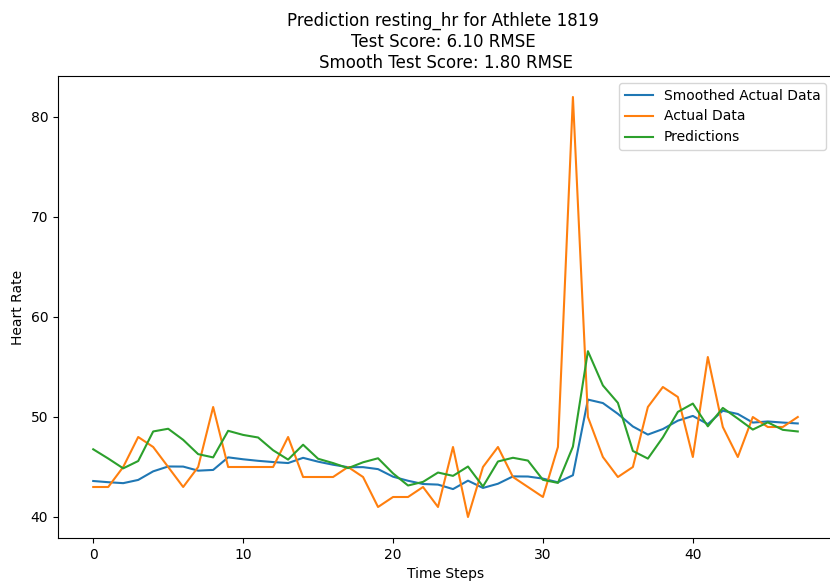


FIGURE 43: Prediction of resting HR for athlete 1819

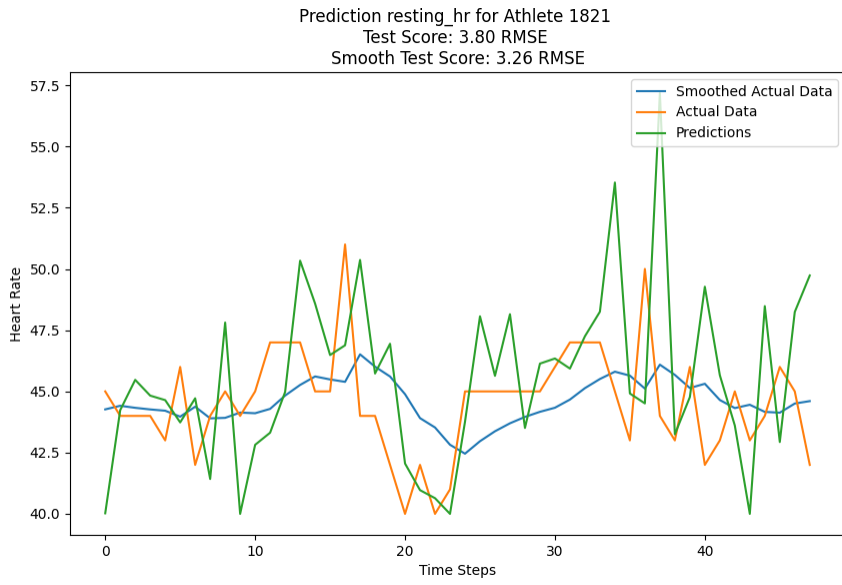


FIGURE 44: Prediction of resting HR for athlete 1821

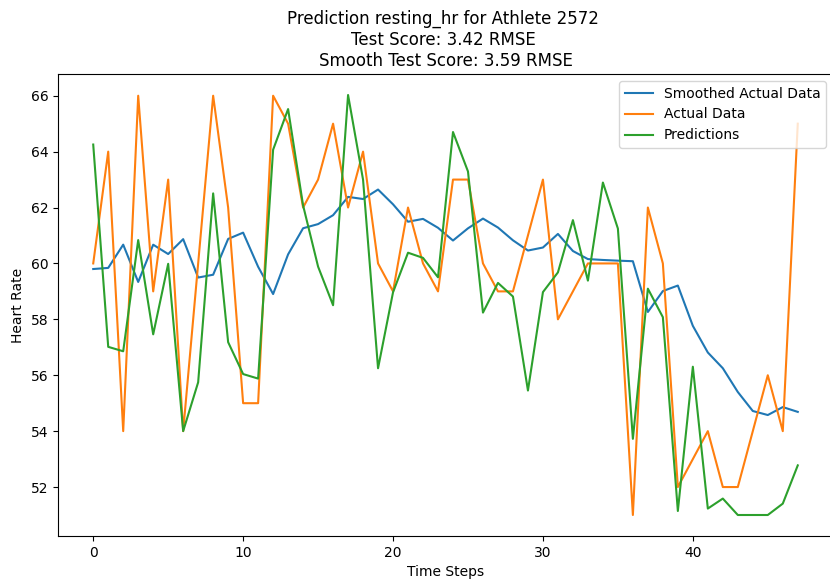


