KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Data Science Approaches for the Analysis and Interpretation of Training Load Data of Athletes

Tim Op De Beéck

Supervisor:
Prof. dr. J. Davis

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

June 2019

Data Science Approaches for the Analysis and Interpretation of Training Load Data of Athletes

Tim OP DE BEÉCK

Examination committee:
Prof. dr. ir. J. Vandewalle, chair
Prof. dr. J. Davis, supervisor
Prof. dr. ir. H. Blockeel
Prof. dr. W. Helsen
Dr. ir. W. Meert
Prof. dr. B. Vanwanseele

Prof. dr. B. Negrevergne
(Université Paris Dauphine, France)
Prof. dr. A. Zimmermann
(Université Caen Normandie, France)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

June 2019

Uitgegeven in eigen beheer, Tim Op De Beéck, Celestijnenlaan 200A box 2402, B-3001 Leuven (Belgium)

Preface

For the past couple of years I had the opportunity to work on a project that I was passionate about. The project allowed me to combine my interests in sports training and technology.

Looking back at these past years, I realize that I could not have written this thesis on my own. Therefore, I would like to thank the following people:

My promotor, Jesse, thank you for your guidance and mentorship along the way. Our frequent meetings really helped to drastically improve my technical skills and to develop my critical thinking. I really appreciate that you were always very honest with me and I quickly learned to value your direct way of communicating. You knew to highlight the positive points when I was frustrated and you quickly figured out how you could boost my focus when it was crunch time. It was a real honor to be a part of your research team.

Wannes and Jan, for our countless discussions and your practical tips and tricks. Nevertheless, I have to admit that it was sometimes frustrating that you came up with a solution within the hour for something I had been struggling with for a week.

Kurt, for passing along your passion about running and for making something fun of brainstorming, early morning meetings, collecting data, and even writing project proposals.
Benedicte, for the opportunity to collaborate with your group, for helping me to interpret my results, and for your guidance on how to properly collect data. I must admit, it is not as easy as I thought it was, especially during winter...

All runners, that I convinced to run and suffer for me during winter. Standing outside in the cold for three months would not have been possible without your enthusiasm, warming conversations and your motivation to push your limits. Mike, for teaching me a lot of useful English expressions, and for helping me to put on the sensors in the meantime.

Arne, for the pleasant collaboration throughout the many revisions of our papers and the countless meetings.
Michel, Filip, and Wouter, for your critical feedback and suggestions that helped to improve both manuscripts.
Werner, for your guidance, and expertise, and for helping me to establish a collaboration between your research group and the one of Jesse. Since the master thesis that I worked on together with Jeroen, I'm glad to see that the collaboration moved forward and both groups now better understand how they can complement each other.

My jury members, Hendrik, Albrecht, Benjamin, Werner, Benedicte, and Wannes, for your critical feedback that really helped to improve the end result of this thesis.
My chair Joos, for chairing both my preliminary and public defense.

TopSportsLab, and in particular Jan and Steven, for helping Jeroen and me to explore the possibilities for our master thesis topic, for providing us with the necessary data, and for the many meetings that followed. This master thesis really sparked my interest to become a researcher.
Jeroen, for turning a master thesis into something fun, for setting the time record in my outdoor data collection study and for all the good times we had together so far!

Peter, Arjen and Maarten, for our collaboration on our ILP paper and for hosting me during my research visit. Working in a different research environment for a month was a very inspiring and instructive experience in many ways.

My colleagues at DTAI and in particular, Jan, Tom, Irma, Vincent, Arne, Pieter, Jessa, Toon, and Elia for creating such a nice working atmosphere. Vincent, I really enjoyed our lunch meetings, we should schedule a next one soon.

Finally, I do not only want to thank the people in my direct working environment, but also the people that surrounded me during the past years and made this a time to never forget:

My house mates, Wancho, Bendrick, Ryvers, Robainz, Krikke, Jackie, Michel, Makke, and frequent visitors, Fax, Thokke, Ruben, and Akke, for all the interesting discussions over dinner, heated card games, basement swimming, hot-lightning meals, and BBQs.

My fellow team mates of the Belgian National ultimate open team, the coaching staff, my team mates and fellow captains of Jetset Open, due to your enthusiasm and our shared goals, it was never too much effort to go to practices and to give a 100% at all times. Playing ultimate with you was a welcome way to clear my mind and replenish my energy levels.

My family, mama, papa, Bram and Nele, and my recently acquired family Lieve, Jan, Rogier, Carolina, Vincent, and Ciska for being always there to help me out, to support me, to motivate me, and for the workouts, family dinners, coffee breaks, flip-turns, turtle encounters, and all other fun moments we shared together.

My wife, and best friend, Laure, for always watching my back, for reminding me to put things into perspective and to enjoy whatever I'm doing, even if there are a lot of different things going on at the same time.

My son, Rover, for making me laugh when I'm not supposed to approve your behavior, for lengthening my days so that I had more time to enjoy life and work on my thesis, and for enthusiastically waving at people until it makes them feel uncomfortable. Rovermans, you are my hero!

Abstract

Research on the analysis of real-world sports data dates back at least to 1958 (Lindsey 1959; Rubin 1958). Advances in technology have caused an explosion of the amount of sports-related data about sports. The abundance of data has attracted the interest of both the academic community and the industry. The aim of this sports analytics community is to leverage the available data to help decision makers to gain a competitive advantage (Alamar and Mehrotra 2011). The advent of wearable technology has yielded a new data source that still has a lot of unexplored potential. These data can assist practitioners to monitor athletes during daily life activities (Kwapisz et al. 2011) and rehabilitation (Um et al. 2017; Whelan et al. 2016), to quantify their training loads (Bourdon et al. 2017; Halson 2014; Jaspers, Brink, et al. 2017), and to analyze their risk of injury (Gabbett and Ullah 2012).

From a data science perspective, these continuous monitoring data pose several interesting data challenges. First, combining the data of different athletes is non-trivial due to inter-individual differences. Second, because the behavior of athletes can change and because often only limited individual data are available, it is also non-trivial to model the data on an individual level. Third, the use of subjective measures to quantify certain aspects of the athlete (e.g., perceived wellness), confounding factors (e.g., running speed), and missing values further complicate the analysis of these data.

In this thesis we evaluated how data science techniques can provide value to the analysis and interpretation of athletes' training load data. Our main focus is on the analysis of training load data from soccer players and outdoor runners. Specifically, we examined three relevant relationships. First, we studied how soccer players perceive external loads. Second, we modeled the relationship between external and internal load, and perceived wellness of soccer players. Third, we analyzed the relationship between biomechanical movement data of outdoor runners and their perceived fatigue status.

We presented three types of evidence to support the dissertation statement. First, we found that both data-driven feature selection methods and simple statistical features can complement expert knowledge. Second, we illustrated that group models can be used to individually monitor an athlete when limited-to-no prior data are available for that athlete. Third, we showed that machine learning techniques are well suited to model the complex relationships that are relevant for the analysis of athletes' training load data: non-linear relationships, relationships between objective and subjective variables, and relationships where multicollinearity exists among the input variables.

Additionally, we formulated some lessons learned for data scientists. We argued that modeling the context of and athlete's data, either explicitly or implicitly, can improve the performance of predictive models by adjusting for inter- and intra-subject differences and external factors. We presented several such strategies: standardizing features relative to an individual baseline, predicting a normalized target variable instead of the originally reported target variable, and adding the previous state as a feature. Moreover, we identified subtle data dependencies, that hinder obtaining an unbiased estimation of a model's ability to generalize to unseen data.

We identified three limitations of the current thesis. First, we evaluated the methodologies to monitor soccer players on the data of only one club. Second, the data collection protocol to collect outdoor data from runners experimentally controlled for total distance, intensity, and running surface and might have introduced a bias towards reporting higher fatigue scores near the end of the protocol. Third, RPE, a subjective measure used in every relationship of this thesis, quantify muscular fatigue, as well as cardiovascular and psychological fatigue.

Future research in this area can benefit from an interdisciplinary collaboration between data scientists, sports scientists and other domain experts. A close collaboration throughout all phases of the data science process can further advance the state of the art. First, it will improve the quality of the data that is being collected. Second, it can help to properly contextualize the data when modeling relevant relationships. Third, it will allow obtaining an unbiased estimation of these predictive models.

Beknopte Samenvatting

Onderzoek omtrent de analyse van gegevens van sporters tijdens competitie gebeurde reeds in 1958 (Lindsey 1959; Rubin 1958). Technologische vooruitgangen zorgden voor een explosie aan gegevens binnen de sport. De overvloed aan gegevens in de meeste sporten heeft de interesse gewekt van zowel de academische gemeenschap als van de industrie. Het doel van deze sportanalyse gemeenschap is om de beschikbare gegevens te gebruiken om beslissingsmakers een competitief voordeel te bezorgen (Alamar en Mehrotra 2011). De opkomst van draagbare technologie heeft voor een nieuwe gegevensbron gezorgd die nog veel potentieel heeft. Deze gegevens kunnen beslissingsmakers helpen om: atleten op te volgen tijdens dagelijkse activiteiten (Kwapisz e.a. 2011), revalidatie (Um e.a. 2017; Whelan e.a. 2016), om hun trainingsbelastingen te quantificeren (Bourdon e.a. 2017; Halson 2014; Jaspers, Brink e.a. 2017), en om hun risico op blessures te analyseren (Gabbett en Ullah 2012).

Vanuit een data science perspectief zorgen deze continue monitoring gegevens van atleten voor verscheidene uitdagingen. Ten eerste is het niet triviaal om gegevens van verschillende atleten te combineren omwille van individuele verschillen. Ten tweede is het niet triviaal om deze gegevens op een individueel niveau te modelleren omdat het gedrag van atleten kan veranderen en omdat er vaak maar een beperkte hoeveelheid individuele gegevens beschikbaar zijn. Ten derde wordt de analyse verder bemoeilijkt door het gebruik van subjectieve maatstaven om bepaalde aspecten van atleten te quantificeren (v.b., welzijn) en de aanwezigheid van verstorende factoren (v.b., loopsnelheid).

In deze thesis evalueerden we hoe data science technieken een meerwaarde kunnen bieden voor de analyse en interpretatie van trainingsbelastingsgegevens van atleten. We spitsten ons toe op de analyse van trainingsbelasting van voetballers en de bewegingen van outdoor lopers. Meer specifiek onderzochten we drie relevante relaties. Ten eerste bestudeerden we hoe voetballers externe belasting waarnemen. Ten tweede modelleerden we de relatie tussen externe en interne belasting enerzijds, en het gerapporteerde welzijn van voetballers

anderzijds. Ten derde analyseerden we de relatie tussen biomechanische bewegingen van outdoor lopers en hun gerapporteerde vermoeidheidstoestand.

We onderbouwden de thesisstelling op drie manieren. Ten eerste vonden we dat gegevensgestuurde feature selectie methodes en eenvoudige statistische features complementair kunnen zijn aan de kennis van experten. Ten tweede illustreerden we dat groepsmodellen gebruikt kunnen worden om een atleet individueel op te volgen, zelfs wanneer er weinig gegevens van die atleet beschikbaar zijn. Ten derde toonden we aan dat machine learning technieken goed geschikt zijn voor het modelleren van de complexe relaties die relevant zijn voor de analyse van trainingsbelastingsgegevens van atleten: niet-lineaire relaties, relaties tussen objectieve en subjectieve variabelen, en relaties waarbij de invoer variabelen aan elkaar gecorreleerd zijn. Daarnaast formuleerden we ook enkele lessen voor data scientists. We argumenteerden dat de prestaties van voorspellende modellen verbeterd kunnen worden door de context van atleten expliciet of impliciet mee in rekening te nemen. We stelden verschillende strategieën voor: door het standardiseren van features ten opzichte van een individuele baseline, door een genormaliseerde doelvariabele te voorspellen in plaats van de origineel gerapporteerde variabele, en door de vorige toestand als feature toe te voegen. Verder identificeerden we ook subtiele afhankelijkheden in de gegevens die een correcte evaluatie van hoe goed een model veralgemeent verhinderen.

We lichtten ook drie limitaties van de huidige thesis toe. Ten eerste evalueerden we de methodologieën voor het monitoren van voetballers met de gegevens van één club. Ten tweede controleerde het outdoor lopers protocol experimenteel voor de totale afgelegde afstand, intensiteit en ondergrond en mogelijk zorgde het protocol ervoor dat lopers de neiging hadden om naar het einde toe hogere RPE scores te rapporteren. Ten derde is RPE, een subjectieve schaal die in elke gemodelleerde relatie in deze thesis werd opgenomen, een maat die niet alleen spiervermoeidheid, maar ook de cardiovasculaire en psychologische vermoeidheid meet.

Toekomstig onderzoek in dit domein kan profiteren van een interdisciplinaire samenwerking tussen data scientists, sports scientists en andere domein experten tijdens alle stappen van het data science proces. Ten eerste zal het de kwaliteit van gegevens verbeteren. Ten tweede zullen de gegevens in de juiste context geplaatst kunnen worden om relevante relaties te modelleren. Ten derde zal het toelaten om een correcte evaluatie van de voorspellende modellen te bekomen.

A	arm
AAD	average absolute difference
AFL	Australian football league
AM	all runners model
ANN	artificial neural network
AU	arbitrary unit
AUC	area under the curve
CI	confidence interval
DFA	detrended fluctuation analysis
dRPE	differential RPE
ELI	external load indicator
EN	elastic net
FPR	false positve rate
FPW	future perceived wellness
GBRT	gradient boosted regression tree
GPS	global positioning system
h	hour
HR	heart rate
HSR	high speed running
Hz	Hertz
IM	individual model
IMU	inertial motion unit
km	kilometer
LASSO	least absolute shrinkage and selection operator
m	meter
MAE	mean absolute error
ML	machine learning
NRPE	normalized rating of perceived exertion
OM	other runners only model
POMS	profile of mood states
PPW	pre-session perceived wellness
RHIE	repeated high intensity efforts
ROC	receiver operating characteristic
RPE	rating of perceived exertion
s	second
SD	standard deviation
sRPE	session RPE
Stat.	statistical
Symm.	symmetry
T	tibia
TD-baseline	trial dependent baseline
TPR	true positive rate
VHSR	very high speed running
W	wrist

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Introduction

Research on the analysis of real-world sports data dates back at least to 1958 (Lindsey 1959; Rubin 1958). Over the years, advances in technology have caused an explosion of the amount of data collected about sports. Written scouting reports, box scores and raw camera footage have evolved to much richer data sources:

Event stream data These data capture all game events (e.g., passes, dribbles, tackles) in a structured format. For each event, the format lists, the players involved, its timestamp, and location (Opta Sports 2018).

Positional data Motion capture systems (SciSports 2018; STATS' SportVU 2018) allow tracking the position of athletes up to 100 times per second during competition (Alamar and Mehrotra 2011).

This abundance of data has attracted the interest of both the academic community and the industry. The aim of the sports analytics community is to leverage the available data to help decision makers to gain a competitive advantage (ibid.).

Over the years, the analysis of these data have evolved from descriptive statistics to advanced predictive models. Data science techniques have found their way into several popular forms of data analytics:

Match outcome prediction Models to predict the outcome of games have been proposed for soccer (Van Haaren and Davis 2015), basketball (Zimmermann et al. 2013), ice hockey (Marek et al. 2014) and tennis (Spanias and Knottenbelt 2013).

Player projections Models to forecast how players will perform in the future were proposed for soccer (Vroonen et al. 2017), basketball (Hwang 2012), and ice hockey (Liu and Schulte 2018).

Action rating Researchers developed models to quantify actions of players in soccer (Bransen et al. 2018; Decroos, Bransen, et al. 2018), basketball (Cervone et al. 2014), ice hockey (Schulte et al. 2017), and volleyball (Bagley and Ware 2017).

Strategy analysis The available spatio-temporal data showed the potential to reveal strategical patterns in soccer (Lucey et al. 2013; Van Haaren, Dzyuba, et al. 2015), basketball (Miller et al. 2014), volleyball (Van Haaren, Horesh, et al. 2016), tennis (Wei et al. 2015), and marathon running (Smyth and Cunningham 2018).

Advanced metrics Advanced metrics such as the expected goals metric have been introduced. This metric computes the likelihood of scoring given a description of the game state. Variants of the metric exist for soccer (Decroos, Dzyuba, et al. 2017), ice hockey (Macdonald 2012), American football (Pasteur and Cunningham-Rhoads 2014), and basketball (Chang et al. 2014).

While the sports analytics community has largely focused on the analysis of performance during competition, the advent of wearable technology (see section 2.1) has yielded a new data source that still has a lot of unexplored potential. These continuous monitoring data capture other aspects such as:

GPS data GPS sensors (Catapult 2018; STATSports 2018) track the position of athletes up to 10 times per second. These sensors can quantify distance covered, running speed, accelerations and decelerations during competition and training.

Biomechanical data Inertial motion units quantify biomechanical parameters (Shimmer 2019) (e.g., impacts, accelerations, jumps) and daily life activities (Fitbit 2019).

Physiological data Heart rate monitors (Polar 2019) and other wearable sensors track physiological signals (e.g., heart rate, skin temperature, and galvanic skin response) (Li et al. 2017).

Questionnaire data Mobile apps are a convenient tool to allow athletes to track other aspects (e.g., nutritional data, wellness data, ratings of perceived exertion) in the form of questionnaires and digital diaries.

These data can assist practitioners to monitor athletes during daily life activities (Kwapisz et al. 2011) and rehabilitation (Um et al. 2017; Whelan et al. 2016), to quantify their training loads (see section 2.2) (Bourdon et al. 2017; Halson 2014), and to analyze their risk of injury (Gabbett and Ullah 2012).

From a data science perspective, these continuous monitoring data from athletes pose several interesting data challenges:

Small sample sizes It is not always feasible to collect a lot of data about each individual. In a research context, it can be time consuming to attach the sensors and to record and read out the data. Moreover, it is not always possible to collect data of multiple athletes at the same time. In a real-world sports setting the total number of sessions that can be monitored during a season is limited because of regulations, player transfers, injuries, international weeks, or coach preferences. Thus, when these longitudinal data are aggregated per session, individual data sets are often small. Small data sets in combination with intra-subject differences, noise and confounding factors can make it non-trivial to construct accurate individual models. Therefore, it is often necessary to construct models using data of multiple athletes.

Individual characteristics of athletes Individual characteristics (e.g., age, anthropometrics, strength, injury history) of athletes add complexity to the data analysis because they introduce inter- and intra-subject differences in the data. Inter-subject differences in the data arise because these individual characteristics determine how athletes move (e.g., running style), how they respond to a load that they are subjected to, and what type of injuries they are susceptible to. Thus, combining data of different athletes is non-trivial. Intra subject differences arise because some individual characteristics of an athlete fluctuate over time. Thus, an athlete's movements, load response, and injury risk profile evolve over time which complicates the analysis on an individual level.

Subjective data Subjective measures are popular to quantify an athlete's perception of training load as well as aspects of the athlete's wellbeing. However, the subjectivity of these data complicates their analysis because an athlete's perception is individual and can evolve over time.

External factors External factors further complicate data analysis. Sensor positioning and attachment can introduce noise and motion artifacts

into the data. Confounding factors such as running speed, terrain, and temperature can influence the data as well. In a real-world setting it is not always feasible to control for these variables. Thus, it is often necessary to adjust for these variables when analyzing the data.

Missing data Several factors can introduce missing data. Some missing values are introduced at random (e.g., no GPS signal, sensor defect) and can typically be imputed. However, other types of missing data are non-trivial to impute (e.g., missing values because players did not fill out their questionnaire on match days and recovery days).

In this thesis we focus on the analysis of training load data of soccer players and outdoor runners.

The core principle of training is that it causes biological adaptations in athletes. These changes improve the fitness of the athlete and as a result can improve the athlete's performance potential (Soligard et al. 2016).

Unfortunately, both poor management of training loads and non-sport stressors can cause maladaptations in athletes. These maladaptations can lead to decreased performance and injuries. This underscores the importance of monitoring loads (i.e., stimuli that are applied to a human biological system) that are applied to athletes, understanding how athletes respond to these loads and assessing whether they need recovery or extra training (ibid.).

Measures of training load (i.e., loads applied during a training session) measure either external load (i.e., all locomotor activities performed by athlete) or internal load (i.e., how the athlete responds to a given external load) (Bourdon et al. 2017). External load can be characterized using objective variables (e.g. total distance covered). Internal load can be measured using both objective (e.g., heart rate) and subjective measurements (e.g., rating of perceived exertion). Yet, the subjective measures are often more practical to use, more specific (Brink et al. 2010) and more sensitive (Saw, Main, et al. 2016) compared to available objective measures for internal load.

Different subjective internal load measures are being used. The global rating of perceived exertion (i.e., RPE) is a subjective and holistic measure to capture an overall feeling of fatigue. This global RPE score simultaneously captures the cardiovascular, mechanical and psychological fatigue of an athlete. Differential RPE or dRPE (Vanrenterghem et al. 2017) breaks down the global concept of fatigue into different more specific components that are scored separately (e.g., cardiovascular, biomechanical and technical demands). Three popular scales to measure ratings of perceived exertion are being used to study runners and soccer players: the original Borg RPE scale (Schütte, Seerden, et al. 2018), the CR10-scale, and the CR100-scale (Fanchini et al. 2016). Wellness questionnaires

can be used to capture acute and chronic changes in athlete wellness (e.g., fatigue, general muscle soreness and sleep quality) in response to load (Saw, Main, et al. 2016).

While seemingly similar, the different internal load measurements can actually complement each other. Typically, the global RPE measure is used to quantify the internal load of an entire training session. Measuring the global RPE score repeatedly during a session allows studying the evolution of the internal load within a session. The dRPE can either replace the global RPE measurement, or can be used jointly with the global RPE to explain the global RPE measurement. Wellness questionnaires are typically filled out by athletes before the first training session of the day and capture the response to all sports and non-sports loads of the preceding days (Thorpe et al. 2017). The wellness state of the athlete influences how the athlete responds to the subsequent external load that the athlete is subjected to.

While measuring the external loads helps to accurately dose the load an athlete is subjected to, the internal load helps to assess whether the response to the external load matches the intended response. The internal-load-to-external load ratio can be used to infer the training status of an athlete (i.e., does the athlete need more recovery or more training) (Buchheit, Racinais, et al. 2013). An increase in internal load to a standardized external load indicates that the athlete is more fatigued, whereas a decrease in internal load reveals an increase in fitness of the athlete (Bourdon et al. 2017). In practice, standardized external loads can be approximated by the use of standardized conditioning tests (e.g., the yo-yo intermittent recovery test (Bangsbo et al. 2008)). However, it is not feasible to frequently conduct these tests as they are time consuming and they can disrupt an athlete's training program. Therefore, it would be valuable to assess the training status of athletes from the internal-to-external-load ratio of normal training sessions. However, external load is characterized by many external load indicators. Thus, the ratio between the internal load and all these indicators should be evaluated simultaneously. This is a non-trivial task as no such normative data exists.

In this thesis we use data science approaches to model this relationship. The use of data science techniques to analyze the data that are collected in these fields is still limited.

In the field of training load monitoring most models are hand crafted and only consider one input variable at a time (Banister et al. 1975; Buchheit, Racinais, et al. 2013; Hulin et al. 2016). These models make simplifications or assumptions that might not hold in a real world setting (Hellard et al. 2006; Jobson et al. 2009). However, recently, the training load community acknowledged the

potential of machine learning (see section 2.3) techniques (Bourdon et al. 2017). Previously, artificial neural networks (ANNs) were used to predict the training response of Australian football players (Bartlett et al. 2017). Yet, these models only used a small set of hand selected input variables.
Over the past years, the field of biomechanics showed an increasing interest in machine learning techniques (Halilaj et al. 2018). Yet, the protocols employed in these studies often experimentally control for confounding factors such as running speed (E. Mitchell et al. 2015) or artificially induce asymmetries (ibid.) or fatigue (Buckley et al. 2017). Therefore, analyzing more realistic data might reveal new interesting data challenges.

1.1.1 Dissertation Statement

In this thesis we evaluate the following statement:

"Data science techniques can provide value to the analysis and interpretation of training load data of athletes".

To evaluate this statement we examine the following questions:

1. Is a data driven approach to identify important variables complementary to expert knowledge?
2. Do group models provide value for the individual monitoring of athletes?
3. Are machine learning models well suited to model continuous monitoring data of athletes?

1.1.2 Contributions

To summarize, the main contributions of this thesis are:

- We identify the external load indicators that are perceived as most exerting by soccer players
- We show that group models can be used for individual monitoring of the training load response of soccer players.
- We find that acute external and internal load in combination with preceding perceived wellness are most predictive of future perceived player wellness.

- We present a methodology that effectively accounts for running speed, the subjectivity of the target variable and inter- and intra-individual differences between runners.
- We show that the fatigue status of a runner can accurately be predicted with limited or no prior labeled data of a runner using a set of simple features computed on the data of one IMU-sensor attached to the wrist.

1.1.3 Structure of the Thesis

This thesis is structured as follows:

Chapter 2 provides the reader of this thesis with the necessary background to read the subsequent Chapters.

In Chapters 3 through 5 we focus on three applications where we apply data science approaches to analyse training load data in close collaboration with domain experts.

All three applications adhere to the same general methodology that consists of five steps. First, we define research questions that are relevant for the domain. Second, we decide on the most suitable data set to analyse. Third, we define relevant subsets of the data that will be used to learn predictive models. Fourth, we design an evaluation strategy to obtain an unbiased estimation of the accuracy of the predictive models. Fifth, we construct predictive models for each subset using various popular traditional machine learning algorithms. These algorithms learn a model from a set of feature vectors. This approach increases the buy-in of domain experts as it allows the use of both meaningful features and interpretable models. This is important because the use of data science approaches to analyse training load data is still novel.

More specifically, Chapter 3 contributes to the evaluation of the training status of players. In this Chapter we model the relationship between the rating of perceived exertion (i.e., RPE) (Borg 1982; Borg 1998) reported 30 minutes after a training session and the external load performed during the training session. We examine the importance of the available external load indicators and assess the role of individual characteristics of players. This chapter was previously published as:

Jaspers, A.*, Op De Beéck, T.*, Brink, M., Frencken, W., Staes, F., Davis, J.+, Helsen, W.+ (2018). Relationships between the External and Internal Training

Load in Professional Soccer: What can We Learn from Machine Learning? International journal of sports physiology and performance, 13(5), 625-630.
(*) denotes shared first authorship, ($^+$) denotes shared last authorship

Chapter 4 contributes to the longitudinal wellness monitoring of athletes. In this Chapter we model how an athlete's perceived wellness is affected by the preceding training load. We evaluate the impact of both acute and chronic training loads. We quantify these training loads using internal and external load indicators. We also examine the potential of extra contextualization of the data. This Chapter was previously published as:

Op De Beéck, T.*, Jaspers, A.*, Brink, M., Frencken, W., Staes, F., Davis, J.$^+$, Helsen, W.$^+$ (2019). Predicting Future Perceived Wellness in Professional Soccer: The Role of Preceding Load and Wellness. International journal of sports physiology and performance, 1-25.
(*) denotes shared first authorship, ($^+$) denotes shared last authorship

Chapter 5 adds to the real-time monitoring of athletes. We collected a longitudinal outdoor data set of runners wearing multiple IMU sensors during an all out 3200m test. We examine how the fusion of these multiple biomechanical motion sensors reflects the fatigue status of the runner. We show how to deal with the subjectivity of the target label and the inter-individual differences between runners. We also show that the evaluation of these data is non-trivial. This Chapter was previously published as:

Op De Beéck, T., Meert, W., Schütte, K., Vanwanseele, B., Davis, J. (2018). Fatigue Prediction in Outdoor Runners Via Machine Learning and Sensor Fusion. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 606-615). ACM.

Chapter 6 reflects on Chapters 3 to 5. We revisit the dissertation statement, formulate lessons learned, discuss the limitations and point out potential directions for future research.

1.1.4 Other Research Conducted

To present a coherent story, this thesis focuses on the research related to the analysis of continuous monitoring data of athletes using data science techniques, where I am a first author. Yet, this is only a subset of the research that I have performed during my PhD. This section gives a short summary of the other publications that I have contributed to.

Mining Hierarchical Pathology Data using Inductive Logic Programming. In this study (Op De Beéck et al. 2015), we proposed a methodology based on inductive logic programming to extract novel associations from pathology excerpts. We discussed the challenges posed by analyzing these data and discussed how we addressed them. As a case study, we applied our methodology to Dutch pathology data for discovering possible causes of two rare diseases: cholangitis and breast angiosarcomas.

Published as: **Op De Beéck, T.**, Hommersom, A., Van Haaren, J., van der Heijden, M., Davis, J., Lucas, P., Nagtegaal, I. (2015, June). Mining Hierarchical Pathology Data using Inductive Logic Programming. In Conference on Artificial Intelligence in Medicine in Europe (pp. 76-85). Springer, Cham.

Data Fusion of Body-worn Accelerometers and Heart Rate to Predict VO2max during Submaximal Running. In this study (De Brabandere et al. 2018) we presented a model for recreational runners to estimate their VO2max from submaximal running on a treadmill. It requires two body-worn sensors: a heart rate monitor and an accelerometer positioned on the tibia.

Published as: De Brabandere, A., **Op De Beéck, T.**, Schütte, K., Meert, W., Vanwanseele, B., Davis, J. (2018). Data Fusion of Body-worn Accelerometers and Heart Rate to Predict VO2max during Submaximal Running. PloS One, 13(6).

AMIE: Automatic Monitoring of Indoor Exercises. In this study (Decroos, Schütte, et al. 2018) we examined the feasibility of a system that automatically provides feedback on correct movement patterns for patients performing physical therapy exercises. We introduced AMIE, a machine learning pipeline that detects the exercise being performed, the exercise's correctness, and if applicable, the mistake that was made.

Published as: Decroos, T., Schütte, K., **Op De Beéck, T.** , Vanwanseele, B., Davis, J. (2018, September). AMIE: Automatic Monitoring of Indoor Exercises. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 424-439). Springer, Cham.

Surface Effects on Dynamic Stability and Loading During Outdoor

Running using Wireless Trunk Accelerometry. The purpose of this study (Schütte, Aeles, et al. 2016) was to investigate outdoor surface effects on dynamic stability and dynamic loading during running using tri-axial trunk accelerometry. The results suggested that woodchip trails disrupt aspects of dynamic stability and loading that are detectable using a single trunk accelerometer.

Published as: Schütte, K. H., Aeles, J., **Op De Beéck, T.**, van der Zwaard, B., Venter, R., Vanwanseele, B. (2016). Surface Effects on Dynamic Stability and Loading During Outdoor Running using Wireless Trunk Accelerometry. Gait & posture, 48, 220-225.

Monitoring the Crus for Physical Therapy. The goal of this study (Van Craenendonck et al. 2014) was to investigate whether a 3D camera, such as the Microsoft Kinect, can be used to monitor the lower limbs of patients performing physical therapy exercises. This study presented two particle-filtering based algorithms for accurate tracking.

Published as: Van Craenendonck, T., **Op De Beéck, T.**, Meert, W., Vanwanseele, B., Davis, J. (2014). Monitoring the Crus for Physical Therapy. In 1st International Workshop on Machine Learning for Urban Sensor Data (pp. 1-16).

Chapter 2

Background

In this chapter we provide an overview of the background that is necessary to read the subsequent chapters of this thesis. First, we give more information on the wearables that were used for this thesis. Second, we introduce the concept of training load and motivate why the load of athletes should be monitored. Third, we explain several important machine learning concepts and provide an overview of the different machine learning algorithms that we use in chapters 3 to 5.

2.1 Wearable Sensors in Sports

This section gives more background on GPS sensors, accelerometers, and gyroscopes in the context of data collection for sports applications. Typically, two sensor characteristics should be considered when evaluating a sensor.

First, the sample rate (i.e., the number of measurements per second) is important because it determines how fine grained the data are. Higher sampling rates will allow capturing events that happen in a short amount of time (e.g., heel strike while running). However, there is a trade-off between the sampling rate on the one hand, and the battery life and storage capacity of the sensor on the other hand.

Second, the range of the sensor is important as it defines the minimum and maximum values that can be measured. When the actual values exceed this range, the sensor will clip the measured values to either the minimum or maximum value. Clipping introduces inaccuracies as noise in the data.

Throughout this thesis we analyzed data collected using the following three sensors:

GPS A GPS sensor can measure the speed and position of athletes. For runners, a typical GPS sensor has a sampling rate between 1/10 Hz and 1 Hz. A GPS to track athletes in team sports requires a sampling rate of 10 Hz to capture the changes of direction as well as accelerations and decelerations (M. Scott et al. 2016).

Accelerometer A (tri-axial) accelerometer can measure accelerations in three directions (i.e., in Gs). The sampling rate should be high (i.e., 500-1000 Hz (Willy 2018)) when it is important to detect gait events (e.g. initial foot contact during a running step). The range of the sensor should be between -20 Gs and 20 Gs (ibid.) to avoid clipping of the acceleration signals measured at the tibia (i.e., the shin) while running.

Gyroscope A gyroscope can measure rotations per second (i.e., degrees/s) along three axes (i.e., pitch, roll, yaw). Both the sampling rate and range of the sensor determine its applicability. Gyroscopes used to monitor movements of athletes during teamsports and research applications currently use a range between 250 and 2000 degrees per second and a sampling rate between 100 and 1000 Hz to capture subtle movements (Catapult 2018; Shimmer 2019).

When collecting data in sports, several other factors influence the quality of the data of accelerometers and gyroscopes.

First, the sensor location affects the data that can be recorded. To capture impacts while running for example, an accelerometer on the foot will register higher impacts compared to an accelerometer on the upper back. When the runner's foot hits the ground, a shock wave will travel upwards throughout the runner's body. As a mechanism to protect the brain from shocks, the runner's body will have already absorbed part of the shock wave by the time it reaches the sensor on the upper back.

Second, the orientation of the sensor relative to the structure it is attached to affects the data. Accelerometers and gyroscopes measure according to a local reference system (i.e., three axes). To not complicate the interpretation of the data it is important to align these axes with the structure the sensor is attached to (e.g. the shin). To avoid variation in the data, the same sensor position and sensor orientation should be used for each recording (e.g., when collecting longitudinal data).

Third, it is important that the sensor remains in place during movement. A combination of moving at high speeds, high accelerations or decelerations,

and sweating have the potential to move the sensor (i.e., sensor rotation or translation). High impacts can cause elastic movement of the sensor (i.e., the sensor moves relative to the structure, but returns to its original position with some delay). These type of movements can create artificial movement artifacts in the data.

2.2 Training Load Monitoring

Getting athletes in top shape while keeping them injury free is not an easy task. Practitioners should use a holistic approach to monitor their athletes (Verhagen and Gabbett 2019). First, they need to track the load that they apply to their athletes. Second, they should assess their load capacity (i.e., how much load can an athlete tolerate). Third, they have to follow up on their health and performance statuses. Fourth, they need to consider the individual context and environment of each athlete (ibid.). These four categories are not independent of each other. Figure 2.1 illustrates these dependencies on a high level. In this simplistic model, some relationships reflect a positive influence from one category on another (shown by solid lines), whereas other relationships reflect a negative influence from one category on another (shown by dashed line). In the following sections we will zoom in on different aspects of this model.

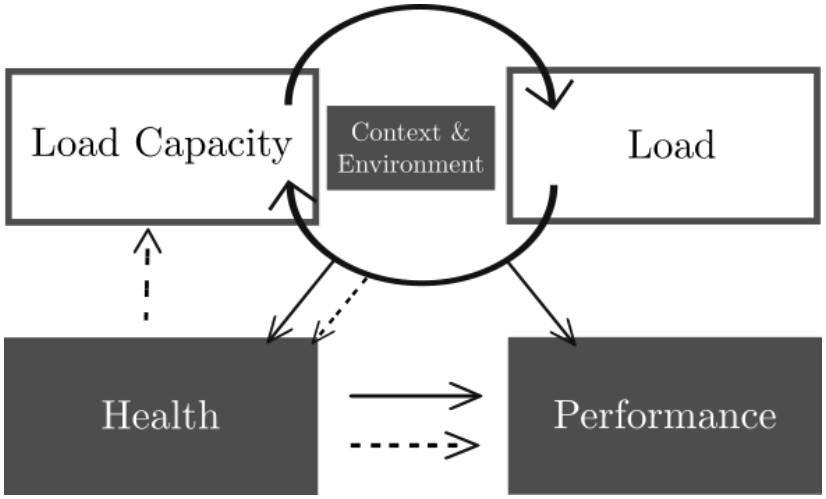


Figure 2.1: Relationships between load, load capacity, health and performance, redrawn from (Verhagen and Gabbett 2019). Solid lines indicate a positive relationship. Dashed lines indicate a negative relationship.

2.2.1 Training Load

Training load is a stimulus applied to a human biological system. This stressor can be physiological, psychological, or mechanical in nature (Soligard et al. 2016). The concept of training load can be broken down into external and internal load (see Figure 2.2). External load represents the dose performed by the athlete. Internal load represents the psychophysiological stress experienced by the athlete. The model in Figure 2.2 describes the training outcome as the consequence of the internal load. Two athletes performing the same external load can experience a different internal load. Individual characteristics of these athletes can explain this difference (Impellizzeri, Rampinini, and Marcora 2005).

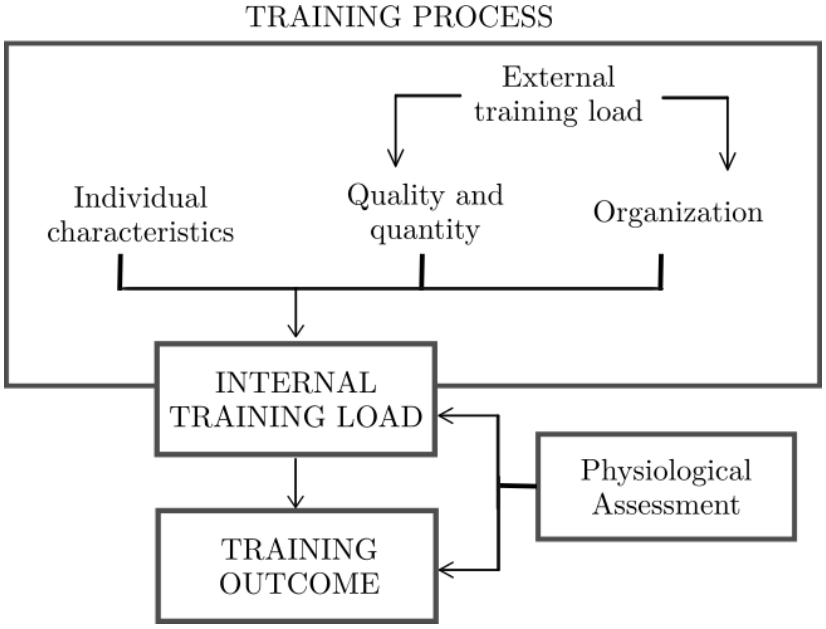


Figure 2.2: Model describing the relationship between individual characteristics of athletes, the external load, internal load and training outcome, redrawn from (Impellizzeri, Rampinini, and Marcora 2005).

2.2.2 Balancing Training Load and Training Load Capacity

A load that exceeds an athlete's load capacity, will induce training adaptations. The athlete's body will try to adapt to better deal with similar loads in the

future. This load versus load capacity relationship is delicate. A correct load results in health and performance benefits. In turn, these will increase the load capacity of the athlete. Yet, an excessive load or a load that is too low increases the risk for detrimental health effects (Verhagen and Gabbett 2019). When the health of an athlete drops, this is detrimental to the athlete's performance and it harms the athlete's load capacity (ibid.).

2.2.3 The Temporal Aspects of Training Load Monitoring and Injury Risk

Practitioners need to decide what the correct load is for a particular athlete at a given time. Yet, the athlete's context and environment influence load, load capacity and their balance. Both the context and environment change over time. These interactions influence the athlete's health status and risk for injury. The injury aetiology model in Figure 2.3 (Windt et al. 2017) describes these temporal relationships in more detail.

The athlete's internal risk factors or individual characteristics determine the athlete's context. This context predisposes the athlete for certain type of injuries. A training load results in a training outcome and elicits training effects. These training effects influence the athlete's modifiable internal risk factors. Other internal risk factors remain unaffected by these training effects.

Training load also exposes athletes to external risk factors (i.e., the environment). Yet, to sustain an injury, there needs to be an inciting event (e.g., an off-balance between load and load capacity).

A more recent conceptual framework (see Figure 2.4) breaks down the concept of training load into physiological and biomechanical load. Both types of load lead to adaptations of the athlete's biological system. In turn, these adaptations will influence, the individual characteristics of the athlete.

2.2.4 Appropriate Training Load Management: Structure-specific Load Capacity versus Sport-specific Load Demands

Assessing an athlete's risk of specific injuries, requires detailed monitoring of athletes. In this thesis we do not try to predict the risk of an athlete directly. Yet, we focus on the relationships between the different aspects of training load monitoring. A better understanding of these relationships can help to further

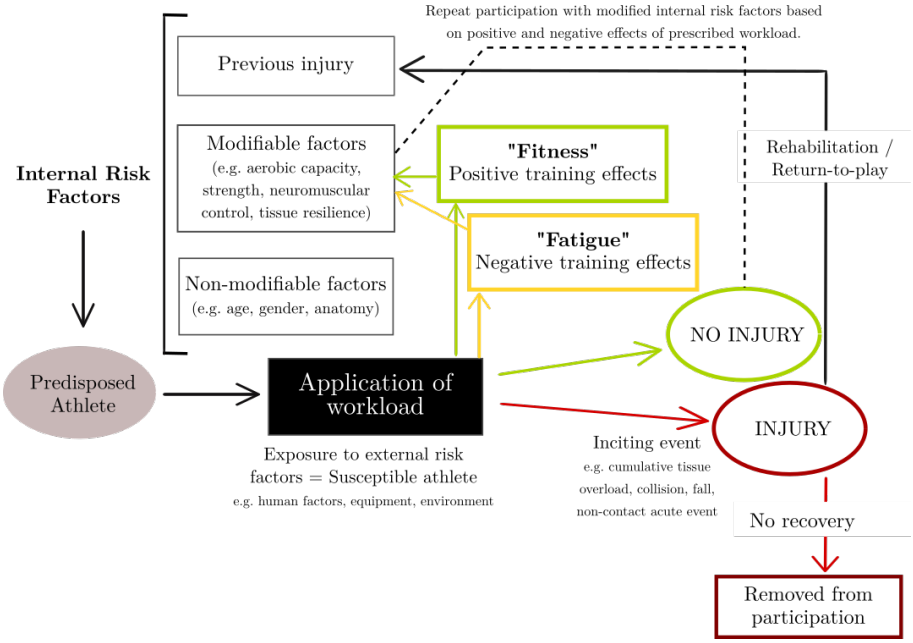


Figure 2.3: Model describing the aetiology of injuries and the temporal aspects of individual characteristics, redrawn from (Windt et al. 2017).

optimize and individualize the training programs of athletes by prescribing more appropriate loads to each athlete.

To understand why it is non-trivial to determine appropriate loads, it is important to realize that a practitioner should assess two aspects of his athlete. First, he should track structure (i.e. muscle, tendon or bone) specific loads. Second, he should assess the structure-specific load capacities. To avoid injuries each structure should tolerate the sport-specific load demands. If these are out of sync, the athlete is at high risk of sustaining an injury to that structure. Robust physical capabilities provide an athlete with high structure-specific load capacities. Yet, to develop these physical capabilities, the athlete needs to endure high loads. This circular causation (see Figure 2.5) motivates a gradual increase in training loads (Gabbett, Nielsen, et al. 2018).

In an applied setting, it is not always possible to measure structure-specific load. Nor is it possible to measure structure-specific load capacities. Yet, it is possible to track these two aspects through several proxies. First, a practitioner could track capacity related variables (e.g., strength). Second, he could measure

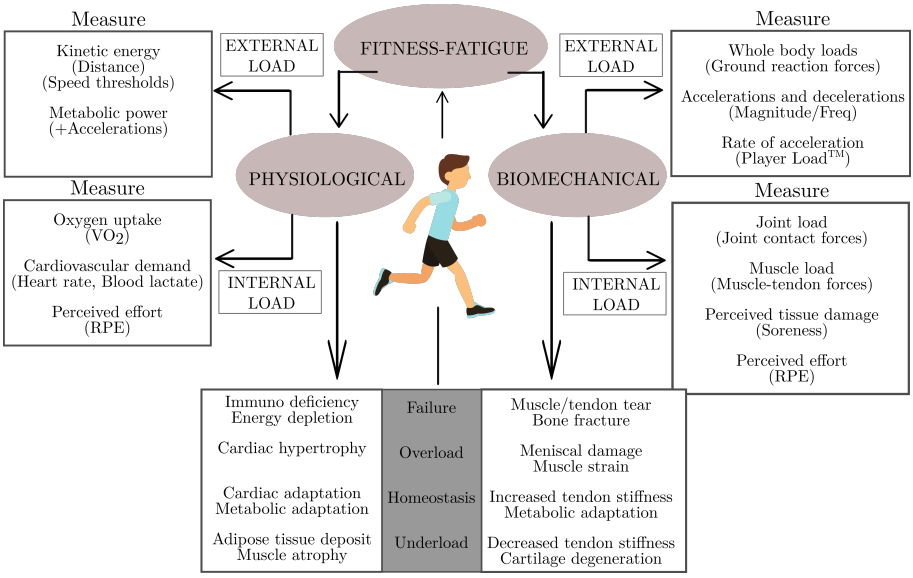


Figure 2.4: The model illustrates that both physiological and biomechanical loads lead to fitness and fatigue effects in athletes, redrawn from (Vanrenterghem et al. 2017).

variables that influence the magnitude of the structure-specific load (e.g. body weight, impacts). Third, he could track how the athlete distributes the load (e.g. scapular control while throwing). The practitioner can then examine the relationships between these variables and training load. The conceptual model in Figure 2.6 visualizes these interactions within a training session (Nielsen et al. 2018).

2.3 Machine Learning

This section provides an overview of several important machine learning concepts and introduces the machine learning algorithms that are used in the subsequent chapters.

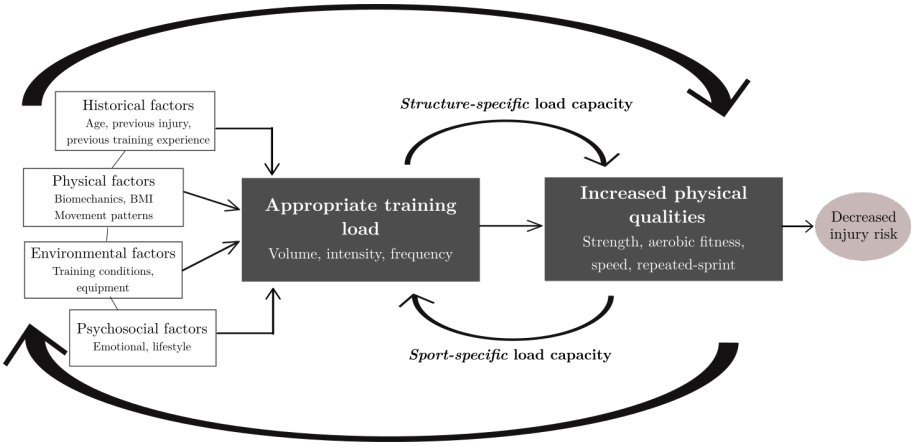


Figure 2.5: The circular causation between physical capabilities of an athlete and load tolerance motivates a gradual increase in training load, redrawn from (Gabbett, Nielsen, et al. 2018).

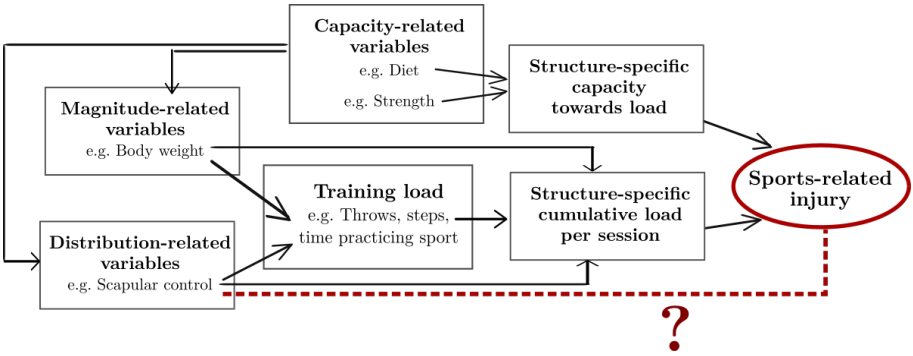


Figure 2.6: Within session relationships related to structure-specific load and load capacity, redrawn from (Nielsen et al. 2018).

2.3.1 Supervised Learning

The goal of a supervised machine learning algorithm is to learn a predictive model M. This model should mimic the behavior of a real-world system S. The system S itself, is a black box. Yet, a learning algorithm can learn more about S by observing its behavior. The learner can observe the output y_i that S produces, given a set of input values or features $x_{i_1}, x_{i_2}, x_{i_3}, ..., x_{i_n}$. Thus,

such an observation provides a learning example for the learner. After observing enough learning examples (i.e., the learning set), the learner should be able to learn a model M that mimics S. To learn M, the learner should be able to generalize from the provided learning examples to deal with unseen examples (i.e., the testing set) that are sampled from the same population. To perform this generalization step the learner needs to make assumptions about the behavior of M (i.e., the learner has a certain bias for choosing one generalization over another (T. Mitchell 1980)).