FIGURE 45: Prediction of resting HR for athlete 2572

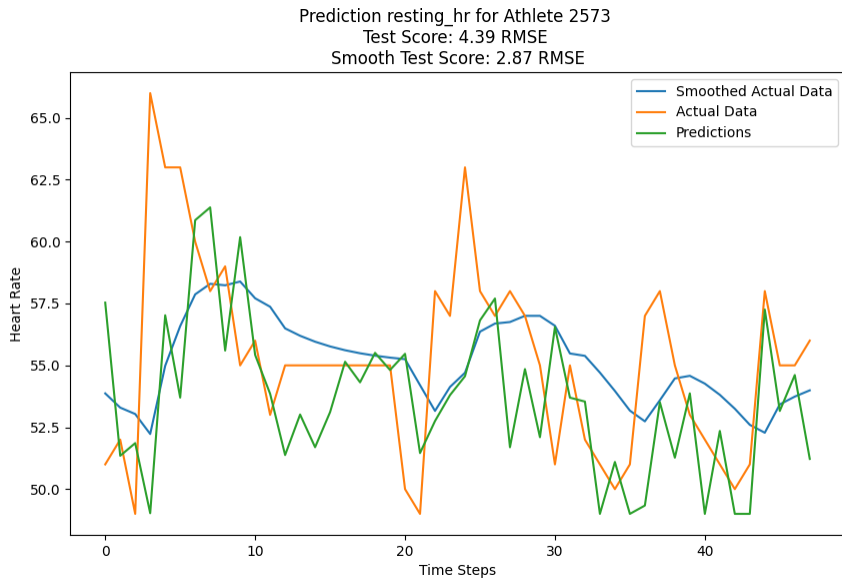


FIGURE 46: Prediction of resting HR for athlete 2573

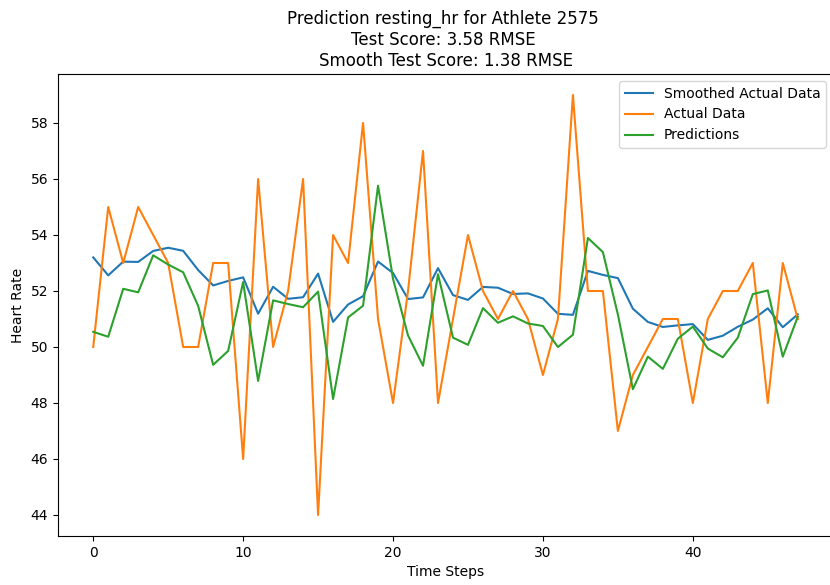


FIGURE 47: Prediction of resting HR for athlete 2575