2.3.2 Model Evaluation

When the predictions of a model M are close to the observed outputs of S, M models the behavior of S well. A set of held-aside examples provide an estimation of M's ability to generalize to unseen data. This testing set is not consulted by the learning algorithm while learning.

In the context of continuous monitoring data there are several dependencies among the learning examples. First, there are time dependencies between trials (i.e., a data recording session of an athlete). Second, there can be multiple trials per athlete and per trial there can be multiple learning examples. These dependencies should be respected when selecting a learning set and testing set. Otherwise, information about the testing examples is available when learning the model. This would result in an unrealistic assessment of the performance of M.

Two popular evaluation methodologies exist:

Train-Test. This approach splits the set of all observations into a learning set and a testing set while respecting the dependencies that are present.

Cross-validation This strategy evaluates the performance of M using disjoint folds of the set of examples. Each fold selects one part of the examples for testing, the other part for learning. Two cross-validation approaches are in particular relevant for this thesis, because they allow to maximally use the data for both learning and model evaluation.

Leave-one-subject-out This approach creates one fold per subject. Each fold uses the examples of one subject for testing and the examples from all other subjects for learning.

Leave-one-trial-out The approach creates a fold per trial. Each fold selects the examples of one trial for testing and the examples from all other trials for learning.

2.3.3 Learning Algorithms

This section provides an overview of the different machine learning algorithms that are used in this thesis.

Least Absolute Shrinkage and Selection Operator (LASSO) LASSO is a more advanced linear regression technique (Tibshirani 1996). It employs a mechanism that forces the regression coefficients for the variables toward zero. Thus the model will consist of all variables that have non-zero coefficients. More technically, LASSO, minimizes the following loss function:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|,$$

where y denotes the output of the system, x, the input variables, β the regression coefficients, p the number of coefficients, and λ_1 the weight assigned to the L1-penalty. Thus, the technique has a preference towards associating coefficients with each variable subject to two competing factors:

1. the coefficients should model the data well, as measured by the squared error of the predictions on the learning data; and
2. it wants weights that are the coefficients need to be small, that is, close to zero.

The motivation behind (2) is that small coefficients should help the model to make better predictions on future data. Specifically, LASSO's main innovation is on (2) where it imposes what is known as an L1 penalty on the magnitude of the coefficients. This contrasts with the older Ridge Regression technique (Hoerl and Kennard 1970) which imposes what is called an L2 penalty ($\lambda_2 \sum_{j=1}^{p} \beta_j^2$) on the magnitude of the regression coefficients, which also has a preference for small coefficients, but is less likely to set coefficients to zero. The coefficients are determined via a mathematical optimization procedure that determines the optimal coefficients subject to the trade off between these two factors.

Elastic Net Elastic Net (Zou and Hastie 2005) combines the L1 penalty of LASSO with the L2 penalty of Ridge regression and minimizes the following loss function:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2,$$

where y denotes the output of the system, x, the input variables, β the regression coefficients, p the number of coefficients, λ_1 the weight assigned to the L1-penalty, and λ_2 the weight assigned to the L2-penalty.

The L1 penalty encourages the learner to set coefficients to zero. Yet, in case of a group of correlated variables, L1 will keep at most one of the variables of that group. The L2 penalty allows the learner to shrink coefficients of correlated variables together. Yet, L2 is unable to set the coefficients to zero and discard variables. By combining L1 and L2, Elastic Net leverages the strengths of both methods. Thus, Elastic Net can keep or remove entire groups of correlated variables.

Artificial Neural Networks (ANN) An artificial neural network (ANN) (Basheer and Hajmeer 2000) consists of a network of neurons (see Figure 2.7). Each neuron is a simple processing unit that combines and transforms its inputs into an output. The output of one neuron can serve as the input of the neurons of other layers. When the neurons use a non-linear activation function, connecting multiple neurons yield the ability to express complicated non-linear functions (Cybenko 1989).

To learn such a non-linear function, the ANN learning algorithm learns weights for the input of each neuron. As small changes to these weights can have a big impact on the output of the network, one needs efficient computational techniques (i.e., stochastic gradient descent and backpropagation) to optimize these weights. While ANNs often perform well, they have two main drawbacks. First, the networks are hard to interpret because it is non-trivial to identify which features are important and it quickly becomes complicated to understand how features are combined to come to a prediction. Second, the learner requires a high number of learning examples to learn an accurate model.

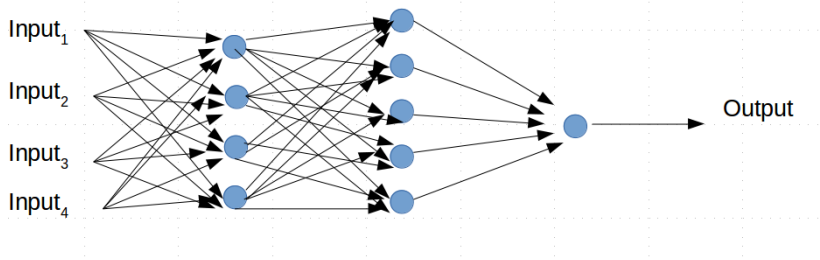


Figure 2.7: High level overview of a possible ANN network. Circles represent neurons. This network consists of one input layer, two hidden layers and one output layer.

Gradient Boosted Regression trees (GBRT) A GBRT model consists of multiple simple decision tree models that are learned sequentially using boosting (Friedman 2001).

Decision trees are learned using a top-down stepwise process. Each step selects the single best input variable according to some score criteria (e.g., mean squared error) and adds it to the model. Then, it partitions the data based on this variable's value, and recursively finds the best variable in each partition. This process helps with multi-collinearity because highly-correlated variables will have similar scores. Therefore, after adding one of these variables to the model, the others are unlikely to be included because they will not help to further partition the data. A drawback of decision trees is their high variation (i.e., small changes in the data result in a completely different model) (Hastie et al. 2009).

Combining a set of decision trees has been shown to be more robust to overfitting (i.e., the model fits the learning data too closely and fails to generalize to unseen data) compared to a single tree (Domingos 2000).

Boosting is a popular method to combine decision trees into a model. The resulting model can make a prediction by adding the prediction of each tree. Intuitively, boosting learns trees sequentially, such that each model can compensate for the errors made by the previous model.

More technically, in each iteration, the boosting method fits a decision tree to the "pseudo"-residuals (i.e., the negative gradient of the loss function being minimized) for each of the data points that are being selected in that iteration (Friedman 2002). Furthermore, each iteration only uses a random subset of the learning set as this has been shown to increase the model's performance (ibid.). This iterative process stops when a predefined number of trees are learned.

Chapter 3

Relationships Between the External and Internal Training Load in Professional Soccer: What can We Learn from Machine Learning?

Published as: Jaspers, A., Op De Beéck, T., Brink, M., Frencken, W., Staes, F., Davis, J., Helsen, W. (2018). Relationships between the external and internal training load in professional soccer: what can we learn from machine learning? International journal of sports physiology and performance, 13(5), 625-630.
A.J. and T.O.D.B share the first authorship;
J.D. and W.H. share the last authorship

Author contributions T.O.D.B, A.J, J.D, and W.H conceived and designed research;
A.J. collected the data;
A.J. and T.O.D.B preprocessed the data;
T.O.D.B performed the machine learning experiments;
T.O.D.B, A.J, J.D and W.H interpreted results of experiments;
A.J and T.O.D.B drafted manuscript;
A.J, T.O.D.B, M.B., W.F., F.S., J.D., and W.H. revised manuscript;

A.J, T.O.D.B, M.B., W.F., F.S., J.D., and W.H. approved final manuscript.

3.1 Abstract

Purpose: Machine learning may contribute to understanding the relationship between the external load and internal load in professional soccer. Therefore, the relationship between external load indicators and the rating of perceived exertion (RPE) was examined using machine learning techniques on a group and individual level.

Methods: Training data were collected from 38 professional soccer players over two seasons. The external load was measured using global positioning system technology and accelerometry. The internal load was obtained using the RPE. Predictive models were constructed using two machine learning techniques, artificial neural networks (ANNs) and least absolute shrinkage and selection operator (LASSO), and one naive baseline method. The predictions were based on a large set of external load indicators. Using each technique, one group model involving all players and one individual model for each player was constructed. These models' performance on predicting the reported RPE values for future training sessions was compared to the naive baseline's performance.

Results: Both the ANN and LASSO models outperformed the baseline. Additionally, the LASSO model made more accurate predictions for the RPE than the ANN model. Furthermore, decelerations were identified as important external load indicators. Regardless of the applied machine learning technique, the group models resulted in equivalent or better predictions for the reported RPE values than the individual models.

Conclusions: Machine learning techniques may have added value in predicting the RPE for future sessions to optimize training design and evaluation. Additionally, these techniques may be used in conjunction with expert knowledge to select key external load indicators for load monitoring.

3.2 Introduction

Nowadays, professional soccer clubs monitor training and match load to optimize physical fitness and reduce injury risk (Akenhead and Nassis 2016). When

considering training and match loads, it is typical to distinguish between the external and internal load (Impellizzeri, Rampinini, and Marcora 2005). The external load represents the dose performed and the internal load represents the psychophysiological stress experienced by the player (ibid.). The external load is generally defined as all locomotor and non-locomotor activities performed by players (Gabbett 2016; Impellizzeri, Rampinini, and Marcora 2005. Global positioning systems (GPS) and inertial sensors are used for monitoring external load indicators (ELIs) such as the distance covered and jumps (Gabbett 2016). The internal load can be quantified using the rating of perceived exertion (RPE), which is often considered a good indicator of the global internal load (Impellizzeri, Rampinini, Coutts, et al. 2004). Due to differences in individual characteristics (e.g., training history and actual physical fitness), similar external loads can result in different internal loads for players. Insights into the relationship between the external and internal load can improve load management and help to optimize physical fitness and support injury prevention (Drew et al. 2016).

To date, several studies about team sports have focused on the relationship between the external and internal load. In these studies, the data were analyzed using traditional statistical methods such as Pearson correlation coefficients, multiple regression and general linear models with partial correlation coefficients (Gaudino et al. 2015; T. Lovell et al. 2013; B. Scott et al. 2013). Recently, a study in Australian football (AFL) found that artificial neural networks (ANNs), a machine learning approach, more accurately predicted the RPE in response to ELIs compared to traditional statistics (Bartlett et al. 2017). Other machine learning techniques could be used for this task as well, and each technique has strengths and weaknesses (Bishop 2006).

In general, the data-driven approach of machine learning is able to capture linear and non-linear relationships between various ELIs and the response variable RPE (ibid.). Given a large set of ELIs, machine learning approaches can automatically identify the specific ELIs that are most predictive of the RPE, often without correcting for multicollinearity or using expert knowledge to hand select a set of ELIs. This can aid in evaluating newly developed external load metrics that come with improved tracking systems such as GPS technology and inertial movement sensors (J. Malone, R. Lovell, et al. 2017).

Another advantage of machine learning is its ability to detect possible inter-player differences. In the AFL study using machine learning techniques, various ELIs were examined to determine their predictive value for each player's RPE (Bartlett et al. 2017). Inter-player differences were found for ELIs and their contribution to an individual's RPE (ibid.). For most players, the distance covered was the most predictive ELI for the RPE. However, for some players, the distance covered per minute or distance covered at high-speed ($>14.4\ km/h$) had a higher predictive value, indicating that individual differences should be considered

when evaluating dose and response to training load (Bartlett et al. 2017).

Even though AFL and soccer are both running-based team sports, each sport imposes different physical demands on players due to differences in rules, pitch dimensions, player rotations versus substitutions, and playing time (Varley, Gabbett, et al. 2014). In comparison to soccer players, AFL players typically cover 2.6 times greater distance (1322 m versus 517 m) at very high-speed (19.8-25.1 km/h) and 3.5 times greater distance (328m versus 93m) at sprinting speed (>25.2 km/h) in matches. When comparing the absolute number of maximal acceleration efforts (>2.78 m/s^2) to the absolute number of high-speed efforts (19.8-25.1 km/h), AFL players show a 1:1 ratio whereas soccer players exhibit a ratio of 1.7, indicating that numerous accelerations during matches do not result in high-speed efforts. Based on this comparison, it may be unlikely that the results regarding the most predictive ELIs and inter-player differences in AFL will generalize to professional soccer. To our knowledge, no prior study in professional soccer has investigated the relationship between ELIs and RPE using machine learning techniques to determine which ELIs are most predictive of the RPE or to examine possible inter-player differences.

In summary, the current study evaluates the ability of machine learning techniques to:

1. Predict the RPE from a given set of ELIs;
2. Identify which ELIs for soccer players contribute most to the RPE;
3. Evaluate both group and individual models to examine possible inter-player differences regarding the relationship between ELIs and RPE.

3.3 Methods

3.3.1 Subjects

Data from 38 professional soccer players (22.7 ±3.4 years, 1.83 ±0.06 m, 77.0 ±6.7 kg, and 10.3 ±1.8% body fat) competing for a team in the highest league in the Netherlands were included. Goalkeepers' data were excluded from the study due to different physical demands. The study was conducted according to the requirements of the Declaration of Helsinki and was approved by the KU Leuven ethics committee (file number: s57732).

3.3.2 Design

Data were collected from pre-season and in-season training sessions over two seasons (2014-2015 and 2015-2016). Similar to Bartlett et al., (Bartlett et al. 2017) this study focused on the relationship between ELIs and RPE in training sessions. Therefore, data from matches, on-field recovery sessions, and rehabilitation sessions were excluded from the analysis. For each training session, the external load was measured using 10 Hz GPS and 100 Hz accelerometer technology (Optimeye S5, Catapult Sports, Melbourne) in accordance with the recommendations for collecting and processing GPS data in sports (J. Malone, R. Lovell, et al. 2017). The internal load was measured using the RPE. Each player reported his RPE approximately 30 minutes after the training session using the modified Borg CR-10 scale (Foster et al. 2001). All players were familiarized with the use of RPE before the beginning of the study and were instructed to rate their perceived effort for the whole training session (Impellizzeri, Rampinini, Coutts, et al. 2004). Furthermore, each player was asked in isolation for his RPE to minimize the influence of factors such as peer pressure (J. Malone, Di Michele, et al. 2015).

The first season contained data from 23 players. The number of sessions recorded per player ranges from 35 to 160 with a mean and standard deviation of 125 $\pm$34 sessions. The second season contained data from 28 players. The number of sessions recorded per player ranged from 51 to 163 with a mean and standard deviation of 109 $\pm$33 sessions. As players frequently switch teams in professional football, only 13 players appeared in both seasons.

3.3.3 Methodology

To examine the relationship between the external load and RPE using machine learning, a set of 67 ELIs that could be exported from the manufacturer's software (Sprint version 5.1.7, Catapult Sports, Melbourne, Australia) was selected to capture the external load of a training session. The set of ELIs can be divided into high-level categories about duration, distance, speed, acceleration and deceleration, PlayerLoad (i.e., a metric based on accelerometry), and repeated high-intensity effort (RHIE) activity (Table 3.1). The first goal was to identify the ELIs that are most predictive of the RPE. Therefore, a model was constructed that accurately predicts what a player's reported RPE (internal load) will be based on the observed value for all ELIs in a training session.

The mean absolute error (MAE) was used to assess a model's predictive performance. This metric calculates the mean of the absolute errors (i.e., |(reported RPE value) – (predicted RPE value)|) over all predictions. The

Table 3.1: Set of ELIs. Abbreviations: #, number of; ELI, external load indicator; RHIE, repeated high-intensity effort.

Category (# ELIs)	**Definition**
Duration (1 ELI)	This ELI defines the duration of the training session.
Distance (17 ELIs)	These ELIs capture the total distance covered, distances covered in different speed zones, and percentages of distances covered at different speeds. The different speed zones considered are: 0-1 km/h, 1-7 km/h, 7-12 km/h, 12-15 km/h, 15-20 km/h, 20-25 km/h, and >25 km/h.
Speed (8 ELIs)	This group contains ELIs that describe the distance covered per minute and the number of efforts in different speed zones.
Acceleration and deceleration (18 ELIs)	These ELIs capture the accelerations and decelerations, as well as the accelerating and decelerating distance. The ELIs regarding accelerating and decelerating efforts and distance are divided into different zones based upon magnitude (0-1 m/s^2, 1-2 m/s^2, 2-3.5 m/s^2 and >3.5 m/s^2).
PlayerLoad (10 ELIs)	This category consists of ELIs based on measures of PlayerLoad. PlayerLoad 3D is calculated based on the changes in accelerations of a player in the X, Y and Z axis. PlayerLoad per meter (i.e, PlayerLoad 3D per total distance covered) and the PlayerLoad per minute are included as well. Furthermore, it includes PlayerLoad 1D (i.e., PlayerLoad values per axis).
RHIE (13 ELIs)	An RHIE bout was defined as three or more sprints, high-magnitude accelerations or a combination of both within 21 seconds (Austin et al. 2011; Spencer et al. 2004). This category included measures based on RHIE such as RHIE bout recovery, RHIE duration, RHIE per bout, and RHIE total bouts.

MAE is easy to interpret as it uses the same unit as the RPE value: a MAE of 1 means that, on average, the predicted RPE is one value below or above the reported RPE. While a MAE of zero is unrealistic, the goal is to minimize a model's MAE.

To construct predictive models, two standard machine learning techniques were considered as well as one naive baseline method:

Artificial neural networks (ANN) ANNs are a standard approach for constructing non-linear models that often exhibit good predictive performance (Bishop 2006). However, a disadvantage of ANNs is that the resulting models are difficult to interpret (i.e., they do not provide insight into the interactions that are modelled among ELIs).

Least absolute shrinkage and selection operator (LASSO) This technique is an advanced version of linear regression (Tibshirani 1996). When setting the regression coefficients, LASSO contains a mechanism that biases many of them to be zero. Consequently, LASSO only selects a subset of the ELIs, those with a non-zero coefficient, to be included in the model. This results in both better interpretability and more robustness to multicollinearity among the input variables than traditional linear regression. As LASSO constructs a linear model, it is more robust to small sample sizes compared to the more expressive ANNs.

Additionally, a well-known LASSO-based approach can be used to compute importance scores of the ELIs (Meinshausen and Bühlmann 2010). The importance scores are calculated as the probability that an ELI is selected by the LASSO model and fall in the range of zero to one. Higher scores denote more important ELIs. In general, the presence of collinearity among the input ELIs tends to result in lower importance scores.

Baseline This model does not consider the external load and always predicts the average RPE value over all training sessions used to construct the model. This model assumes that none of the ELIs are predictive of the RPE. While a MAE of zero is a lower bound (i.e., a perfect predictive model), the baseline provides a realistic upper bound for the MAE. A valuable predictive model should have a lower MAE than this baseline.

3.4 Data Analysis

Two experiments were conducted. Each one employed standard machine learning methodology and subdivided the data into two disjoint sets: the learning set and

testing set. Each machine learning approach used the data in the learning set to construct a model. The independent testing set was used to estimate a model's predictive performance on unseen (that is, future) data. Specifically, each model made a prediction for the reported RPE associated with every training sessions in the testing set, and the MAE was computed for these predictions. In addition, 90% confidence intervals (CI) and effect sizes were calculated (Hopkins 2002; Hopkins et al. 2009).

The first experiment evaluated the value of group models. The temporal nature of the data was preserved by partitioning the data based on seasons: data from the first season served as the learning set and the data from the second season as the testing set. A consequence of the seasonal split was that each model made predictions for unseen players, that is, players who had no data in the learning set. One group model was constructed using each learning approach. The most predictive ELIs were identified by inspecting the most accurate learned model.

The second experiment examined the impact of accounting for inter-player differences. As only a few players appeared in both seasons, there was insufficient data to consider season-based partitioning of the data. Therefore, season 1 and season 2 were treated separately. Each season's data was subdivided temporally such that the first 75% served as the learning set and the last 25% served as the testing set. Using each learning approach, both one group model and an individual model for each player was constructed. The group model was constructed using data from all the players in the learning set. An individual model for each player was constructed by only considering that specific player's training session data in the learning set. A global mean of the absolute errors of all individual models was calculated so that the metric aligned with how the group model's MAE was computed.

For automated preprocessing and advanced analysis, custom Python scripts were developed using Python Pandas for data handling and Sklearn for machine learning (McKinney 2010; Pedregosa et al. 2011).

3.5 Results

The average RPE for all 5917 analyzed training sessions was 3.59 ±1.46 AU. The following descriptive statistics were calculated for these commonly reported ELIs: duration 70 ±16 minutes, total distance covered 4614 ±1576 m, distance covered at high-speed (>15 km/h) 426 ± 351 m, and total distance covered per minute 65 ±14 m.min-1.

Table 3.2 shows the MAEs and 90% CIs for the group models constructed using

the data from season 1 and evaluated on the data from season 2. In addition, the effect sizes are shown for the MAEs of ANN and LASSO group models compared to the baseline's MAE. Both the ANN and LASSO models outperform the baseline. Compared to the baseline, the LASSO model resulted in a 29.8% reduction in the MAE when predicting the RPE of unseen training sessions from season 2. Moreover, the LASSO model made more accurate predictions than the ANN model. A trivial effect size was found for ANNs compared to the baseline, while a small effect size was found for the LASSO group model compared to the baseline.

Table 3.2: Machine learning group models and baseline constructed on season 1 and evaluated on season 2: MAEs, 90% CIs, % diff vs LASSO, and effect sizes of MAEs vs baseline. Abbreviations: % diff, percentage difference; ANN, artificial neural networks; CI, confidence interval; d, standardized difference; LASSO, least absolute shrinkage and selection operator; MAE, mean absolute error; vs, versus.

Method	**Aggregation**	**MAE (90% CI)**	**% diff vs LASSO**	**d**	**Effect size**
ANN	Group	1.09 (1.07-1.11)	26.6	0.06	trivial
LASSO	Group	0.80 (0.78 - 0.82)		0.44	small
Baseline	Group	1.14 (1.12 - 1.16)	29.8		

Table 3.3 displays the ELIs, and their corresponding importance scores, selected by the LASSO group model (learned on the data from season 1) that most contribute to predicting the RPE.

Table 3.4 reports the MAEs and 90% CIs for individual and group models that were constructed and evaluated on season 1 and season 2 separately. Additionally, the effects sizes are presented for the comparison of the MAEs of ANN and LASSO models (i.e., both individual and group models) with the baseline. In all eight cases, the learned models had a lower MAE score than the baseline. Regardless of learning method, the group models resulted in equivalent or even more accurate predictions of the reported RPE values than the individual models.

3.6 Discussion

This study aimed to evaluate the ability of machine learning techniques to predict the RPE of soccer training sessions from a set of ELIs. Additionally, it aimed

Table 3.3: Overview of ELIs and importance score selected by the LASSO group model. Abbreviations: ELI, external load indicator; LASSO, least absolute shrinkage and selection operator; RHIE, repeated high-intensity effort.

ELI	**Importance score**	**Definition**
Acceleration zone 4 efforts	0.515	Number of acceleration efforts above 3.5 m/s^2
RHIE per bout - mean	0.513	Average of repeated high-intensity efforts per bout of 21 seconds
Deceleration zone 3 distance	0.510	Decelerating distance between -3.5 and -2 m/s^2
Velocity zone 5 distance	0.507	Distance covered between 15-20 km/h
Acceleration zone 3 efforts	0.507	Number of acceleration efforts between 2 and 3.5 m/s^2
PlayerLoad	0.487	Accumulated PlayerLoad measured by accelerometry
Velocity zone 4 distance	0.487	Distance covered between 12-15 km/h
Minutes	0.471	Training duration
Deceleration zone 4 distance	0.466	Decelerating distance below -3.5 m/s^2
PlayerLoad 1D side	0.458	Accumulated PlayerLoad for sideways movements (or medio-lateral axis) measured by accelerometry
Velocity zone 6 efforts	0.428	Efforts between 20-25 km/h
PlayerLoad 2D	0.384	Accumulated PlayerLoad with exclusion of up- and downwards movements (or longitudinal axis) measured by accelerometry

to identify the ELIs which are most predictive of RPE within a professional soccer context. Finally, it attempted to explore inter-player differences for how ELIs contribute to each player's RPE.

The constructed ANN and LASSO models outperformed the baseline indicating that it is possible to construct machine learning models that capture a part of the relationship between ELIs and RPE in professional soccer. Additionally, it

Table 3.4: Machine learning models and baseline for season 1 and season 2: MAEs, 90% CIs, % diff vs LASSO, and effect sizes of MAEs vs baseline. Abbreviations: % diff, percentage difference; ANN, artificial neural networks; CI, confidence interval; d, standardized difference; LASSO, least absolute shrinkage and selection operator; MAE, mean absolute error; vs, versus.

Season	Method	Aggregation	MAE (90% CI)	% diff vs LASSO	d	Effect size
1	**ANN**	Individual	0.84 (0.82 - 0.86)	3.6	0.21	small
		Group	0.81 (0.79-0.83)	2.5	0.26	Small
	LASSO	Individual	0.81 (0.76 - 0.86)		0.26	Small
		Group	0.79 (0.75 - 0.83)		0.30	Small
	Baseline	Group	0.99 (0.94 - 1.04)	20.2		
2	**ANN**	Individual	0.85 (0.83 -0.87)	0	0.33	Small
		Group	0.83 (0.81 - 0.85)	-2.4	0.34	Small
	LASSO	Individual	0.85 (0.80 - 0.90)		0.33	Small
		Group	0.85 (0.80 - 0.90)		0.33	Small
	Baseline	Group	1.11 (1.05 - 1.17)	23.4		

suggests that a good strategy is to start with a large set of ELIs, as opposed to hand selecting a small number of ELIs to reduce the chance of discarding a relevant ELI. Moreover, a strength of machine learning techniques is their ability to automatically select a subset of predictive ELIs, often without correcting for multicollinearity. Therefore, this method may provide new insights and support expert knowledge in the selection of key load indicators for monitoring strategies.

The LASSO technique identified various ELIs as contributing the most to the perceived exertion in professional soccer (Table 3.3). These ELIs are partly in line with earlier findings in professional soccer using a smaller set of ELIs (Gaudino et al. 2015; B. Scott et al. 2013). However, as GPS devices from

different manufacturers are used in the other studies, it is difficult to compare findings (J. Malone, R. Lovell, et al. 2017).

The novel important ELIs are indicators regarding decelerations. The results of this study indicate that this type of load, next to other ELIs, may contribute to a player's RPE. Previously, mainly concentric, energy-demanding efforts were associated with higher RPE values in professional soccer (Gaudino et al. 2015; B. Scott et al. 2013). Decelerating efforts are related to eccentric activity (Nédélec et al. 2012). This type of muscle activity has a lower energy cost in comparison with concentric muscle activity (Lindstedt et al. 2001). However, this type of eccentric contractions might more easily induce muscle damage (Lindstedt et al. 2001; Nédélec et al. 2012). Therefore, monitoring ELIs concerning decelerations can be particularly important.

Both individual and group models captured part of the relationship between ELIs and RPE. In contrast to Bartlett et al., (Bartlett et al. 2017) we found that group models using ANN and LASSO techniques demonstrate an equivalent or superior accuracy for both season 1 and 2 compared to individual models when predicting RPE based on ELIs. A combination of diverse underlying factors may explain these results.

First, these findings are in contrast to the theoretical model of Impellizzeri et al., which states that the internal load (RPE) results from the interaction between the external load (ELIs) and individual characteristics (Impellizzeri, Rampinini, and Marcora 2005). The results of our study may indicate that there is less variation in the external loads and individual characteristics of professional soccer players than in AFL so there is less impact on the reported RPE. It is possible that there are greater differences in positional activity profiles and in individual characteristics (e.g., body composition and aerobic capacity) in AFL compared to professional soccer, which result in a more heterogeneous group in AFL (Coutts et al. 2015; Varley, Gabbett, et al. 2014). The descriptive statistics for the ELIs and RPE clearly exhibit lower average values and less variation for professional soccer training sessions compared to AFL training sessions (Bartlett et al. 2017). These inter-sport differences may partly explain the results indicating the presence of other ELIs that mutually determine the RPE for (most of) the players within a professional soccer team.

On the other hand, the sample size (i.e., the number of data points used to construct the model) is another factor which may have contributed to the equivalent performance of the group models. The group models are learned using a much larger sample size of more than 2000 data points compared to the individual models which typically relied on less than 100 data points. Nonetheless, we find that individual models constructed with the LASSO method perform similarly to the group models as the technique is robust to small sample

sizes. If more data were available for each player, we would expect the individual models' performance to improve. However, from a practical perspective this does not seem realistic. In professional soccer, only 100-150 training sessions (i.e., data points) are conducted per season per player. Additionally, players are often transferred which makes it difficult to obtain data over multiple seasons.

The current study focused on the relationship between ELIs and RPE for training sessions and matches were thus excluded. In future research, the same method could be applied to examine if similar ELIs influence the RPE for matches, or if different ELIs determine the RPE values of matches. However, as mentioned, machine learning requires sufficient amounts of data to build accurate predictive models. This could be a limitation due to the relative small number of games in a season. Additionally, the RPE for matches may be influenced by contextual factors (Brito et al. 2016).

Recently, the differential RPE (dRPE) has demonstrated its added value by quantifying respiratory and muscular perceived exertion (Los Arcos et al. 2014; McLaren et al. 2017; Weston et al. 2015). Using the dRPE may further clarify if specific ELIs have a higher impact on central (i.e., breathlessness) or local (i.e., leg muscle exertion) perceived exertion. These insights can aid in optimizing load and adaptation in terms of physiological (i.e., cardiorespiratory system) and biomechanical (i.e., musculoskeletal system) pathways (Vanrenterghem et al. 2017).

Additionally, measures of recovery and psychosocial factors were not considered. Therefore, the inclusion of measures such as pre-training perceived wellness and recovery may further clarify the RPE outcome for a given external training load (Gallo et al. 2016; Saw, Main, et al. 2016).

The identification of key ELIs may aid in the evaluation of players' training dose and response over time using efficiency ratios (i.e., the proportion between RPE and ELIs) (Akubat et al. 2014; Buchheit, Cholley, et al. 2016). For example, some ELIs may be perceived as less exerting at the end of pre-season or a rehabilitation process compared to the beginning due to improvements in physical fitness. Consequently, a consistent deviation between the expected and reported RPE may be used as an efficiency ratio. This ratio could be used to exhibit if players evolve over time in their ability to deal with the external load. However, further research is needed regarding efficiency ratios relating to changes in fitness or fatigue.

3.7 Practical Applications

Machine learning techniques may have added value in predicting the RPE for future training sessions and in selecting key ELIs for load monitoring in professional soccer. This study identified novel ELIs that should be considered such as high-magnitude decelerations that contribute to the RPE.

In addition, group models may have an added value in predicting the RPE for individual players: they can be applied to any player whereas an individual model is only applicable to that specific player. Hence, group models can make predictions for newly transferred or youth players, for whom there is often little (or no) available data. From a monitoring perspective, a dashboard for player monitoring may initially be made with similar ELIs for the players within a team. In case more data is available for a specific player, an individual model can be constructed and a customized dashboard can be monitored.

3.8 Conclusion

Our study confirmed that machine learning techniques are able to predict RPE based on a large set of ELIs collected during two seasons in professional soccer. Secondly, these techniques can be applied to support expert knowledge for the selection of key ELIs such as decelerations and, accordingly, improve load management strategies. Lastly, group models predicted the RPE with an equivalent or even better accuracy than individual models. Possible limitations of the applied machine learning approaches were discussed. In addition, guidelines for future machine learning research and practical applications were provided.

Chapter 4

Predicting Future Perceived Wellness in Professional Soccer: the Role of Preceding Load and Wellness

Published as: Op De Beéck, T., Jaspers, A., Brink, M. S., Frencken, W. G., Staes, F., Davis, J. J., Helsen, W. F. (2019). Predicting Future Perceived Wellness in Professional Soccer: The Role of Preceding Load and Wellness. International journal of sports physiology and performance, 1-25.
A.J. and T.O.D.B share the first authorship;
J.D. and W.H. share the last authorship

Author contributions T.O.D.B, A.J, J.D and W.H conceived and designed research;
A.J. collected the data;
A.J. and T.O.D.B preprocessed the data;
T.O.D.B performed the machine learning experiments;
T.O.D.B, A.J, J.D and W.H interpreted results of experiments;
A.J and T.O.D.B drafted manuscript;
A.J, T.O.D.B, M.B., W.F., F.S., J.D., and W.H. revised manuscript;
A.J, T.O.D.B, M.B., W.F., F.S., J.D., and W.H. approved final manuscript.

4.1 Abstract

Purpose: The influence of preceding load and perceived wellness on the future perceived wellness of professional soccer players is unexamined. This chapter simultaneously evaluates the external and internal load for different time frames in combination with pre-session wellness to predict future perceived wellness using machine learning techniques.

Methods: Training and match data were collected from a professional soccer team. The external load was measured using global positioning system technology and accelerometry. The internal load was obtained using the RPE multiplied by duration. Predictive models were constructed using gradient boosted regression trees (GBRT) and one naive baseline method. The individual predictions of future wellness items (i.e., fatigue, sleep quality, general muscle soreness, stress levels, and mood) were based on a set of external and internal load indicators in combination with pre-session wellness. The external and internal load was computed for acute and cumulative time frames. The GBRT model's performance on predicting the reported future wellness was compared to the naive baseline's performance by means of absolute prediction error and effect size.

Results: The GBRT model outperformed the baseline for the wellness items fatigue, general muscle soreness, stress levels and mood. Additionally, only the combination of external load, internal load, and pre-session perceived wellness resulted in non-trivial effects for predicting future wellness. Including the cumulative load did not improve the predictive performances.

Conclusions: The findings may indicate the importance of including both acute load and pre-session perceived wellness in a broad monitoring approach in professional soccer.

4.2 Introduction

Monitoring team-sport athletes is considered important for understanding responses to training and match load, and accordingly, for optimizing loads to ensure competition readiness (Halson 2014). Consequently, various player tracking tools are employed to continuously monitor training and match load (Bourdon et al. 2017). Furthermore, these loads elicit responses, such as fitness, fatigue and a certain need for recovery (Bourdon et al. 2017; Buchheit, Racinais, et al. 2013). These athletes' responses are often measured by perceived

wellness questionnaires (Bourdon et al. 2017; Buchheit, Racinais, et al. 2013). In professional soccer, several studies have provided evidence for using perceived wellness questionnaires to quantify the outcome of a training or match load by assessing players' fatigue statuses (Buchheit, Cholley, et al. 2016; Fessi, Nouira, et al. 2016; Moalla et al. 2016; Thorpe et al. 2015, 2017). It is assumed that changes in perceived wellness influence both on-field performance and injury risk (Laux et al. 2015; Saw, Main, et al. 2016).

Two studies have evaluated the external load in relation to changes in perceived player wellness, and both focused on the distance covered at high speed (HSR; $>14.4\ km/h$) (Thorpe et al. 2015, 2017). Other external load indicators such as total distance, distance covered at very high speed (VHSR; $>20.0\ km/h$), accelerations, and decelerations remain unexamined. Most studies examining the relationship between load and perceived wellness use the session rating of perceived exertion (sRPE), (Fessi, Nouira, et al. 2016; Moalla et al. 2016) which is derived by multiplying the RPE by duration, and is considered a global measure of the internal load (Impellizzeri, Rampinini, Coutts, et al. 2004).

To date, perceived wellness studies in professional soccer have focused on either external or internal load indicators. A simultaneous evaluation of external and internal load indicators has not been conducted yet. Thus, a combined approach that simultaneously evaluates different load indicators and their relationship with perceived wellness can help identify relevant load indicators. This may improve load management strategies for optimizing perceived player wellness in professional soccer.

Similarly, the impact of loads accumulated over several days on perceived wellness needs further exploration. One study in professional soccer focused on the cumulative external load as measured by HSR over the previous 2, 3, and 4 days (Thorpe et al. 2017). However, considering the cumulative load did not improve the strength of the relationship between HSR and changes in perceived player wellness (ibid.). Still, evaluating load indicators beyond HSR over different time periods has not been conducted and could help better understand the influence of cumulative loads on perceived wellness.

Recently, research in Australian rules football, (Gallo et al. 2016) American college football, (Govus et al. 2018) and professional soccer (S. Malone et al. 2018) has provided evidence that perceived pre-training wellness influences the subsequent training output. In view of the model of Impellizzeri and colleagues, (Impellizzeri, Rampinini, and Marcora 2005) the pre-training wellness status may be considered as an individual characteristic that impacts the performed external load but also the main stimulus for the training outcome, the perceived internal load. Following the rationale of the training process model, one can argue that pre-training wellness may also influence

the outcome of training or match load. Consequently, it is possible that pre-training wellness, in addition to training and match load, may influence future perceived wellness (Impellizzeri, Rampinini, and Marcora 2005). However, to our knowledge, the influence of pre-training wellness on future perceived wellness remains unexplored.

Finally, the relationships between load and perceived wellness can be examined for both each individual wellness item on the questionnaire, (Buchheit, Cholley, et al. 2016; Buchheit, Racinais, et al. 2013; Fessi, Nouira, et al. 2016; Moalla et al. 2016; Thorpe et al. 2015, 2017) and a global wellness measure computed as the summed score over all items (Buchheit, Racinais, et al. 2013; Moalla et al. 2016). One limitation of a global wellness measure is the limited capability to identify specific relationships between load indicators and wellness items (Gallo et al. 2016; Govus et al. 2018). Relationships between load indicators and various perceived wellness items have been examined for different season periods in professional soccer. However, except for a frequently observed relationship between higher loads and an increased perceived fatigue, the relationships between load and other wellness items such as sleep quality and general muscle soreness are less clear (Moalla et al. 2016; Thorpe et al. 2015, 2017). Furthermore, the relationships between diverse load indicators and wellness items have not been investigated over the course of a full season. Therefore, an explorative examination of relationships between load and wellness items over a longer period can provide additional insights into typical load-wellness response profiles for each wellness item over a season.

It is generally recognized that the relationship between load and perceived wellness may be non-linear (Gallo et al. 2016; S. Malone et al. 2018). Therefore, linear statistical techniques used in earlier research may be incapable of elucidating these relationships. Non-linear statistical models or machine learning techniques may provide additional insights in relationships between load and training outcomes. Machine learning (ML) techniques are suited for these analyses and corresponding data because they often account for multicollinearity and can model non-linear relationships among large sets of variables (Bishop 2006).

This study will apply ML techniques to construct individual predictive models for professional soccer players to:

1. Examine simultaneously the relationship between external (EL) and internal load (IL) indicators on future perceived wellness (FPW) items as measured on the next day;

2. Investigate the impact of both acute and cumulative loads on FPW items;

3. Evaluate the influence of pre-session perceived wellness (PPW) on FPW items.

4.3 Methods

4.3.1 Subjects

Data from 26 professional male soccer players (mean ± SD age: 23.2±3.7 years, weight: 77.5±7.4 kg, height: 1.82±0.06 m, body fat: 10.4±1.9%) competing for the same team at the highest level in the Netherlands were collected during the 2015-2016 season, both pre-season and in-season. Written informed consent was obtained according to the Helsinki declaration. The study was approved by the ethical committee of KU Leuven (file number: s57732).

4.3.2 Training and Match Load

External load was measured individually during all field training sessions and matches throughout the season. Data were obtained using an athlete tracking system with an integrated 10 Hz global positioning system (GPS) and accelerometer technology (Optimeye S5, Catapult Sports, Melbourne, Australia). This system is considered a reliable tool for measuring external load that obtains an acceptable level of accuracy for quantifying various locomotor activities (M. Scott et al. 2016). The minimum effort duration to detect velocity was 0.6 seconds, and 0.4 seconds for acceleration with a smoothing filter of 0.2 seconds (J. Malone, R. Lovell, et al. 2017; Varley, Jaspers, et al. 2017). The data were processed using the manufacturer's software (SprintTM version 5.1.7, Catapult Sports, Melbourne, Australia). Based upon earlier research, (Barrett et al. 2014; Jaspers, Kuyvenhoven, et al. 2018) the included external load indicators were training and match duration, total distance covered, PlayerLoad, distance covered at high speed (>20 km/h), the number of acceleration efforts >1 m/s^2 and deceleration efforts <-1 m/s^2.

The internal load was obtained for all players after the training sessions and matches using the sRPE method (Impellizzeri, Rampinini, Coutts, et al. 2004). In order to ensure that the perceived effort would reflect the session in total, rather than the most recent exercise intensity, each player was separately asked 30 minutes after every training session or match to rate his perceived exertion using a category ratio scale of 0-10 with verbal anchors (with 0 rated as 'rest', 1 rated as 'very, very easy' and 10 rated as 'maximal') (Foster et al. 2001). All players were familiarized with the scale before the study commenced. Each

player's sRPE in arbitrary units (AU) was derived by multiplying the RPE with the training or match duration in minutes (Foster et al. 2001). The entire duration of a training session was used including the transition time between drills. For matches, the sum of the warm-up and match time was used. The time between the warming-up and the start of the match as well as the half time break were excluded.

4.3.3 Perceived Player Wellness Questionnaire

The perceived player wellness data were individually collected using a custom-designed iPad-based electronic survey (TopSportsLab^{TM}, Leuven, Belgium) each morning prior to any session. Players were not asked to report wellness scores on match and rest days. The survey contained five questions about fatigue, sleep quality, general muscle soreness, stress levels, and mood that were used in earlier research (Buchheit, Cholley, et al. 2016; Buchheit, Racinais, et al. 2013). The responses were reported on a 5-point scale (with 1 and 5 representing poor and very good ratings), with 0.5-point increments (Buchheit, Racinais, et al. 2013). The players were familiarized with the questionnaire before the start of the study.

4.3.4 Data Analysis

This study applied a widely used machine learning pipeline to construct individual predictive models for each player (Bishop 2006). An individual model was constructed by ignoring the data from all other players. The goal was to predict a training session's outcome, which was represented by the future value of a perceived wellness (FPW) item. Specifically, the models predicted what perceived wellness score a player would report for an item prior to the next day's first session. Combinations of three sets of input variables were considered: external load indicators (EL), internal load indicators (IL), and pre-session perceived wellness (PPW) items.

Figure 4.1 illustrates the input variables that were computed to predict the FPW prior to the first session on day D_{FPW}. Based upon earlier research, the EL and IL variables of training sessions and matches were summed over four different time frames: 1 day (acute), 2 days, 3 days and 4 days (Thorpe et al. 2017). Additionally, because the weekly load is often related to an increased injury risk, the EL and IL variables were summed over the previous 7 days (Gabbett 2016). The PPW was defined as the pre-session perceived player wellness that was reported before the first session on day D_{FPW}-1 (i.e., a time frame of 1 day).

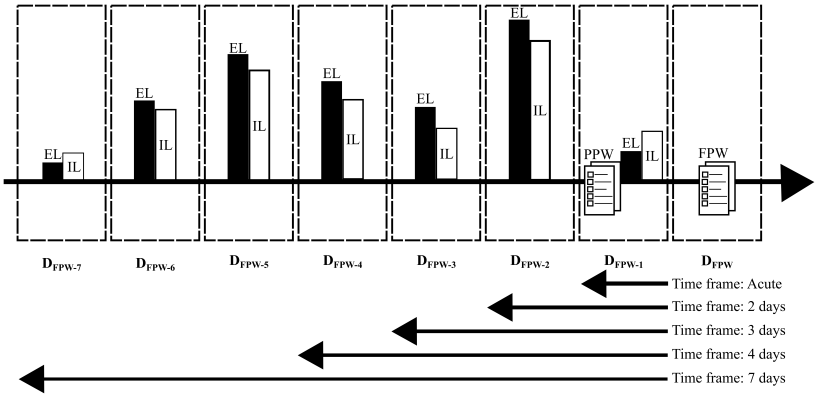


Figure 4.1: Overview of the parameters that are computed to predict future perceived wellness.

The data was split chronologically to respect its sequential nature: the first 80% of a player's data was used to construct the model (i.e., the learning set). The remaining 20% was used for model evaluation (i.e., the testing set). For each of the five time frames, seven combinations of variable classes were considered: EL, PPW, IL, EL + PPW, IL + PPW, EL + IL, and EL + IL + PPW. For each of the five FPW items (fatigue, sleep quality, general muscle soreness, stress levels, and mood), one model per player was learned for each of the 35 input variable time frame combinations.

The individual predictive models were constructed from the learning set using the Gradient Boosted Regression Tree (GBRT) algorithm in Scikit Learn (Friedman 2001; Pedregosa et al. 2011). GBRTs can handle both high-dimensional data and mixed variable types. A GBRT model contains a number of decision trees. Decision trees are learned using a top-down stepwise process. Each step selects the single best input variable according to some score criteria and adds it to the model. Then, it partitions the data based on this variable's value, and recursively finds the best variable in each partition. This process helps with multi-collinearity because highly-correlated variables will have similar scores. Therefore, after adding one of these variables to the model, the others are unlikely to be included because they will not help to further partition the data. Additionally, ensembles of decision trees tend to be robust to overfitting (Domingos 2000).

To assess if the learned individual models captured any dependencies between the input variables and the FPW, a naive baseline model was constructed that ignores all input variables. This model simply predicted a player's FPW as the

average of all FPW values in his learning set. A learned model only outperforms this baseline if it captures some relationship between the input variables and the FPW.

An individual model's predictive performance was evaluated by making a prediction for each of the player's reported wellness scores in the testing set and then computing the mean absolute error (MAE) for these predictions. The predictive performance for a given set of input variables was computed as the macro average of all the MAEs for the individual models that were constructed using that set of input variables. Per wellness item, and for each combination of input parameters and time frames, two comparisons were done. First, the macro MAE of the GBRT models was compared to the macro MAE of the baseline models. Second, the effect sizes between the macro MAE of the GBRT models and the macro MAE of the baseline models were calculated to evaluate the meaningfulness of the predictive performances using Cohen's d: $d = (macro\ MAE_{BASELINE} - macro\ MAE_{GBRT})/(pooled\ SD_{BASELINE,GBRT})$. The threshold values for effect sizes were trivial (0.0-0.19); small (0.2-0.59); moderate (0.6-1.19); large (1.2-1.99); and very large (>2.0) (Hopkins 2002).

Initially, the dataset contained data collected from 6110 training sessions or matches across all 26 players. Before the above methodology was applied to the dataset, four preprocessing steps were required, as illustrated in Figure 2.

First, perceived wellness scores were not reported on most rest and match days. Consequently, these days FPW value was unknown. Hence, these days were excluded from the learning and testing set. However, the EL and IL variables were monitored on these days and were used to calculate the cumulative external and internal loads.

Second, sometimes it was not possible to calculate the 7-day cumulative load for EL or IL due to missing EL and IL data (e.g., the first week after the off-season, international qualifiers, etc.). While these instances did not occur at random, they were excluded because the missing loads could not be realistically imputed.

Third, even if the FPW was known, the PPW was missing sometimes. The PPW was imputed using the last observation carried forward method, and hence set to be the reported perceived wellness score on day D_{FPW}-2 (Engels and Diehr 2003). If no scores were reported on D_{FPW}-2, then the session was excluded. While a match or training session on D_{FPW}-2 affects the perceived wellness of the player on D_{FPW}-1, this is a common imputation approach for temporal data because it respects the chronological dependencies present in the data. This necessary imputation step should be taken into account when analyzing the results. Other popular imputation strategies were also considered. However, because the data was not missing at random and its chronological

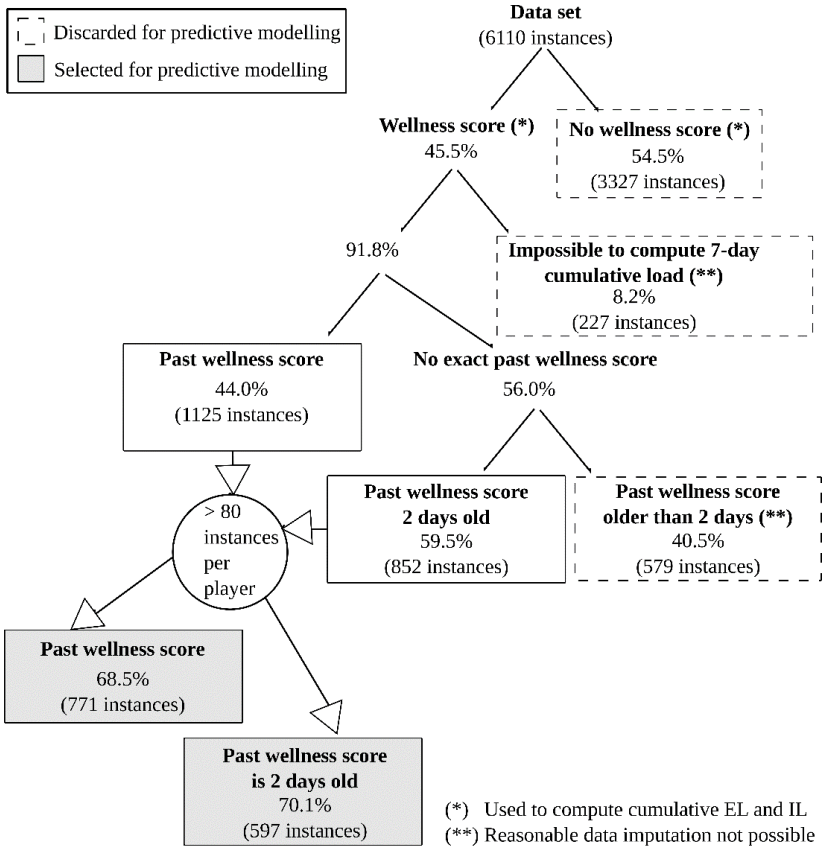


Figure 4.2: Overview of the preprocessing steps before application of GBRT.

dependencies need to be respected, not enough data instances were available to apply potentially more accurate imputation strategies.

Fourth, models were only learned for players where 80 data instances could be constructed to ensure that sufficient data was available for learning and evaluating the models. After preprocessing, the final dataset contained data from 14 players with an average of 98 data instances per player (range 84-119). On average each player's learning data contained 78 data instances (range 67-95) and testing data contained 20 data instances (range 17-24).

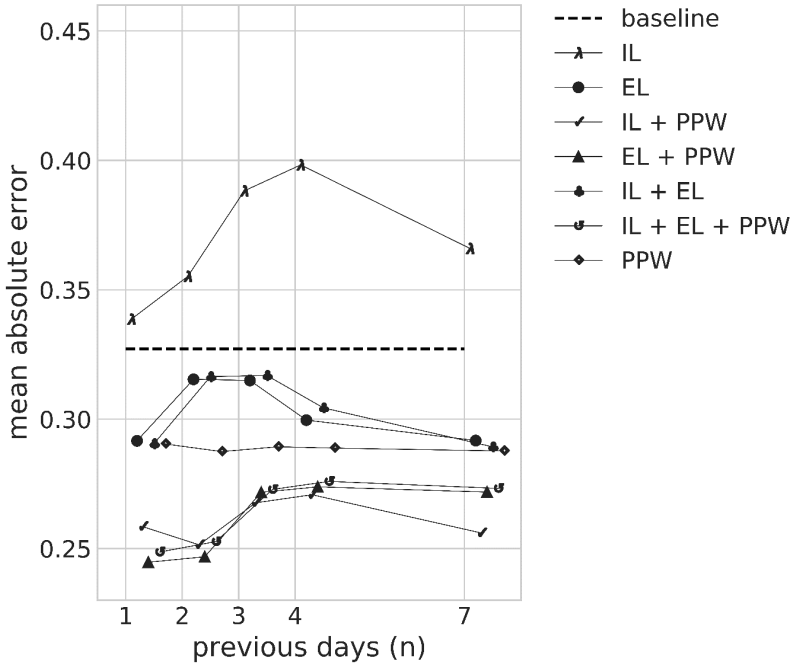


Figure 4.3: Mean absolute errors for each of the combinations per time frame for perceived wellness item "fatigue".

4.4 Results

Figures 4.3, 4.4, 4.5, and 4.6 shows graphs for the four wellness items (fatigue, general muscle soreness, stress levels and mood) with at least one small effect size found for one of the five considered time frames. Because only trivial effect sizes were found for sleep quality, no plot is shown for it. A small effect size indicates that the GBRT model obtained better predictive performance than the baseline model. For each wellness item, the plot shows the MAEs for each of the seven combinations of EL, PPW and IL as a function of the time frame. A decrease in the MAE over time indicates a better predictive performance when including the cumulative load over the previous days.

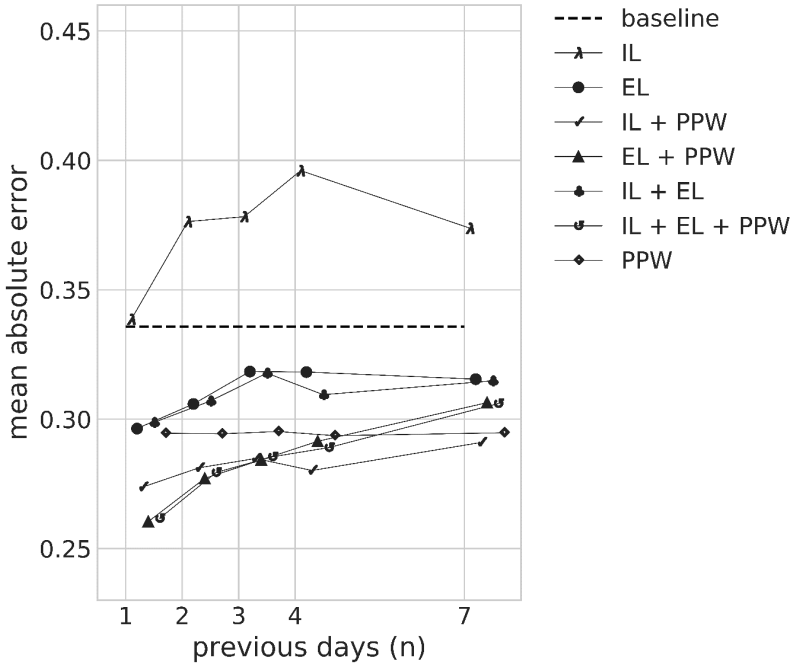


Figure 4.4: Mean absolute errors for each of the combinations per time frame for perceived wellness item "general muscle soreness".

4.5 Discussion

This study applied machine learning techniques to evaluate the influence of external and internal load indicators, both for acute and cumulative loads, along with pre-session perceived wellness on changes in future perceived wellness. When comparing EL and IL by absolute prediction error, EL exhibited a better performance for fatigue, general muscle soreness and stress levels. In general, the combination of EL and IL did not result in better predictive performances than EL alone.

Moreover, none of the predictive performances for EL, IL or EL+IL exhibited effect sizes above the trivial level. These effect sizes indicate that the external load and internal load, separately and in combination, do not have sufficient predictive ability for FPW items. However, in earlier research, these external and internal load indicators were related to changes in perceived wellness items and revealed various results, including non-significant and significant correlations

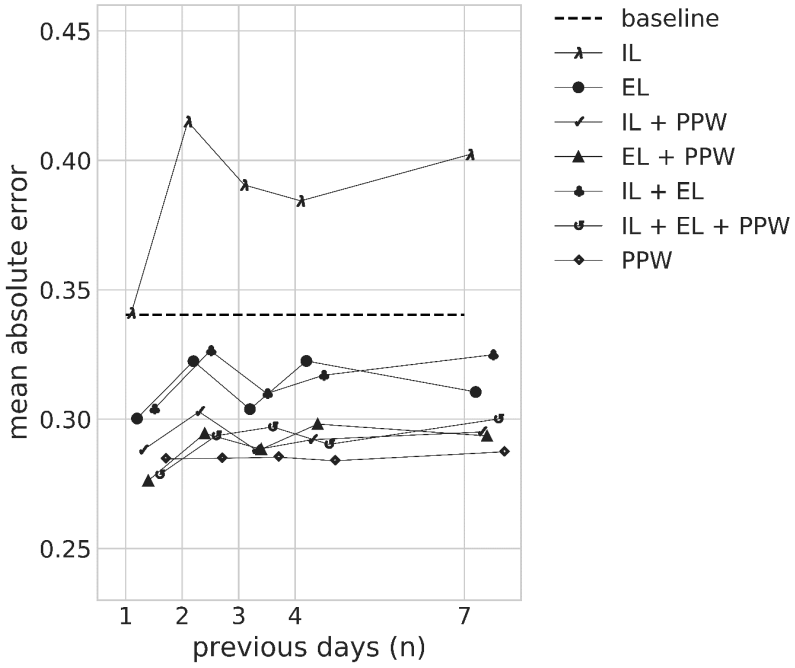


Figure 4.5: Mean absolute errors for each of the combinations per time frame for perceived wellness item "stress levels".

with the magnitude of correlation ranging from trivial to large (Fessi, Nouira, et al. 2016; Moalla et al. 2016; Thorpe et al. 2015, 2017). The difference with earlier findings could arise from the type of analysis performed. Prior work used analyses to quantify the strength of the linear associations among variables. In contrast, our study uses predictive models, that given EL and IL data collected at some future time point would make accurate predictions for that data's FPW values. Therefore, the current study's findings complement the earlier works.

Cumulative loads alone did not result in better predictive performances, which is in accordance with earlier findings that loads beyond the previous day's training are not meaningfully linked to wellness responses (Thorpe et al. 2017). As Thorpe and colleagues suggest, (ibid.) professional soccer's periodization of training and match load with an alternation between demanding sessions and easy or recovery sessions, may be responsible for the large influence of the previous day's training or match load.

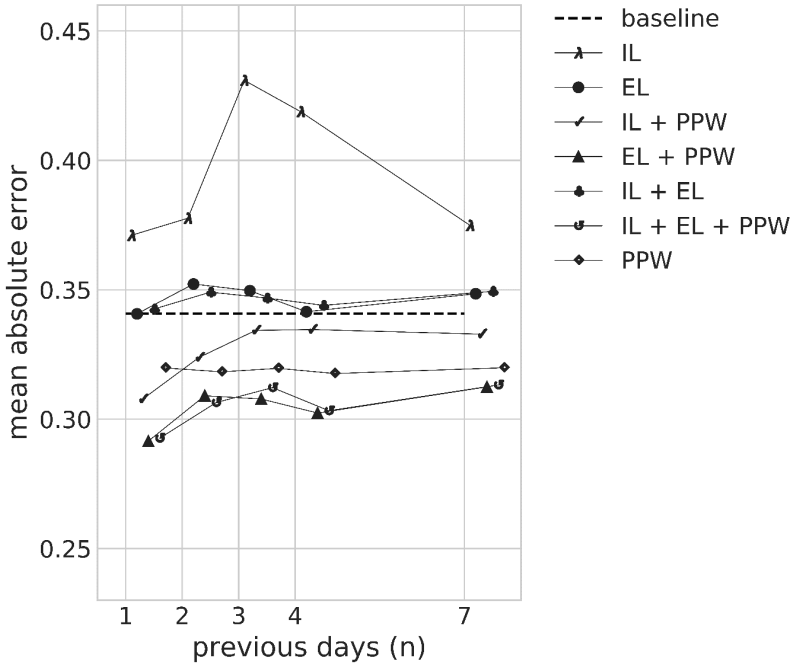


Figure 4.6: Mean absolute errors for each of the combinations per time frame for perceived wellness item "mood".

Including PPW in combination with EL, IL and EL+IL clearly showed small effect sizes for most time frames for fatigue, general muscle soreness, and stress levels. For mood, the results were more ambiguous and only the combination of acute load for EL and PPW and EL+PPW+IL resulted in a small effect size. To date, no research in professional soccer has focused on the relationship between load and mood, therefore, little information is available to compare results. Additionally, other factors such as match result, match location and quality of opposition may influence mood (Abbott et al. 2018). Potentially, mood is influenced after prolonged overload and therefore it might be interesting to study periods longer than 7 days. In conclusion, the findings reveal that PPW along with EL and/or IL resulted in the best predictive performances for FPW, thereby indicating the usefulness of monitoring perceived wellness. Therefore, PPW in combination with training and match load may be considered for a broad monitoring approach to improve training prescription and evaluation.

The perceived wellness items fatigue, general muscle soreness and stress levels

were predicted by the input variables. For the perceived wellness items sleep quality and mood, almost all predictive performances exhibited trivial effect sizes. Some studies found small to large positive correlations between sRPE and sleep quality, (Fessi, Nouira, et al. 2016; Moalla et al. 2016) while other studies revealed trivial relationships between HSR and sleep quality (Thorpe et al. 2015, 2017). This may indicate that factors beyond load and PPW have a greater impact on these items. Recent research in professional soccer has indicated that the match result, location and quality of opposition impact sleep quality and mood (Abbott et al. 2018; Fessi and Moalla 2018). Nevertheless these items can be useful for assessing a player's status and to support decision-making regarding load management.

A strength of the current study is the using of GBRT machine learning technique, which can capture non-linear relationships, to construct an individual predictive model per player (De'Ath 2007). Furthermore, GBRTs can handle long tailed distributions, outliers and are robust to the presence of irrelevant input variables (Friedman 2001). Furthermore, GBRTs allowed evaluating a broad monitoring approach by examining simultaneously the impact of EL, IL and PPW on FPW. These techniques and corresponding findings complement the statistical methods used in earlier research (Fessi, Nouira, et al. 2016; Moalla et al. 2016; Thorpe et al. 2015, 2017) and help to evaluate the usefulness of perceived wellness in monitoring strategies.

The analysis revealed that individual predictive models are more accurate than average player thresholds, which are commonly used. Therefore, such models could improve monitoring strategies, by comparing the reported wellness to the predicted player wellness after each practice. If the reported wellness and predicted wellness differ substantially (i.e., higher or lower scores), this may be a sign to zoom in on the load and responses of a player for detailed interpretations. Moreover, it may aid in individualizing a training program as the models can simulate how a player with a certain wellness status will respond to a given external load.

Some limitations should be acknowledged.
First, a large part of the data could not be used to construct and evaluate the predictive models because the wellness scores were not reported on match and rest days. Since these days do not occur at random, an imputation strategy was necessary to examine the impact of past wellness. This solution provides a reasonable estimation while respecting the data's chronological ordering. Moreover, using this imputation outperformed the baseline method (i.e., small effect sizes were found) which can be considered as the current state of the art when predicting wellness scores for held-aside data samples. Currently, the applied models are not designed to make predictions when the previous three days only contain a combination of match and rest days. However, they do

support all combinations of match, rest- and practice days, when at least one of the previous three days is a practice day. Thus, these models are already versatile enough to be practically useful and the results underscore the importance of daily wellness monitoring.

Second, the load of strength training sessions was not included and may influence the perceived wellness. However, besides the normal injury prevention programs, there were only a small number of separate strength training sessions, and therefore, their influence on the results may be limited.

Third, the perceived wellness questionnaire used in the current study was previously examined in various studies, revealing relationships between load and the wellness items (Buchheit, Cholley, et al. 2016; Buchheit, Racinais, et al. 2013). The custom items of this perceived wellness questionnaire have not been extensively studied concerning their reliability and validity (Saw, Kellmann, et al. 2017). Therefore, there possibly exists a more adequate composition of perceived wellness items for a questionnaire to monitor fatigue and recovery status (ibid.).

Finally, the direction of the relationship between input variables (i.e., EL, IL, and PPW) and FPW is not presented in the current study. In earlier research, higher loads were related to lower perceived wellness (Fessi, Nouira, et al. 2016; Moalla et al. 2016; Thorpe et al. 2015, 2017). The correlation and interactions of input variables complicate the interpretation of non-linear models (Auret and Aldrich 2012). Nevertheless, the findings indicate that a combination of EL and/or IL together with PPW resulted in the best predictive performances of FPW. As presented by Bittencourt and colleagues, (Bittencourt et al. 2016) a complex interaction among a web of determinants may be related to injury occurrence and adaptation. Similarly, this may be the case for perceived wellness. In future research, more extensive analyses using partial dependence plots (Auret and Aldrich 2012) and including other mediating or moderating factors (Windt et al. 2017) may provide additional insights in the direction of relationships between EL, IL, PPW, and FPW.

4.6 Practical Applications

The current study's findings indicate the importance of including both load and preceding perceived wellness in a broad monitoring approach. Additionally, the wellness items fatigue, general muscle soreness and stress levels are the most useful items for assessing the combined impact of load and current wellness status on future wellness. These insights may improve load management strategies in professional soccer. Machine learning techniques may have added value

for analyzing load-wellness relationships and daily practice by the comparison of predicted/expected versus actual wellness scores. Meaningful differences between these scores may be used for load management strategies. However, more research is warranted to indicate the direction of relationships and the influence of specific load indicators.

4.7 Conclusion

The current chapter focused on predicting of future perceived wellness based on preceding load and perceived wellness in professional soccer using individual machine learning models. It was found that the external and/or internal load in combination with preceding perceived wellness resulted in the best predictive performances, indicating the importance of daily wellness status assessment. Including cumulative load for previous days did not improve the predictive performances.

Chapter 5

Fatigue Prediction in Outdoor Runners via Machine Learning and Sensor Fusion

Published as: Op De Beéck, T., Meert, W., Schütte, K., Vanwanseele, B., Davis, J. (2018, July). Fatigue Prediction in Outdoor Runners Via Machine Learning and Sensor Fusion. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 606-615). ACM.

Author contributions T.O.D.B, W.M, B.V. and J.D conceived and designed research;
T.O.D.B collected the data;
T.O.D.B and W.M. preprocessed the data;
T.O.D.B performed the machine learning experiments;
T.O.D.B, W.M, K.S, B.V and J.D interpreted results of experiments;
T.O.D.B drafted manuscript;
T.O.D.B, W.M, B.V, J.D revised manuscript;
T.O.D.B, W.M, B.V, J.D approved final manuscript.

5.1 Abstract

Running is extremely popular and around 10.6 million people run regularly in the United States alone. Unfortunately, estimates indicated that between 29% to 79% of runners sustain an overuse injury every year. One contributing factor to such injuries is excessive fatigue, which can result in alterations in how someone runs that increase the risk for an overuse injury. Thus being able to detect during a running session when excessive fatigue sets in, and hence when these alterations are prone to arise, could be of great practical importance. In this chapter, we explore whether we can use machine learning to predict the rating of perceived exertion (RPE), a validated subjective measure of fatigue, from inertial sensor data of individuals running outdoors. We describe how both the subjective target label and the realistic outdoor running environment introduce several interesting data science challenges. We collected a longitudinal dataset of runners, and demonstrate that machine learning can be used to learn accurate models for predicting RPE.

5.2 Introduction

Worldwide, recreational running is one of the most popular forms of physical activity. In the United States alone, 10.5 million people run regularly, and around 36 million people in total participate in running each year (Messier et al. 2008). While running regularly has many health benefits, injuries hamper these benefits and can even be detrimental for the runner. Unfortunately, runners are prone to overuse injuries, with estimates indicating that anywhere from 29% to 79% of runners suffer at least one overuse injury per year (Gent et al. 2007). Overuse injuries arise due to the repetitive nature of the movements performed during running. These movements repeatedly stress (i.e., apply force to) the same structures (e.g., muscle tissues, tendons or joints) in the body. The effects of the stress accumulate over time, and may eventually exceed the structure's stress tolerance, resulting in an injury (Hreljac et al. 2000; L Bertelsen et al. 2017). Typical overuse injuries in running include pain under the foot (plantar fasciitis) and pain on either the front (patellofemoral pain syndrome) or side (iliotibial band friction syndrome) of the knee (Taunton et al. 2002).

While current research is inconclusive, three categories of factors have been linked to overuse injuries. First are anatomical factors such as high arches, which are inherent to a person. Second are training factors such as excessive long-distance running. Third are biomechanical factors such as the symmetry between the right and left side in a person's running movement. For the third factor, the onset of running fatigue plays a crucial role as it can alter

a person's running style, that is, the movement pattern performed from one step to the next while running. Changes in style can introduce irregularities such as asymmetries between the left and right side, which can elevate the risk of an overuse injury (Schütte, Seerden, et al. 2016). Because these movement alterations are very subtle, they are often not consciously observed by the runner.

Thus, *predicting an individual's fatigue state* based on his currently observed running style has the potential to reduce the risk of overuse injury. While this task can naturally be posed as a supervised machine learning problem, several factors make this an extremely challenging task. First, we need to monitor and characterize running style in "the wild", that is, in a real-world outdoor setting (e.g., variations in weather, running speeds, etc.) in contrast to traditional, controlled laboratory conditions (e.g., running at a fixed-speed on a treadmill). Second, due to several inherent physiological and morphological differences, individuals will respond differently to the same type of exercise. Third, measuring the fatigue state is highly non-trivial. Within the sport science literature, researchers distinguish between different types of fatigue, such as cardiovascular fatigue, biomechanical fatigue, respiratory fatigue, or mental fatigue among others. Which type of fatigue is relevant depends on the task. While heart rate can capture cardiovascular fatigue, overuse injuries are related to biomechanical fatigue. Therefore, we focus on measuring biomechanical fatigue, which can be invasive and expensive (e.g., blood lactate), or represent a subjective measurement of fatigue (e.g., rating of perceived exertion or RPE).

In this chapter, we present a machine learning approach for predicting a runner's RPE, a subjective fatigue measure, based on fusing inertial motion data. We introduce this as an interesting and important data science challenge. In particular, it involves challenges such as analyzing noisy real-world data, handling partial ground truth labels, and reasoning about subjective judgments that vary over time. Our approach is based on defining a variety of biomechanically relevant features that characterize a person's running style. We then build regression models to predict the RPE value at a specific point in time. We evaluated our approach on a longitudinal data set of 29 runners. Each subject completed at least three maximal effort running tests on an outdoor track while wearing four inertial motion units (IMU). We found that, on average, we are able to accurately predict a runner's fatigue state. We found no substantial benefits to fusing the data from multiple sensors compared to using inertial motion data captured from the wrist. Furthermore, we showed that we could effectively deal with the subjectivity of the target variable and the noise introduced by variable running speeds and inter and intra individual differences.

To summarize, this chapter's contributions are:

1. Introducing fatigue prediction in runners as an interesting and important data science problem;
2. Highlighting a number of data science challenges that we encountered while working on this problem;
3. Describing a supervised learning pipeline for this problem that addresses these challenges;
4. Presenting the results of predicting RPE on a real-world longitudinal data set; and
5. Illustrating that several techniques can, to some extent, account for the subjectivity of the target variable and inter and intra individual differences.

5.3 Fatigue Prediction: Definition and Data Science Challenges

This chapter aims to solve the following problem:

Given: Multiple signals collected by inertial sensors placed on a runner.

Do: Learn a model to predict the runner's fatigue state at a given point in time.

This section begins by defining fatigue and how to measure it, then describes our data, and finishes with a discussion of the challenges posed by this task.

5.3.1 Measuring Fatigue

The first issue is measuring an individual's running-based fatigue state, which can be thought of as a hidden variable with a continuum of possible values. Running induced fatigue implies a decrease in running performance (i.e., decreased average speed) due to physiological limitations (i.e., low aerobic capacity, low lactate threshold, or poor running economy) that bring about biomechanical compensations (i.e., alterations in the running kinematics). Several possibilities or markers exist for capturing a runner's fatigue state, yet not all of them are appropriate or suitable for our task. Hence, we developed four primary criteria for selecting the most ideal measure of running fatigue:

Non-invasive. That is, the measurement method or device does not involve the introduction of instruments into the runner's body. Examples of invasive measurements of fatigue include blood lactate (Stoudemire et al. 1996), creatine kinase (Kobayashi et al. 2005), or rectal temperature (Crewe et al. 2008).

Unobtrusive. That is, the measurement method or device does not hinder the runner's comfort in any way and does not interfere with the fluidity of the runner's movement. For instance, obtrusive measurements may include metabolic systems that measure gas exchange (i.e., volume of oxygen consumed (VO_2) or carbon dioxide produced (VCO_2)). Although some of these more portable metabolic systems are wearable, they require a constrained harness, a heavy battery pack, and an uncomfortable face mask that often hinder a runner's comfort.

Non-interruptive. That is, collecting the measure does not interfere with the runner's performance or continuity. Interruptive measures would include both invasive such as blood lactate, as well as non-invasive measurements such as heart rate variability which is known to be inaccurate during dynamic activity (Dong 2016). Interruptive also implies unnecessary physical or mental effort is required by the runner. For instance, more sophisticated rating scales that subjectively quantify fatigue include the *Hooper's Index* (Hooper and Mackinnon 1995) or the *profile of mood states (POMS)* (Williams et al. 1991), which are time consuming and require cognitive loads that force measurements to be attained prior or post running.

Fatigue Specificity. That is, while running the measurement or device provides insights into the musculoskeletal response, which has closer links to overuse injury. For instance, at low to medium aerobic intensities, a runner's biomechanical loading can gradually accumulate and movement compensations may arise while heart rate (HR) can remain relatively stable, suggesting a "mismatch" in fatigue between the musculoskeletal and cardiovascular systems. Thus, although other measures such as HR may fulfill the criteria of being non-invasive, unobtrusive, and non-interruptive, it lacks specificity by only providing insights into the cardiovascular, rather than the musculoskeletal response to running.

Consequently, we measure fatigue using the rating of perceived exertion (RPE), a subjective measure of fatigue that is widely used in running research specifically, and within sport science more generally. Specifically, we use the Borg scale (Borg 1982), where subjects indicate their perception of exertion between 6 (i.e., no exertion) and 20 (maximal exertion). Because it is subjective,

RPE should be viewed a partial truth label. However, RPE has several advantages because it is non-invasive, unobtrusive, and non-interruptive due to its measurement simplicity. Importantly, RPE also has fatigue specificity, given that it provides a more holistic view of fatigue that is said to represent feedback from cardiovascular, respiratory and musculoskeletal systems (Crewe et al. 2008). Furthermore, RPE has been shown to model a runner's performance better in the real-world compared to heart rate which is less responsive to different terrain types (Borg 1998). Thus, RPE is an appropriate and validated marker of a runner's fatigue (Borg 1982).

5.3.2 Data

The data used to train our model consists of longitudinal data for 29 runners. Of these 29 runners, six runners were self-identified as novice runners, 19 as recreational, and four as sub-elite. Moreover, six runners were self-reported as untrained, 17 as moderately trained, and six runners as well-trained. The average age of the runners was 24 with a standard deviation of 6.6 years and a range between 18 and 55. In total, data from 98 trials was collected, where 20 runners completed three trials, seven completed four trials, and two completed five trials. Each trial consists of completing a 3200 meter run on an outdoor track (one lap is 400 meters). Each runner was instructed to use a self-selected pacing strategy to run the trial such that they were fatigued by the end of the run and would reach a RPE between 16 and 20 (very exerted). Running outdoors means the test more naturally mimics the running style of a runner's regular training sessions compared to running on a treadmill at a controlled speed. The study protocol was designed in collaboration with biomechanics researchers with extensive expertise in collecting and analyzing running data. From a sports science perspective, this is a larger data set than usual because data collection is very time consuming, with this collection effort taking > 4 months. The study was conducted according to the requirements of the Declaration of Helsinki and was approved by the KU Leuven ethics committee (file number: s59353).

Prior to starting the run, we explained the RPE scale. The scale ranges from 6 through 20 inclusive and the runner is free to pick any integer value in this range. To help the runners understand the scale, we provided verbal fatigue anchors (e.g., 15 = "hard", 17 = "very hard", 19 = "extremely hard") for every odd number on the scale. Then each participant ran one warm up lap followed by the 3200 meter test. The runners reported their RPE after each lap, including the warm up lap, yielding nine RPE values per trial. Figure 5.1 visually illustrates the protocol. Per lap RPE was thought to be a reasonable time-frame to capture fatigue changes without hampering the runner's performance (e.g., additional mental fatigue or distractions caused by frequent RPE measurements).

During the test, six 1024Hz inertial motion unit (IMU) (Shimmer 3, Shimmer, Dublin) and a strap-based heart rate monitor (Garmin Forerunner 210, Garmin, Schaffhausen) at 1 Hz were attached to the runner. Each IMU contains an accelerometer, gyroscope and magnetometer that measures one signal for each of the three orthogonal axes per sensor type, resulting in nine signals per IMU. One IMU was attached to each of the left and right: shin bone (anteromedial aspect for the distal tibia), wrist (dorsal carpal ligament) and arm (at the level of the mid-point between the acromiale and the radiale, on the mid-line of the lateral surface of the arm). Unfortunately, sensor errors sometimes caused the data from one of the wrist or arm sensors to be lost. Therefore, only data from one wrist and one upper arm sensor was used. If available, we used the left wrist and left arm sensors. Otherwise, we used the right wrist or right arm sensors. Thus four sensors were considered in total. Finally, each lap time was recorded by a hand-held stop watch.

5.3.3 Challenges in Fatigue Prediction

Using a subjective measure of fatigue as the target label introduces several challenges:

C1: Subjectiveness of Target Label. Different runners will rate their exertion level differently. Moreover, it is often hard for runners to accurately assess the gradual and subtle increases of their exertion level throughout the test.

C2: Accommodation to the Test Protocol. Runners were instructed to run in a way that resulted in a RPE score between 16 and 20 by the end of the trial. Consequently, some runners likely reported a high RPE score at the end because this was the expected behavior, and not a score that reflected their true RPE.

C3: Evolution in Reporting RPE. Most subjects were unfamiliar with the RPE prior to the study, and were perhaps unsure how to use it at first. The longitudinal nature of the study means that runners became more familiar with the scale as they ran more tests, and therefore their use of the RPE potentially evolved across consecutive running tests. This issue is similar to problems associated with working on rating data (e.g., for movie prediction) in machine learning (Koren 2010).

Employing a study protocol that mimics normal running (e.g., outdoors, self-select speed), introduces a number of challenges into the data that should be accounted for:

C4: Pacing Strategies Runners apply different pacing strategies during the test. Experienced runners are able to maintain a nearly constant speed over a test. In contrast, many novice runners start fast, slow down in the middle, and speed up at the end. Furthermore, subjects use past experience to alter their running strategy for subsequent tests. Thus there are both inter and intra subject differences in pacing strategies.

C5: Variable Running Speed Running speed, and changes in it, impact the measured inertial motion data (e.g., higher speed means higher acceleration measurements). As we are only interested in fatigue induced changes in the data we need methods that are robust to speed changes.

C6: Individual Running Style Individual characteristics (e.g., weight, height, fitness level, strength, flexibility and training background) mean that runners will have different running styles. These unique styles affect parameters such as step length, step frequency, and arm movement.

5.4 Our Approach to Fatigue Prediction

In this section, we outline our approach to fatigue prediction. First, we discuss which signals we consider. Second, we describe how to construct examples and how to address the challenges described in Subsection 2.3. Third, we describe which features we compute for each example. Fourth, we discuss how to build models. Figure 5.1 provides an overview of our approach.

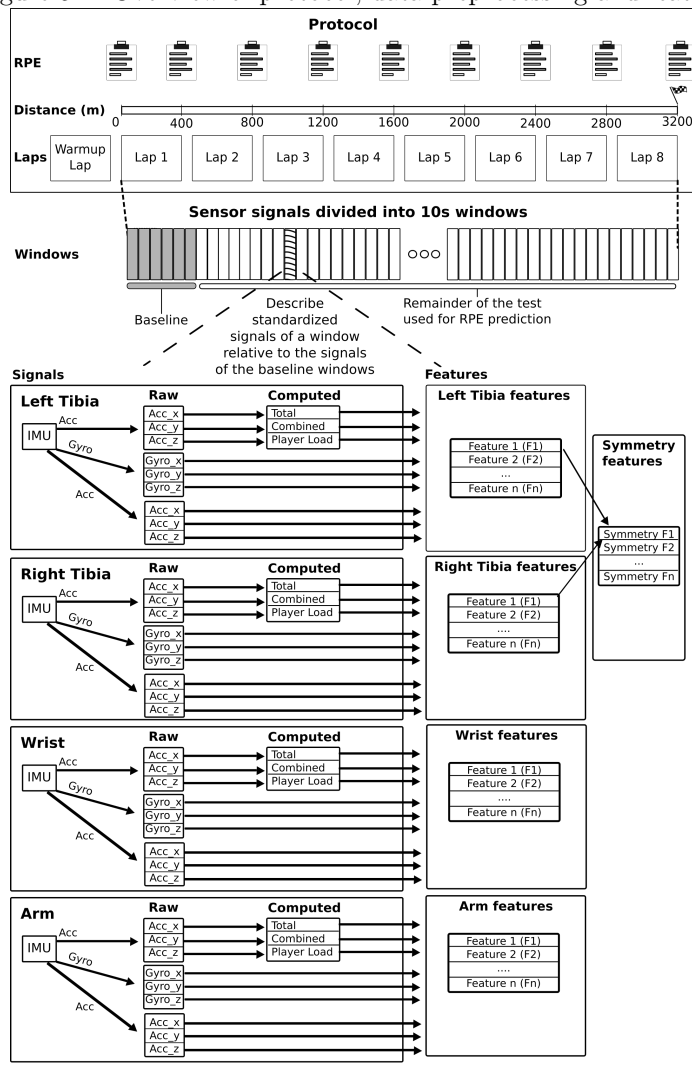
5.4.1 Signals Considered

Our data originally contained heart rate, accelerometer, gyroscope, and magnetometer signals. We altered this in three ways. First, as discussed in Subsection 5.3.1, the heart rate was of limited value because it plateaus quickly. Therefore, we omitted the heart rate data. Second, we also omitted the magnetometer data as this signal, in isolation, does not provide information about running style. Third, we augmented the data by deriving five additional signals from the accelerometer data from an IMU sensor that are commonly used in sport science:

Total Acceleration. This signal is less dependent on the exact attachment of the sensor as it combines the x, y and z acceleration signals at time t_i, and is defined as:

$$\sqrt{{a_{x_i}}^2 + {a_{y_i}}^2 + {a_{z_i}}^2}.$$

Figure 5.1: Overview of protocol, data preprocessing and feature extraction.



Combined Acceleration. The following three signals were found to work well for gait identification because they are less sensitive to the device's attachment (Gafurov et al. 2006). Each signal computes the alignment of the accelerations along one particular axis with respect to the total acceleration. We compute these combined signals, by comparing each of

the x, y, or z axes to the total acceleration:

$$C(v_i) = \arcsin\left(\frac{a_{v_i}}{\sqrt{{a_{x_i}}^2 + {a_{y_i}}^2 + {a_{z_i}}^2}}\right).$$

Player Load™. This signal was developed by Catapult to monitor the changes of accelerations in team sports using an IMU-unit attached to the upper back (Boyd et al. 2011). However, to our knowledge, this signal has neither been used for runners nor calculated based on data collected on the tibia, wrist or arm. In contrast to team sports, where player loads are often aggregated over a training session, we compute an instantaneous change at time t_i as: $\sqrt{\frac{(a_{x_i}-a_{x_{i-1}})^2+(a_{y_i}-a_{y_{i-1}})^2+(a_{z_i}-a_{z_{i-1}})^2}{100}}$.

Alternatively, the raw accelerometer signals can be rotated from a device reference system to an earth reference system (Madgwick 2010). As a result of this process, the z-axis is always perpendicular to the earth's surface. However, early experiments showed that this rotation was not valuable for solving this task. Therefore, it was omitted from all experiments. The reason it was not necessary is that in our protocol, each sensor was firmly attached to a limb. Hence, the sensor moved minimally, if at all, during the course of the trial, meaning that the sensor's relative position was fixed. The relative movement contains all the relevant information and it directly expresses movement in, for example, the left-right direction while the rotation would obfuscate this information.

In summary, each IMU generated three raw and five computed accelerometer signals and three raw gyroscope signals. Thus, with four IMUs, each trial is described by 44 time series signals.

5.4.2 Example Construction and Data Preprocessing

We construct examples by dividing the collected sensor signals into non-overlapping 10 second windows. A 10 second window was chosen because it is sufficiently small to process the data quickly, and represents the typical amount of time used previously with respect to fatigue and running biomechanics (Morin et al. 2011).

Because runners only report the RPE every 400 meters (mean and standard deviation of lap time: 110s ± 18s and range: 72s–177s), we assign an RPE to each example by linearly interpolating between the RPE values reported at the

end of the previous and current lap. RPE is known to linearly change with exercise intensity and running fatigue (Borg 1982).

To deal with the challenges **C1**-**C3** related to the subjective nature of the RPE outlined in Subsection 2.3, we apply min-max normalization to the RPE value based on the current test. This normalization helps to account for inter and intra subject differences in RPE. First, each runner may interpret the scale differently and report a different range of values. Second, the first RPE value reported in a trial may serve as an anchor for subsequent ratings in that trial. The minimum value is the RPE reported after the warm up lap (mean and standard deviation of the RPE after first lap: 10.57 ± 1.97 and range; 7-15) and the maximum value is 20, which is the highest RPE value on the Borg scale. We used this value because using the final RPE value from the test would cause the current label to depend on future data, which is not methodologically sound.

Two other challenges mentioned in Subsection 2.3 are that runners employed different pacing strategies (**C4**) and varied their running speed during the trials (**C5**). To help mitigate the effect of these issues, we standardized every signal within an example. For each signal in an example, we subtract the example's mean value for that signal and divide it by the signal's standard deviation in that example.

On average, each trial generates 78 examples, which results in 7,607 examples in total across all runners and trials. For ten second windows and a sampling rate of 1024Hz, an example consists of 44 time series signals, with 10,240 measurements per signal, and 450,560 measurements in total.

5.4.3 Feature Construction

We want to define features that describe fatigue-related changes in a runner's style. Specifically, because running is a cyclical repetitive movement, and deviations from a runner's pattern may arise due to excessive fatigue, we want to design features that capture changes in the movement pattern.

We consider three broad categories of features: (1) Simple statistical features, which describe aspects of a runner's movement pattern, (2) more advanced sport science features (Jordan et al. 2007; Moe-Nilssen and Helbostad 2004; Tochigi et al. 2012), which capture to what extent a runner is able to copy his movement from one stride (i.e., cyclical motion of one leg) to another, and (3) expert-defined symmetry features (Schütte, Seerden, et al. 2016), which explicitly compare the movement of the left and right leg. While the first two categories are computed by analyzing one sensor's signal, the symmetry features are computed based on two signals (i.e., one from each tibia sensor).

Statistical Features. First, we compute for each signal a set of 15 basic features. We consider four standard features of the signal: *The minimum*, *maximum*, *skew*, and *kurtosis*. We also compute the *average absolute difference (AAD)*, which computes the average absolute difference between each value in a signal and the signal's mean value (Kwapisz et al. 2011)

We compute ten features based on constructing a *binned distribution* of the signal (ibid.). The signal is divided into ten equal sized bins based on its minimum and maximum value. There is one feature per bin which is equal to the proportion of the signal's values that fall in that bin.

Additionally, for the total acceleration, we construct two features based on the *time between peaks*, which was found to be a useful feature in activity recognition (ibid.). Since we only have one activity which is cyclical, a window can be more accurately partitioned into consecutive strides by applying peak detection on the total acceleration signal. The average stride duration and the consistency of the stride durations within a window are then captured by two features: the mean and standard deviation of the stride durations.

Sport Science Features. Second, we compute for each signal three more advanced self-similarity features:

Sample Entropy. This feature measures the complexity of a time-series $T = t_1, t_2, ..., t_n$ as $-log\frac{A}{B}$. Given a length m subsequence in T $seq(x) = t_x \dots t_{x+m}$, B is the number of length m pairs such that $d(seq(i), seq(j)) < r$ where d is the Chebychev distance, and r is a tolerance threshold. Given the set of similar length m pairs, A is the number of pairs that after being extended to length $m + 1$, remain similar (i.e., $d(seq(i), seq(j)) < r$) (Richman and Moorman 2000). Furthermore, it was shown to capture running-fatigue related decline in physiological variability of movement patterns (Schütte, Maas, Exadaktylos, et al. 2015).

Detrended Fluctuation Analysis (DFA). This feature divides the signal into segments of equal length l and quantifies the fluctuations of the signal after subtracting local trends (i.e., by fitting a polynomial curve) for each segment. This process is repeated for multiple values of l to plot the signal's fluctuations as a function of l. The feature's value is the slope of the linear curve fitted through these points (Bryce and Sprague 2012).

Stride Regularity. This feature captures the similarity of consecutive strides. It calculates the value of the first peak in the unbiased autocorrelation signal, which corresponds to comparing the original signal with a copy that was shifted by one stride (Moe-Nilssen and Helbostad 2004). The unbiased autocorrelation

signal is constructed by varying $m = 1 \dots N$, and for each m computing:

$$\frac{1}{N - |m|} \cdot \sum_{i=0}^{N-|m|} x_i \cdot x_{i+m},$$

where N is the number of data points.

Symmetry Features. Third, we compute symmetry features by fusing the signals from the two tibia sensors to capture to what extent a runner is able to replicate his movement from one leg to the other as asymmetries between the left and right side can elevate the risk of an overuse injury (Schütte, Seerden, et al. 2016). Specifically, each symmetry feature is computed as the log difference of the absolute value of the single leg feature calculated on the right side and the absolute value of the single leg feature calculated on the left side (Wetherell 1986).

Normalization. We express the value of each feature relative to a trial-specific baseline for two reasons. First, we expect gradual changes over time of the feature values relative to a non-fatigued state to capture alterations in running style due to fatigue. Second, individual characteristics may affect the observed signal and hence the derived features. After exponentially smoothing all feature values ($\alpha = 0.4$), we use the first six windows to derive a range (i.e., min and max) for each feature. This represents a feature's baseline value for the runner's starting fatigue state. Using this range, we apply min-max normalization to all subsequent values of the feature. To account for inter and intra individual differences we take the absolute value of the normalized feature values.

Summary. To summarize, each IMU has 11 signals. We compute 15 basic features and three sport science features per signal. For the total acceleration signal, two additional features are derived. This means that there are 200 features per IMU. If both tibia sensors are used, then there are 200 additional symmetry features.

5.4.4 Learning Models

We consider three different learning settings, each learned based on different subsets of the data:

1. **All Runners Model (AM).** This setting learns a model using data from all runners. This model attempts to leverage all the data with the assumption that multiple subjects will have similar changes in style as a response to fatigue.

2. **Other Runners Only Model (OM).** This setting builds one model for each runner using only data from other runners. That is, no data is used about the runner for whom predictions will be made. The goal of this setting is to assess how accurate predictions will be if we have no training data available for a specific runner. This is interesting because for first time runners, there will not be data. Furthermore, some runners may not provide RPE value, which are needed to train an individual (or group) model.

3. **Individual Model (IM).** This setting builds one personalized model for each subject using only data from that subject. This model would work well if each subject has a unique alteration in style in response to increasing fatigue.

5.5 Experimental Evaluation

The goal of the empirical evaluation is to assess the viability of predicting RPE in a real-world outdoor setting, provide insights into the input data, and discuss the practical impact of the results. Specifically, we address the following questions:

Q1: How accurately can we predict a runner's RPE based on inertial motion signals?

Q2: How does the location of the sensor's placement on the body affect predictive performance?

Q3: Does fusing the data from multiple sensor locations improve predictive performance?

Q4: Can runners rate their RPE consistently and according to the BORG scale?

Q5: Can we further improve the results using more advanced sport science features and expert knowledge?

Q6: What preprocessing steps are important for accurately predicting RPE?

5.5.1 Experimental Details

We now describe the details of our experiments.

Learners We evaluated four regression techniques: Gradient Boosted Regression Trees (GBRT), Artificial Neural Network (ANN), Linear Regression with Elastic Net regularization (EN), and Linear Regression with Least Absolute Shrinkage and Selection Operator regularization (LASSO). For all models, we used the implementation available in scikit-learn (Pedregosa et al. 2011). For the GBRT, we used the default settings for all parameters except for the following two, because changing them was shown to reduce overfitting (Friedman 2002; Ho 1998): $subsample = 0.4$, $max\_features = 0.9$. For ANN, we used the default parameter settings. For EN and LASSO we used the default parameter settings except for one parameter. For EN we tuned the $L1 - ratio$ parameter, while for LASSO we tuned the alpha parameter. Both were tuned using five fold cross validation on the training set.

Additionally, we consider two baseline predictors. The first baseline model (MIDDLE) always predicts 13, which is the value in the middle of the Borg scale. The second is a personalized, trial-dependent baseline (TD-Baseline) that always predicts the average of the runner's RPE score after the warm up lap and the maximum value of the Borg Scale (i.e., 20). Both of these models can be thought of as predicting the average of a range (i.e., full Borg scale or trial-dependent) assuming that each value in the range is reported the same number of times. We have to use the maximum value of the Borg scale for the top end of the range because when the trial starts, we do not know what the subject's highest reported RPE value will be for that trial.

All four regression techniques and the two baseline models were considered for addressing Q1 and Q2. For the subsequent experiments, we only considered GBRTs because the results from Q1 clearly indicated that GBRT outperform the other techniques on this task.

Features and RPE To answer Q1, Q2, Q3, Q4 and Q6, we train all models on the set of statistical features. To answer Q5, we learn models for different combinations of the statistical, sport science and symmetry features. Additionally, we always train the models using the normalized RPE values, except for in Q4 where we consider the original RPE values as well.

Evaluation We evaluate these models using a cross validation scheme that leaves the last trial of one runner out (i.e., the test set consists of all examples

generated from the last trial for one runner). Because our data is longitudinal, this scheme avoids information leakage between the training set and the testing set arising from the future data of a runner appearing in the training set. Moreover, in the Individual Model setting, this ensures that at least two trials can be used to construct the model. All preprocessing (e.g., standardization of the feature values) is solely done on the training data. The predictions of every model are exponentially smoothed ($\alpha = 0.6$).

To assess the model's accuracy, we report the mean absolute error (MAE). Because we train on the normalized RPE values, we need to convert the predicted RPE value, $rpe_{predicted}$, back to the original BORG scale using: $(20 - rpe_{warmup}) \times rpe_{predicted} + rpe_{warmup}$ where rpe_{warmup} is the RPE reported by the runner after the warmup lap. When computing the MAE, there are several factors that may influence the computation. First, the time a runner needs to complete the protocol can vary across trials. Second, within one trial, variations in speed mean that each lap takes a different amount of time to run. Third, runners have completed a different number of trials. As we do not want our calculation to be unduly influenced by one lap, one trial, or one runner, we calculate a global distance based MAE in two steps. First, we compute for each running test the MAE per lap by assigning each window to a lap. When a window spans two laps, we assign it to the lap in which the majority of the time resides. Second, for each runner, we then compute the average MAE over all laps of that runner. The global MAE is then calculated as the average MAE over all runners. The first step, accounts for different pacing strategies and for variable running speeds within a test, that result in variable lap durations. The second step, accounts for the fact that some runners completed more than three tests.

5.5.2 Experiment and Results for Q1 and Q2

The purpose of this experiment is two-fold. First, we want to evaluate the predictive performance of each learner and each learning setting, that is, the All Runners Model (AM), Other Runners Only Model (OM), and Individual Model (IM) on this task. Second, we want to evaluate the efficacy of the different sensor locations (i.e., the arm (A), wrist (W), and tibia (T)) as it is unclear from the sport science literature where sensors should be placed.

Table 5.1 shows the MAE for all learned models. In terms of learners, GBRTs consistently outperform the other approaches irrespective of the model or sensor location. ANNs clearly perform worse on this task compared to the other learners, probably because ANNs typically require very large amounts of training data (i.e,. more than the 7000 examples we have). The higher MAEs

of the LASSO and EN models compared to the GBRT models suggest that the movement alterations are better captured using non-linear relationships between the features and the target or that the GBRT technique is more effectively dealing with the high number of features. It is also reassuring to see that the GBRT model outperforms both baseline models in all nine scenarios. This illustrates that the performance of these models is non-trivial.

From a more theoretical standpoint, reaching an MAE of 0 is probably not realistic either, as the RPE is a holistic measure that simultaneously captures cardiovascular, respiratory and musculoskeletal fatigue, whereas the IMU sensors only measure musculoskeletal movement patterns.

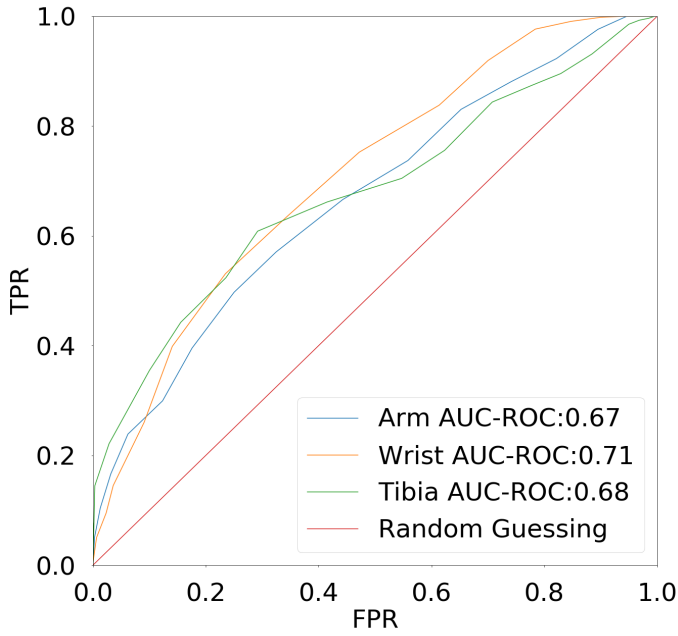
From a learning setting perspective, using data from all runners results in the best predictions, with slight decreases when only data from other runners is considered. In the vast majority of cases, learning an individual model results in worse predictive performance. The predictive performance of the learners in the AM and OM settings are encouraging as it may be difficult to collect large amounts of data for any given individual in practice. Thus, even if a runner provides no labeled fatigue data it is possible to make reasonably accurate predictions. Furthermore, our methodology already takes into account constraints required by the future real-time prediction system (i.e. no future data is used and many of the calculations are parallelizable).

In terms of sensor location, GBRT achieve the best performance using data from the wrist and has slightly worse results on data from the tibia and arm. This contrasts to LASSO and EN which do better on the arm and tibia than the wrist. Practically, it is encouraging that data from the arm and wrist results in accurate predictions as these are locations where a runner may commonly wear a sensor, either in the form of a watch or an attachment of the smartphone to the arm. In contrast, attaching a sensor to the tibia is less common outside of lab setups and possibly more cumbersome, as it might cause pain to the shins and runners may bump into the sensor with their opposite foot while running.

To evaluate the learned GBRT models in a classification setting, we constructed ROC curves by thresholding the predicted RPE to make a fatigued versus not-fatigued prediction. For the ground truth fatigue state, we considered any reported RPE greater than or equal to 16 as representing a truly fatigued runner. This rating corresponds to hard to very hard on the Borg scale. Figure 5.2 shows the ROC curves for the GBRT models learned in the AM setting for each of the three sensor locations. The computed AUC-ROC scores show that all three sensors perform similarly for classifying between *non-fatigued* and *fatigued*. Because each point on a ROC-curve corresponds to a threshold for distinguishing between *non-fatigued*/*fatigued*, the selected threshold could be set according to the individual needs of the runner. When selecting a threshold in practice,

the following are some important considerations. From an injury prevention perspective, a runner might be better off with a threshold that corresponds to a higher TPR at the cost of a slightly higher FPR. As a consequence, the runner will sometimes be advised to stop running before it is actually necessary. Such a threshold choice would be particularly advisable for novice runners, as most running injuries are mainly due to running too far, too fast, too soon (Ballas et al. 1997). More advanced and competitive runners could be less conservative and use a threshold that results in a lower FPR at the cost of a slightly lower TPR.

Figure 5.2: ROC Curves for classifying a runner as being either not-fatigued or fatigued. The results are for the All Runners Model trained using only the statistical features and GBRT for the Arm, Wrist and Tibia sensor locations.



5.5.3 Experiment and Results for Q3

We hypothesized that movement alterations affect the movements of the legs, wrists and upper arms simultaneously while running. Therefore, we assume

Table 5.1: The MAE for predicting RPE for all possible combinations of the four learners, three sensor locations and three learning setting. The three learning settings are the All Runners Model (AM), Other Runners Only Model (OM), and Individual Model (IM).

		AM	**OM**	**IM**
Model	**Sensor**	**MAE**	**MAE**	**MAE**
GBRT	**Arm**	1.99	2.03	1.98
	Wrist	**1.89**	2.04	2.15
	Tibia	1.98	2.08	2.02
ANN	**Arm**	2.92	3.32	14.16
	Wrist	6.48	5.54	19.04
	Tibia	4.37	4.5	42.93
ELASTIC NET	**Arm**	2.28	2.34	2.90
	Wrist	3.16	3.24	2.38
	Tibia	2.09	2.11	3.66
LASSO	**Arm**	2.33	2.38	2.94
	Wrist	2.96	2.92	2.41
	Tibia	2.09	2.12	3.68
MIDDLE BASELINE	**None**	3.00		
TD-BASELINE	**None**	2.60		

that training the GBRT models using features constructed from multiple sensor locations will improve predictive performance. To test this hypothesis, we learned one GBRT model for each learning setting and for each combination of sensor locations.

Table 5.2 shows the MAEs of these models. Combining multiple sensors seems to result in slight improvements in predictive performance, in each learning setting. Practically, there is a trade-off between improved accuracy and the convenience and cost of wearing multiple sensors. It is unlikely that many recreational runners will buy and wear multiple sensors during each run. Therefore, it is reassuring that there are no substantial benefits to fusing the data from multiple sensors. In the future, the evolution of e-textiles means it may be worth revisiting this question as it becomes easier to embed multiple IMU sensors in running apparel.

5.5.4 Experiment and Results for Q4

Because RPE is a subjective measure, different runners might rate their RPE differently. Therefore, we hypothesized that normalizing the RPE values for

Table 5.2: Comparison of the MAE for models learned on all combinations of the four sensor locations: arm, wrist, left tibia, and right tibia. Results for all three learning settings are shown. The models are trained using only the statistical features with a GBRT.

	AM	**OM**	**IM**
SENSORS	**MAE**	**MAE**	**MAE**
Arm (A)	1.99	2.03	1.98
Wrist (W)	1.89	2.04	2.15
Tibia (T)	1.98	2.08	2.02
T-T	1.84	1.90	2.10
W-A	1.89	1.95	2.02
T-A	1.98	2.16	1.89
T-W	1.84	2.01	1.98
T-W-A	1.92	1.98	1.97
T-T-A	1.89	2.00	1.96
T-T-W	**1.74**	1.88	2.06
T-T-W-A	1.83	1.90	1.99

training to account for these inter-individual differences will improve predictive performance.

For each learning setting and sensor location, we trained two GBRT models. The first model was trained using the normalized RPE values, like is done in all other experiments in this chapter. The second model was trained using the originally reported RPE values. Table 5.3 reports the MAE for both approaches for each sensor location. Normalizing the RPE clearly improves the MAEs in the AM and OM learning settings. However, when considering Individual Models, there is no real difference between using NRPE and RPE. These results suggest that runners, at least to some extent, consistently report RPE during consecutive tests. However, different people seem to interpret the BORG scale differently. This might be because the runners had no previous experience with the BORG scale. While previous research found that a learning protocol can improve the validity of the BORG scale (Soriano-Maldonado et al. 2014), it is interesting to see that we can account for these subjective differences between runners, as most runners seem to use their warmup RPE as an anchoring point to rate the remainder of the test.

Table 5.3: The effect of training the model using the normalized RPE (NRPE) values, as is done in all other experiments, and the original RPE values. The normalization is performed to control for the consistency and subjectiveness of the reported RPE values. Results for all three learning settings and each of the three sensor locations are shown. The models are trained using only the statistical features with a GBRT.

	AM		**OM**		**IM**	
	NRPE	**RPE**	**NRPE**	**RPE**	**NRPE**	**RPE**
Sensors	**MAE**	**MAE**	**MAE**	**MAE**	**MAE**	**MAE**
Arm (A)	1.99	2.31	2.03	2.38	1.98	2.11
Wrist (W)	**1.89**	2.24	2.04	2.40	2.15	2.12
Tibia (T)	1.98	2.12	2.08	2.29	2.02	1.97

5.5.5 Experiment and Results for Q5

As the sport science literature has used complex features to study running gait, we hypothesized that these features could complement the set of statistical features and result in improved performance when predicting RPE. Furthermore, we assumed that explicitly describing the symmetry between the movement of the left and the right tibia would capture additional useful information.

In each learning setting, we learned one GBRT model for each combination of feature types: (1) statistical, (2) sports science, (3) statistical and symmetry, (4) sports science and symmetry, (5) statistical and sports science, and (6) statistical, sports science and symmetry. We considered two sensor combinations: wrist (W) and right tibia-left tibia-wrist-arm (T-T-W-A). Note that the symmetry features are only applicable for the second sensor combination.

Table 5.4 reports the MAE for all the different models. The statistical features alone result in the best or close to the best performance in all three learning settings. There are only small changes in the MAE when considering the more advanced features. These results impact the real-world applicability, as the simple statistical features are computationally less expensive to compute. That is, they can easily be computed in real-time and within the resource constraints of a mobile computing platform worn by a runner.

5.5.6 Experiment and Results for Q6

Both running speed and inter and intra individual differences between runners add noise to the computed feature values. Therefore, we hypothesized that we can improve the prediction of RPE by both (1) standardizing the signals per

Table 5.4: The effect of different combinations of statistical, sports science and symmetry features on the MAE. Results for all three learning settings using the data from the wrist (W) and the combined data from the arm, wrist, left tibia and right tibia (T-T-W-A) are shown. The models are trained using GBRT.

	AM		**OM**		**IM**	
Type	**W**	**T-T-W-A**	**W**	**T-T-W-A**	**W**	**T-T-W-A**
Stat.	1.89	1.83	2.04	1.90	1.98	1.99
Sport & Symm.	/	1.99	/	2.06	/	2.02
Stat. & Symm.	/	1.84	/	1.97	/	2.16
Sport	2.14	1.92	2.21	2.07	2.09	2.03
Stat. & Sport	1.99	**1.80**	2.05	2.04	2.09	2.01
Stat. & Sport & Symm.	/	1.84	/	1.91	/	1.97

window before calculating the features and (2) normalizing the feature values with respect to a trial-specific individual baseline for the runner. For each combination of including or excluding these two preprocessing steps, we trained one GBRT model per learning setting using the statistical features calculated on the combined arm, wrist, left tibia and right tibia data.

Table 5.5 reports the MAE for each combination of the two preprocessing steps. The results indicate that normalizing the feature values with respect to the baseline of a runner is an important step that positively impacts predictive performance. This is in accordance with our hypothesis that running style is highly runner specific. However, the standardization of the signal per window has a limited impact on the results.

5.5.7 Discussion

We now revisit the questions posed at the beginning of this section. We can positively answer **Q1** as our evaluation showed that our predictive models have a non-trivial performance when predicting RPE while running. Accurate predictions can be made based on a single sensor that could be located on the wrist, arm or tibia, with the wrist yielding the best results (**Q2**). Furthermore, when evaluating **Q3**, we found that fusing data collected from sensors at multiple locations only resulted in slightly improved predictive performance.

Table 5.5: Impact of standardization and normalization with respect to a trial-specific individual baseline on the MAE. Results for all three learning settings using the combined data from the arm, wrist, left tibia and right tibia are shown. The models are trained using the statistical features and GBRT.

		AM	OM	IM
Standardize Signals	**Normalize w.r.t. Individual Baseline**	**MAE**	**MAE**	**MAE**
yes	yes	**1.83**	1.90	1.99
	no	2.12	2.48	1.95
no	yes	**1.81**	2.00	1.96
	no	2.08	2.50	1.93

Somewhat surprisingly, considering advanced features coming from the sports science literature (**Q5**) did not result in improved performance compared to only considering standard statistical features. We identified several meaningful preprocessing steps that were important to perform in order to account for both inter and intra individual differences and the subjectivity of the RPE scale (**Q4** and **Q6**). To summarize, it is encouraging that promising results are possible using a single sensor attached to the wrist and a set of computationally efficient features.

In terms of moving more towards deploying such a system "in the wild," considering the impact of external factors such as running surface and weather conditions and internal factors such as individual characteristics and pacing strategies would be important. These factor might, for example, influence the interpolation strategy used to assign an RPE value to each window. Furthermore, exploring the relationship between the accumulated load of the impacts endured while running (both in and across multiple training sessions) and RPE, as has been studied for other sports like professional soccer (Jaspers, Op De Beéck, et al. 2018), would be worthwhile.

5.6 Conclusion

This chapter introduced fatigue prediction in runners as a new non-trivial, interesting, and impactful data science problem. Specifically, its non-trivial challenges arise from analyzing sensor data collected in an uncontrolled outdoor environment and the need to resort to a subjective partial and evolving truth label for fatigue. More specifically, we showed that the fatigue status of a runner can accurately be predicted with limited or no prior labeled data of a runner using a set of simple features computed on the data of one IMU-

sensor attached to the wrist. Moreover, our methodology effectively accounts for running speed, the subjectivity of the target variable and inter and intra individual differences between runners. Thus, the results presented in this work are useful and represent a solid start for moving into a real-world application for monitoring the fatigue level of outdoor runners using wearable sensors.

Chapter 6

General Discussion

In this chapter we first discuss the results from Chapters 3 through 5. Afterwards, we revisit this thesis' dissertation statement, followed by a summary of the lessons we have learned. We then identify the limitations of this thesis and discuss potential future research directions. Finally, we end with an overall conclusion.

6.1 Discussion Research Results

This section summarizes the most important results of Chapters 3 through 5. Moreover, we highlight where the methodology of the individual chapters deviates from the general methodology that was outlined in the introduction.

6.1.1 Results of Modeling the Relationships Between the External and Internal Training Load in Professional Soccer

Chapter 3 shows that machine learning models are able to predict RPE based on a large set of external load indicators collected during two seasons from a professional soccer club. Moreover, the use of interpretable models allows identifying the external load indicators that were perceived to be the most exerting by the players. This can help to improve current load management strategies. Lastly, the chapter illustrates the potential of group models as a

tool to individually monitor athletes when limited or no individual data are available.

While this chapter only reports the results of models that are learned using ANN and LASSO, it should be noted that we also experimented with other models such as GBRTs.

Evaluating the GBRT model according to the first experimental setup of Chapter 3 that learns a model on the data of season 1 and evaluates this model on the data of season 2, we obtained an MAE of 0.84.

When the second experimental setup of the chapter is followed that performs a temporal split per season, we obtained an MAE of 0.81 for the GBRT group model and an MAE of 0.78 for the GBRT individual models for the analysis of season 1. For season 2 we found an MAE of 0.81 for the GBRT group model and a MAE of 0.83 for the GBRT individual models.

As the GBRT models did not outperform the LASSO models, we did not include these results in the chapter to not overcomplicate the story. We did report the results of the ANN models to compare to the related work of Bartlett et al. 2017. Furthermore, we did experiment with Elastic Net regularization as well but this technique did not improve the results either.

For the sake of reproducability it is worth adding that we tuned the ANN, LASSO and Elastic Net models on the last 20 percent of the learning set. For ANN we used the scikit-neuralnetwork package (Scikit-neuralnetwork 2019). The ANN network consisted of one hidden sigmoid layer and one linear output layer. Before learning the ANN models we normalized the feature values to values between 0 and 1. We then tuned the number of units [5, 10, 50, 100], learning rate [0.05, 0.01, 0.005, 0.001], and number of iterations [5,10,50,100]. All other parameters were set to the default values. For LASSO and Elastic Net we used the implementation available in Scikit-learn (Pedregosa et al. 2011) and standardized all feature values. For each LASSO model, we tuned the alpha parameter [0.1, 0.2, ...,1] and for each Elastic Net model we trained both the alpha [0.1, 0.2, ..., 1] and the l1-ratio [0.1, 0.2, ..., 1] parameters. For the GBRT models we used the same settings as reported in Chapter 5.

In the second experimental setup, we perform an analysis per season. For each season used a temporal 75-25 train-test split which deviates from the more common 80-20 split. To compare group and individual models on the same test data we chose one shared time split for all players per season. By choosing this time split at 75% instead of at 80%, we increased the size of the test set by five

percent to validate the individual models on more examples, given the limited number of training session a player participates in each season.

The chapter presents similar or better results using LASSO compared to using an ANN. Several factors might explain these results. Since LASSO learns a linear model, it needs fewer learning examples compared to an ANN model. The ANN model is more likely to overfit as it can express more complex relationships between the input variables. Moreover, the L1-regularization of LASSO allows the technique to deal with multicollinearity among the input variables. ANNs do not have this property by default. One obvious extension would be to first use L1-regularization to perform feature selection and reduce the size of the input space before training the ANN network.

According to the first experimental setup of Chapter 3 that learns a model on the data of season 1 and evaluates this model on the data of season 2, we found that this approach does improve the MAE of an ANN group model from 1.09 to 0.85. Yet, when the second experimental setup is followed that performs a temporal split per season, this approach does increase the MAE of the group model and the individual models for the analysis of season 1. For the analysis on season 2 the MAEs of the group model (MAE 0.83) and individual models (MAE 0.82) do not change much. Because these results do not show a clear improvement, we decided to keep the methodology closer to related work of Bartlett et al. 2017.

6.1.2 Results of Predicting Future Perceived Wellness in Professional Soccer

This chapter focuses on predicting future perceived wellness based on preceding load and perceived wellness in professional soccer using individual machine learning models. The chapter shows that the external and/or internal load in combination with the preceding perceived wellness results in the best predictive performances. This highlights the importance of a daily wellness status assessment. Including cumulative load for previous days does not improve the predictive performances.

In this chapter we do not report any results for group models for two reasons. First, we want to focus on the comparison between acute and cumulative loads. Because machine learning techniques are still not commonly used in training load literature, adding a comparison between individual and group models would overcomplicate the chapter's story from a sports science perspective. Only reporting group models is not an option either as domain experts currently believe that individual models are always the best choice. As an extra experiment

we compared group models, that were trained on the data of other soccer players only, with individual models, that were trained separately for each player. To perform this analysis, we used the same time split for all players. We found that these group models did not outperform the individual models nor the baseline model that uses a player's average wellness score as the prediction. Furthermore, we also tried to learn models using LASSO and Elastic Net. Yet, these techniques did not improve the results.

Finally, we select the ELIs based on the results of Chapter 3. We could have used the entire set of available ELIs as we did in the previous chapter. Yet, the main goal of the chapter is to compare the role of different time frames. Computing the entire set of available ELIs for each time frame would have considerably increased the total number of features. While this has the potential to improve the results even further, we already found satisfactory results with the current set.

6.1.3 Results of Fatigue Prediction in Outdoor Runners via Machine Learning and Sensor Fusion

This chapter introduces fatigue prediction in runners as a new non-trivial, interesting, and impactful data science problem. Specifically, its non-trivial challenges arise from analyzing sensor data collected in an uncontrolled outdoor environment and the need to resort to a subjective partial and evolving ground ruth label for fatigue. More specifically, we show that the fatigue status of a runner can accurately be predicted with limited or no prior labeled data of a runner using a set of simple features computed on the data of one IMU-sensor attached to the wrist. Moreover, our methodology effectively accounts for running speed, the subjectivity of the target variable and inter- and intra-individual differences between runners. Thus, the results presented in this work are useful and represent a solid start for moving into a real-world application for monitoring the fatigue level of outdoor runners using wearable sensors.

6.2 The Added Value of Data Science Techniques for the Analysis and Interpretation of Continuous Monitoring Data of Athletes

In this section, we try to support the dissertation statement by focusing on three aspects of the analysis of training load data of athletes where we identified

Table 6.1: MAEs of machine learning group models and baseline constructed on season 1 and evaluated on season 2: Hand picked features versus data-driven feature selection. Abbreviations: ANN, artificial neural networks; LASSO, least absolute shrinkage and selection operator; MAE, mean absolute error.

Method	**Aggregation**	**Bartlett (MAE)**	**Data-driven (MAE)**
ANN	Group	1.45	1.09
LASSO	Group	0.85	0.80
Baseline	Group	1.14	

the added value of data science techniques: to select features, to leverage the data of other athletes, and to model complex relationships.

6.2.1 Complementing Expert Knowledge Using Data-driven Feature Selection Methods

One benefit of using machine learning techniques is that they can learn models that consider multiple input variables. Simultaneously monitoring several variables has been previously identified by the training load monitoring community as one of the keys to successful athlete monitoring (Bourdon et al. 2017). Yet, the question in this case is: "Which variables should be monitored?". Throughout this thesis we evaluated two different hypotheses:

Data-driven Feature Selection or a Set of Hand Picked Features? In Chapters 3 to 5 we consistently used data-driven methods to select a subset of the available features. A second way to select features is to hand select a number of them, which is the approach taken in related work (Bartlett et al. 2017). Tables 6.1 and 6.2 show the comparison of the models presented in Chapter 3 to models that only have access to a set of hand picked ELIs. One can see that the Bartlett setting always results in a worse performance. This illustrates that a data-driven approach to select ELIs can improve the predictive performance of the models.

Simple Statistical Features or Features Based on Domain Knowledge? Chapter 5 evaluates this hypothesis. We compared both sets of features to learn models that predict the RPE of runners. We showed that the simple statistical features outperform both domain features (e.g., stride regularity) and features that explicitly encode domain knowledge (e.g., the asymmetry while running).

Table 6.2: MAEs of machine learning models and baseline for season 1 and season 2. Abbreviations: ANN, artificial neural networks; LASSO, least absolute shrinkage and selection operator; MAE, mean absolute error.

Method	Aggregation	Season 1		Season 2	
		Bartlett (MAE)	Data-driven (MAE)	Bartlett (MAE)	Data-driven (MAE)
ANN					
	Individual	1.10	0.84	0.96	0.85
	Group	1.01	0.81	0.90	0.83
LASSO					
	Individual	0.84	0.81	0.89	0.85
	Group	0.86	0.79	0.88	0.85
Baseline					
	Group	0.99		1.11	

This finding yields three benefits. First, the simple features are more efficient to compute. Second, feature construction is a time-consuming part of the data science pipeline. Since, these features also perform well in other applications on movement data (Decroos, Schütte, et al. 2018; Kwapisz et al. 2011), they have the potential to serve as a go-to set of features in many applications. Third, we can compute these features for every signal of every sensor. Training a model on these features can be a simple method to fuse data of multiple signals and sensors (i.e., on the feature level). Moreover, it can help identify which sensor locations or signals are most predictive. Yet, this does not imply that domain knowledge is unnecessary. The results of Chapter 5 suggest that other parts of the data science pipeline should first deserve our focus. First, domain knowledge should help to design a correct evaluation methodology. Second, understanding the domain helps to understand which context (e.g., by defining an individual baseline, the subjectivity of the target label) could provide value (see 6.3.1).

6.2.2 Individual Monitoring of Athletes Using Group Models

In Chapter 3 and 5, we showed that group models can be valuable to make predictions for athletes with little to no data. The group models of Chapter 3 performed as well as or better than individual models. This suggests that group models can leverage the data of other athletes that behave similarly. In Chapter 5 we obtained the best results when combining the data of a runner with the data of other runners. Here it is important to note that we first applied several strategies to adjust for the differences between runners.

Individual models are likely to outperform group models for cases where a lot of data about the individual are available. Learning individual models can be considered as a strategy to implicitly encode context about the athlete. Yet, in sports, having a lot of individual data is not always feasible (e.g., due to injuries or transfers).

6.2.3 Modeling Complex Relationships to Interpret Continuous Monitoring Data

Responses to training as well as movements while running can be non-linear (Bourdon et al. 2017; Schütte, Maas, Venter, et al. 2015). Therefore, we hypothesized that machine learning techniques that can capture these types of relationships have the potential to work well and outperform linear models. The results from Chapter 5 did indeed support this hypothesis. In Chapter 4 we found a nontrivial effect using a non-linear GBRT model. Yet, in Chapter 3 we found that the linear LASSO technique performed slightly better compared to the ANN models. This shows that linear models can still be useful, especially when relatively limited data is available. We should note here that a traditional least square approach did not work because the input variables are not independent of each other. This highlights another benefit of many machine learning techniques: their ability to deal with multicollinearity among the input variables. While the training load community cautioned that the benefits of machine learning come at the cost of a loss of interpretability (Bourdon et al. 2017), we showed that there exist other advanced techniques beyond ANNs that perform well but are still interpretable to some extent.

6.3 Lessons Learned for Data Scientists

Throughout this thesis we have identified several challenges that arise when analyzing continuous monitoring data of athletes. In this section we try to formulate some lessons learned that can hopefully help other data scientists that want to work in this field.

6.3.1 Contextualization of the Data Is Important

Contextualization of data can help to improve their interpretation. Raw data are hard to interpret, as for most real-world data no normative data exists. Thus, we should evaluate raw data in a broader context. The term context can refer

to many confounding factors. Yet, within the scope of this thesis, we mainly focused on accounting for the intra-and inter-individual differences between athletes that are caused by both the individual characteristics of athletes and the dynamic nature of these characteristics.

We found that incorporating domain knowledge by modeling the context of the data was more effective compared to using domain knowledge to find good features. This thesis employs several technical strategies to encode context either explicitly or implicitly.

In Chapter 5 we have used two explicit strategies to adjust for inter- and intra-individual differences. As one strategy, we defined the first 60 seconds of the data of a runner as the runner's personal baseline on that day. This baseline captures the runner's running style and allows to standardize the feature values of the remaining windows. As another strategy, we corrected for the subjectivity of the target label. We trained the models on normalized RPE scores instead of the reported RPE scores.

In Chapter 4 we employed one explicit strategy to encode context: by adding the previous state as an input (i.e., the wellness score reported before the session). This strategy depends on the quality of the data. If data of the preceding examples are missing, it is not always trivial to impute the previous state. Thus, small details in the data collection protocol can have a big impact on the data analysis (e.g., not reporting wellness scores on match and rest days).

We also experimented with this strategy in Chapter 5 by adding the previously reported RPE score as a feature. While including the previous state improved the predictive accuracy of the models, we found that the predictive models only relied on the previous state feature. Thus, these models learned the study's protocol. Since we wanted to model the relationship between biomechanical movements and RPE, we omitted this strategy.

In Chapter 4 we also applied an implicit strategy. By training individual models for every athlete we implicitly adjusted for a player's subjectivity of the RPE and wellness scores.
In Chapter 3 we employed another implicit strategy by excluding the data from matches from the analysis to focus on the relationship between external load indicators and RPE of training sessions following the methodology of related work (Bartlett et al. 2017).

Including match data would have been only possible for the data of season 2 as GPS data was not collected during matches in season 1. Table 6.3 shows the results of this analysis in case match data would have been included. There are two things to note from these results.

Table 6.3: MAE of machine learning models and baseline for season 2 with and without matches. Abbreviations: ANN, artificial neural networks; CI, confidence interval; LASSO, least absolute shrinkage and selection operator; MAE, mean absolute error.

Method	**Aggregation**	**Practices MAE (90% CI)**	**Practices + Matches MAE (90% CI)**
ANN			
	Individual	0.85 (0.83 - 0.87)	0.87 (0.83-0.91)
	Group	0.83 (0.81 - 0.85)	0.86 (0.82-0.90)
LASSO			
	Individual	0.85 (0.80-0.90)	0.87 (0.83-0.91)
	Group	0.85 (0.80-0.90)	0.87 (0.83-0.91)
Baseline			
	Group	1.11 (1.05-1.17)	1.39 (1.33-1.45)

First, including the match data results in only a slight increase of the MAE for the machine learning models. This illustrates that these models can to some extent account for these implicit context differences. This further strengthens our belief that the machine learning models are able to partly capture the relationship between ELIs and RPE.

Second, the baseline model's MAE increases substantially when match data is included. This increase is because the RPE of matches raises the average RPE of the learning set from 3.48 to 3.96. This average falls somewhere between the training session average and the match average. Therefore, the baseline will both underpredict the RPE of matches and overpredict the RPE of training sessions. This may, in addition to the recent papers of Brito et al. (Brito et al. 2016) and Nassis et al. (Nassis et al. 2017), provide support for a separate evaluation of RPE for matches, as the RPE can be influenced by match-related variables such as location, results, and opponent.

6.3.2 Evaluation of Machine Learning Models in the Context of Continuous Monitoring Data of Athletes Is Non-trivial

In this thesis we have focused on designing evaluation strategies that result in an unbiased assessment of the model's ability to generalize to unseen data. As highlighted in section 2.3.2 it is important that no information is leaked between the learning and testing sets. In Chapters 3 and 4 we used a temporal split of the data. In Chapter 5 we employed a leave-last-trial-of-runner-out validation.

In Chapters 4 and 5 we accounted for the fact that athletes can have a different number of learning examples. In both chapters we required a minimum number of trials per athlete to make sure that enough learning examples were available. We first aggregated the absolute prediction error of each testing example per athlete. Then we aggregated the MAE per athlete over all athletes. In Chapter 5 we added another aggregation step before aggregating the scores per runner. Because the target variable was sampled per lap and not every lap resulted in the same number of examples, we first aggregated the scores per lap. In Chapter 3 we did not explicitly account for the different number of examples per player to more closely match the methodology of related work (Bartlett et al. 2017). Due to the limited overlap of players between seasons and by repeating the analysis for two seasons we believe we still managed to obtain a realistic evaluation.

To interpret the computed performance metrics (e.g., Mean Absolute Error) we compared them to the performance metrics computed for a naive baseline model. In Chapters 3 and 4 we computed effect sizes to assess whether the mean performance of our predictive models was different from the mean performance of the baseline's performance.

6.4 Actionable Insights and Lessons Learned for Sports Scientists and Practitioners

While the previous section highlighted technical contributions that can be useful for data scientists, this section summarizes some actionable insights and lessons learned that can be of value for practitioners and sports scientists:

- Machine learning techniques can provide insight into the relationships between continuous monitoring variables: e.g., decelerations were identified as important to model the relationship between external load and RPE

and data collected from one sensor attached to the wrist could accurately predict the fatigue status of runners.

- Group models can be used when limited or no data are available for an athlete. Yet, this does not imply that an individual approach for the continuous monitoring of athletes is unnecessary. We believe that an individualized approach for monitoring is beneficial. After every practice, match, or window, a practitioner can compare the reported score (e.g., wellness score, RPE score) to the score that was predicted by the predictive models. Large deviations between the predicted and reported values may identify problems early on.

- Machine learning models such as GBRTs can capture non-linear relationships, and are robust to multi-collinearity while still providing interpretability.

- Assessing how predictive models generalize is crucial to draw meaningful conclusions. Yet, the challenges we have identified when evaluating these models were often subtle. Thus, the evaluation methodology should be well thought of and clearly reported.

- As pointed out in the consensus statement on load monitoring (Bourdon et al. 2017), it is non-trivial for practitioners to translate results from literature and apply them to their own population of athletes. Machine learning could provide a good tool to assist with this translation step. A practitioner can copy the methodology from a paper, but retrain and evaluate the models based on data collected in house.

- The quality of the data can have a big impact on the quality of the analysis. Often, simple measures can drastically improve the data quality. Therefore, it is important to already think about the entire data science pipeline during the design of the data collection protocol.

6.5 Limitations

In retrospect, several limitations can be identified when reflecting on Chapters 3 to 5. First, we evaluated the methodology of Chapter 3 and 4 on the data of only one club. Second, while data of two seasons were available for Chapter 3, only data of one season could be used in Chapter 4. Moreover, because wellness-scores were not reported on recovery days and match days this drastically reduced the amount of available data.

We should also acknowledge some limitations of the 3200m all-out protocol of the study on runners. While the outdoor setting introduced interesting challenges that do not arise when collecting treadmill data, one could argue that this protocol still controls the environment of the runners to some extent: all runners ran on the same surface, they ran the same distance and supposedly ran at maximum intensity. In the real-world runners can change surface, distance and intensity. This will likely introduce more challenges. The linear interpolation of the RPE scores for example, might no longer be realistic when the intensity between reported RPE scores varies. Furthermore, this protocol might have introduced a bias toward reporting higher RPE scores near the end of the running test as runners were instructed to run in a way that resulted in a RPE score between 16 and 20 by the end of the trial. Another limitation is that we could not perform the analysis on data captured at the location of the sacrum because during many of the trials, the sensor that we attached on this position came loose due to sweating. This type of analysis has a lot of potential because the more complex features that we computed in Chapter 5 where validated in the literature at this position (Schütte, Seerden, et al. 2018).

Since all three papers use the RPE scale either as a feature or as the target it is interesting to reflect on the use of this measure. RPE is a holistic measure that does not only quantify muscular fatigue, but also cardiovascular and psychological fatigue. As a result, it is unrealistic to expect a perfect RPE prediction based on external load parameters or the description of biomechanical movements. Therefore, it would be interesting to consider more specific subjective measures, such as differential RPE, that differentiate between physiological and biomechanical fatigue pathways (Vanrenterghem et al. 2017).

6.6 Future Work

It would be interesting to evaluate the methodology of Chapters 3 and 4 on data of other clubs. When more data and time would be available this would allow for several additional experiments. First, we could explore the potential of group models to predict player wellness. Second, it could be advantageous to adjust for the subjectivity of the RPE and wellness scores of soccer players, similarly to the strategy used in Chapter 5. Third, adding other individual characteristics of the athlete as features or learning models for athletes that are similar (e.g., based on demographics, position, running style, or more data-driven similarities) have the potential to further improve our results. Fourth, modeling other aspects of the context could be explored as well (e.g. weather conditions, surface, schedule, running speed). Fifth, multi-task learning (Caruana 1997), i.e., where

multiple predictive tasks are learned simultaneously, could be another approach to leverage the similarities between athletes.

While we already explored some aspects of in-session monitoring of athletes in Chapter 5 to analyze the data of runners, more detailed in-session monitoring has the potential to further advance the field of training load monitoring. Currently, most data in this field are aggregated on a session level. As a result, a lot of subtleties are lost. Two athletes with the same aggregated total loads (e.g., total distance) could have reached these loads in a completely different way. Simultaneously monitoring load variables as we did in this thesis is one way to improve the level of detail at which two sessions can be compared. Yet, monitoring how each load variable evolves over time throughout the session will render a more complete picture. Analyzing both the raw GPS and raw accelerometer and gyroscope data of soccer players could help to reveal new interesting patterns that provide insights for training load monitoring. Other improvements would be possible when internal load measurements would be reported more frequently during a training session or when the data of training sessions would be digitally annotated (e.g., with the type of drill). This would for example allow to build models per drill if enough data would be available.

6.7 Conclusion

In retrospect, this thesis contributes to the field of athlete monitoring by showcasing the benefits of using data science techniques to identify important variables, to leverage the data of other athletes to monitor athletes individually, and to model complex (non-linear) relationships that are relevant for the continuous monitoring of athletes. The thesis also contributes to the field of data science by presenting strategies to deal with inter- and intra-subject differences and to correctly evaluate models for the continuous monitoring of athletes. Future research in this area can benefit from an interdisciplinary collaboration between data scientists, sports scientists and domain experts throughout all phases of the data science process to increase the quality of the data that is being collected, to properly contextualize the data, to model relevant relationships and to correctly evaluate and interpret the resulting models.

Bibliography

Abbott, W., T. Brownlee, L. Harper, R. Naughton, and T. Clifford (2018). "The independent effects of match location, match result and the quality of opposition on subjective wellbeing in under 23 soccer players: a case study". In: *Research in Sports Medicine*, pp. 1–14.

Akenhead, R. and G. Nassis (2016). "Training load and player monitoring in high-level football: current practice and perceptions". In: *International journal of sports physiology and performance* 11.5, pp. 587–593.

Akubat, I., S. Barrett, and G. Abt (2014). "Integrating the internal and external training loads in soccer". In: *International journal of sports physiology and performance* 9.3, pp. 457–462.

Alamar, B. and V. Mehrotra (2011). *Beyond "Moneyball": The rapidly evolving world of sports analytics. Analytics Magazine, September/October*.

Auret, L. and C. Aldrich (2012). "Interpretation of nonlinear relationships between process variables by use of random forests". In: *Minerals Engineering* 35, pp. 27–42.

Austin, D., T. Gabbett, and D. Jenkins (2011). "Repeated high-intensity exercise in a professional rugby league". In: *The Journal of Strength & Conditioning Research* 25.7, pp. 1898–1904.

Bagley, C. and B. Ware (2017). "Bump, Set, Spike: Using Analytics to Rate Volleyball Teams and Players". In:

Ballas, M., J. Tytko, and D. Cookson (1997). "Common overuse running injuries: diagnosis and management." In: *American family physician* 55.7, pp. 2473–2484.

Bangsbo, J., F. Iaia, and P. Krustrup (2008). "The Yo-Yo intermittent recovery test". In: *Sports medicine* 38.1, pp. 37–51.

Banister, E., T. Calvert, M. Savage, and T. Bach (1975). "A systems model of training for athletic performance". In: *Aust J Sports Med* 7.3, pp. 57–61.

Barrett, S., A. Midgley, and R. Lovell (2014). "PlayerLoad™: reliability, convergent validity, and influence of unit position during treadmill running". In: *International journal of sports physiology and performance* 9.6, pp. 945–952.

Bartlett, J., F. O'Connor, N. Pitchford, L. Torres-Ronda, and S. Robertson (2017). "Relationships between internal and external training load in team-sport athletes: evidence for an individualized approach". In: *International journal of sports physiology and performance* 12.2, pp. 230–234.

Basheer, I. and M. Hajmeer (2000). "Artificial neural networks: fundamentals, computing, design, and application". In: *Journal of microbiological methods* 43.1, pp. 3–31.

Bishop, C. (2006). "Pattern recognition and machine learning (information science and statistics) springer-verlag new york". In: *Inc. Secaucus, NJ, USA*.

Bittencourt, N., W. Meeuwisse, L. Mendonça, A. Nettel-Aguirre, J. Ocarino, and S. Fonseca (2016). "Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition—narrative review and new concept". In: *Br J Sports Med*, bjsports–2015.

Borg, G. (1982). "Psychophysical bases of perceived exertion". In: *Med Sci Sports Exerc* 14.5, pp. 377–381.

Borg, G. (1998). *Borg's perceived exertion and pain scales.* Human kinetics.

Bourdon, P., M. Cardinale, A. Murray, P. Gastin, M. Kellmann, M. Varley, T. Gabbett, A. Coutts, D. Burgess, and W. Gregson (2017). "Monitoring athlete training loads: consensus statement". In: *International journal of sports physiology and performance* 12.Suppl 2, S2–161.

Boyd, L., K. Ball, and R. Aughey (2011). "The reliability of MinimaxX accelerometers for measuring physical activity in Australian football". In: *International Journal of Sports Physiology and Performance* 6.3, pp. 311–321.

Bransen, L., P. Robberechts, J. Van Haaren, and J. Davis (2018). "Choke or Shine? Quantifying Soccer Players' Abilities to Perform Under Mental Pressure". In: *Proceedings of the 13th MIT Sloan Sports Analytics Conference*.

Brink, M., E. Nederhof, C. Visscher, S. Schmikli, and K. Lemmink (2010). "Monitoring load, recovery, and performance in young elite soccer players". In: *The Journal of Strength & Conditioning Research* 24.3, pp. 597–603.

Brito, J., M. Hertzog, and G. Nassis (2016). "Do match-related contextual variables influence training load in highly trained soccer players?" In: *The Journal of Strength & Conditioning Research* 30.2, pp. 393–399.

Bryce, R. and K. Sprague (2012). "Revisiting detrended fluctuation analysis". In: *Scientific reports* 2, p. 315.

Buchheit, M., Y. Cholley, and P. Lambert (2016). "Psychometric and physiological responses to a preseason competitive camp in the heat with a 6-hour time difference in elite soccer players". In: *International journal of sports physiology and performance* 11.2, pp. 176–181.

Buchheit, M., S. Racinais, J. Bilsborough, P. Bourdon, S. Voss, J. Hocking, J. Cordy, A. Mendez-Villanueva, and A. Coutts (2013). "Monitoring fitness, fatigue and running performance during a pre-season training camp in elite

football players". In: *Journal of Science and Medicine in Sport* 16.6, pp. 550–555.

Buckley, C., M. O'Reilly, D. Whelan, A. Farrell, L. Clark, V. Longo, M. Gilchrist, and B. Caulfield (2017). "Binary classification of running fatigue using a single inertial measurement unit". In: *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, pp. 197–201.

Caruana, R. (1997). "Multitask learning". In: *Machine learning* 28.1, pp. 41–75.

Catapult (2018). https://www.catapultsports.com/. Accessed: 2018-08-03.

Cervone, D., A. D'Amour, L. Bornn, and K. Goldsberry (2014). "POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data". In: *Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA, USA*. Vol. 28, p. 3.

Chang, Y., R. Maheswaran, J. Su, S. Kwok, T. Levy, A. Wexler, and K. Squire (2014). "Quantifying shot quality in the NBA". In: *Proceedings of the 8th Annual MIT Sloan Sports Analytics Conference. MIT, Boston, MA*.

Coutts, A., T. Kempton, C. Sullivan, J. Bilsborough, J. Cordy, and E. Rampinini (2015). "Metabolic power and energetic costs of professional Australian Football match-play". In: *Journal of science and medicine in sport* 18.2, pp. 219–224.

Crewe, H., R. Tucker, and T. Noakes (2008). "The rate of increase in rating of perceived exertion predicts the duration of exercise to fatigue at a fixed power output in different environmental conditions". In: *European Journal of Applied Physiology* 103.5, pp. 569–577. ISSN: 14396319.

Cybenko, G. (1989). "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4, pp. 303–314.

De Brabandere, A., T. Op De Beéck, K. Schütte, W. Meert, B. Vanwanseele, and J. Davis (2018). "Data fusion of body-worn accelerometers and heart rate to predict VO2max during submaximal running". In: *PloS one* 13.6, e0199509.

De'Ath, G. (2007). "Boosted trees for ecological modeling and prediction". In: *Ecology* 88.1, pp. 243–251.

Decroos, T., L. Bransen, J. Van Haaren, and J. Davis (2018). "Actions Speak Louder Than Goals: Valuing Player Actions in Soccer". In: *arXiv preprint arXiv:1802.07127*.

Decroos, T., V. Dzyuba, J. Van Haaren, and J. Davis (2017). "Predicting Soccer Highlights from Spatio-Temporal Match Event Streams." In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 1302–1308.

Decroos, T., K. Schütte, T. Op De Beéck, B. Vanwanseele, and J. Davis (2018). "AMIE: Automatic Monitoring of Indoor Exercises". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 424–439.

Domingos, P. (2000). "A unified bias-variance decomposition". In: *Proceedings of 17th International Conference on Machine Learning*, pp. 231–238.

Dong, J. (2016). "The role of heart rate variability in sports physiology". In: *Experimental and Therapeutic Medicine* 11.5, pp. 1531–1536.

Drew, M., J. Cook, and C. Finch (2016). "Sports-related workload and injury risk: simply knowing the risks will not prevent injuries". In: *Br J Sports Med*, bjsports–2015.

Engels, J. and P. Diehr (2003). "Imputation of missing longitudinal data: a comparison of methods". In: *Journal of clinical epidemiology* 56.10, pp. 968–976.

Fanchini, M., I. Ferraresi, R. Modena, F. Schena, A. Coutts, and F. Impellizzeri (2016). "Use of the CR100 scale for session rating of perceived exertion in soccer and its interchangeability with the CR10". In: *International journal of sports physiology and performance* 11.3, pp. 388–392.

Fessi, M. and W. Moalla (2018). "Postmatch Perceived Exertion, Feeling, and Wellness in Professional Soccer Players". In: *International journal of sports physiology and performance* 20.XX, pp. 1–7.

Fessi, M., S. Nouira, A. Dellal, A. Owen, M. Elloumi, and W. Moalla (2016). "Changes of the psychophysical state and feeling of wellness of professional soccer players during pre-season and in-season periods". In: *Research in Sports Medicine* 24.4, pp. 375–386.

Fitbit (2019). https://www.fitbit.com/. Accessed: 2019-03-07.

Foster, C., J. Florhaug, J. Franklin, L. Gottschall, L. Hrovatin, S. Parker, P. Doleshal, and C. Dodge (2001). "A new approach to monitoring exercise training". In: *The Journal of Strength & Conditioning Research* 15.1, pp. 109–115.

Friedman, J. (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232.

– (2002). "Stochastic gradient boosting". In: *Computational Statistics & Data Analysis* 38.4, pp. 367–378.

Gabbett, T. (2016). "The training—injury prevention paradox: should athletes be training smarter and harder?" In: *Br J Sports Med* 50.5, pp. 273–280.

Gabbett, T., R. Nielsen, M. Bertelsen, N. Bittencourt, S. Fonseca, S. Malone, M. Møller, E. Oetter, E. Verhagen, and J. Windt (2018). *In pursuit of the 'Unbreakable'Athlete: what is the role of moderating factors and circular causation?*

Gabbett, T. and S. Ullah (2012). "Relationship between running loads and soft-tissue injury in elite team sport athletes". In: *The Journal of Strength & Conditioning Research* 26.4, pp. 953–960.

Gafurov, D., K. Helkala, and T. Søndrol (2006). "Biometric Gait Authentication Using Accelerometer Sensor." In: *Journal of computers* 1.7, pp. 51–59.

Gallo, T., S. Cormack, T. Gabbett, and C. Lorenzen (2016). "Pre-training perceived wellness impacts training output in Australian football players". In: *Journal of sports sciences* 34.15, pp. 1445–1451.

Gaudino, P., F. Iaia, A. Strudwick, R. Hawkins, G. Alberti, G. Atkinson, and W. Gregson (2015). "Factors influencing perception of effort (session rating of perceived exertion) during elite soccer training". In: *International journal of sports physiology and performance* 10.7, pp. 860–864.

Gent, B. van, D. Siem, M. van Middelkoop, T. van Os, S. Bierma-Zeinstra, and B. Koes (2007). "Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review". In: *British journal of sports medicine*, pp. 469–480.

Govus, A., A. Coutts, R. Duffield, A. Murray, and H. Fullagar (2018). "Relationship Between Pretraining Subjective Wellness Measures, Player Load, and Rating-of-Perceived-Exertion Training Load in American College Football". In: *International journal of sports physiology and performance* 13.1, pp. 95–101.

Halilaj, E., A. Rajagopal, M. Fiterau, J. L. Hicks, T. J. Hastie, and S. L. Delp (2018). "Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities". In: *Journal of biomechanics*.

Halson, S. (2014). "Monitoring training load to understand fatigue in athletes". In: *Sports medicine* 44.2, pp. 139–147.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*.

Hellard, P., M. Avalos, L. Lacoste, F. Barale, J. Chatard, and G. Millet (2006). "Assessing the limitations of the Banister model in monitoring training". In: *Journal of sports sciences* 24.05, pp. 509–520.

Ho, T. (1998). "The random subspace method for constructing decision forests". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.8, pp. 832–844.

Hoerl, A. and R. Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1, pp. 55–67.

Hooper, S. and L. Mackinnon (1995). "Monitoring Overtraining in Athletes Recommendations". In: *Sports Medicine* 20.5, pp. 321–322.

Hopkins, W. (2002). "A scale of magnitudes for effect statistics". In: *A new view of statistics* 502, p. 411.

Hopkins, W., S. Marshall, A. Batterham, and J. Hanin (2009). "Progressive statistics for studies in sports medicine and exercise science". In: *Medicine+ Science in Sports+ Exercise* 41.1, p. 3.

Hreljac, A., R. Marshall, and P. Hume (2000). "Evaluation of lower extremity overuse injury potential in runners". In: *Medicine & Science in Sports & Exercise* 32.9, pp. 1635–1641.

Hulin, B., T. Gabbett, D. Lawson, P. Caputi, and J. Sampson (2016). "The acute: chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players". In: *Br J Sports Med* 50.4, pp. 231–236.

Hwang, D. (2012). "Forecasting NBA Player Performance using a Weibull-Gamma Statistical Timing Model". In: *MIT Sloan Sports Analytics Conference*.

Impellizzeri, F., E. Rampinini, A. Coutts, A. Sassi, and S. Marcora (2004). "Use of RPE-based training load in soccer". In: *Medicine & Science in sports & exercise* 36.6, pp. 1042–1047.

Impellizzeri, F., E. Rampinini, and S. Marcora (2005). "Physiological assessment of aerobic training in soccer". In: *Journal of sports sciences* 23.6, pp. 583–592.

Jaspers, A., M. Brink, S. Probst, W. Frencken, and W. Helsen (2017). "Relationships between training load indicators and training outcomes in professional soccer". In: *Sports Medicine* 47.3, pp. 533–544.

Jaspers, A., J. Kuyvenhoven, F. Staes, W. Frencken, W. Helsen, and M. Brink (2018). "Examination of the external and internal load indicators' association with overuse injuries in professional soccer players". In: *Journal of science and medicine in sport* 21.6, pp. 579–585.

Jaspers, A., T. Op De Beéck, M. Brink, W. Frencken, F. Staes, J. Davis, and W. Helsen (2018). "Relationships Between the External and Internal Training Load in Professional Soccer: What Can We Learn From Machine Learning?" In: *International journal of sports physiology and performance*, pp. 1–18.

Jobson, S., L. Passfield, G. Atkinson, G. Barton, and P. Scarf (2009). "The analysis and utilization of cycling training data". In: *Sports medicine* 39.10, pp. 833–844.

Jordan, K., J. Challis, and K. Newell (2007). "Speed influences on the scaling behavior of gait cycle fluctuations during treadmill running". In: *Human Movement Science* 26.1, pp. 87–102.

Kobayashi, Y., T. Takeuchi, T. Hosoi, H. Yoshizaki, and J. Loeppky (2005). "Effect of a marathon run on serum lipoproteins, creatine kinase, and lactate dehydrogenase in recreational runners". In: *Research Quarterly for Exercise and Sport* 76.4, pp. 450–455.

Koren, Y. (2010). "Collaborative filtering with temporal dynamics". In: *Communications of the ACM* 53.4, pp. 89–97.

Kwapisz, J., G. Weiss, and S. Moore (2011). "Activity recognition using cell phone accelerometers". In: *ACM SigKDD Explorations Newsletter* 12.2, pp. 74–82.

L Bertelsen, M., A. Hulme, J. Petersen, R. Korsgaard Brund, H. Sørensen, C. Finch, E. Parner, and R. Nielsen (2017). "A framework for the etiology of running-related injuries". In: *Scandinavian Journal of Medicine & Science in Sports*, pp. 1170–1180.

Laux, P., B. Krumm, M. Diers, and H. Flor (2015). "Recovery–stress balance and injury risk in professional football players: a prospective study". In: *Journal of sports sciences* 33.20, pp. 2140–2148.

Li, X., J. Dunn, D. Salins, G. Zhou, W. Zhou, S. Rose, D. Perelman, E. Colbert, R. Runge, and S. Rego (2017). "Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information". In: *PLoS biology* 15.1, e2001402.

Lindsey, G. (1959). "Tatistical Data Useful for the Operation of a Baseball Team". In: *Operations Research* 7.2, pp. 197–207.

Lindstedt, S., P. LaStayo, and T. Reich (2001). "When active muscles lengthen: properties and consequences of eccentric contractions". In: *Physiology* 16.6, pp. 256–261.

Liu, G. and O. Schulte (2018). "Deep reinforcement learning in ice hockey for context-aware player evaluation". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.

Los Arcos, A., J. Yanci, J. Mendiguchia, and E. Gorostiaga (2014). "Rating of muscular and respiratory perceived exertion in professional soccer players". In: *The Journal of Strength & Conditioning Research* 28.11, pp. 3280–3288.

Lovell, T., A. Sirotic, F. Impellizzeri, and A. Coutts (2013). "Factors affecting perception of effort (session rating of perceived exertion) during rugby league training". In: *International journal of sports physiology and performance* 8.1, pp. 62–69.

Lucey, P., D. Oliver, P. Carr, J. Roth, and I. Matthews (2013). "Assessing Team Strategy Using Spatiotemporal Data". In: *Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining*, pp. 1366–1374.

Macdonald, B. (2012). "An expected goals model for evaluating NHL teams and players". In: *Proceedings of the 2012 MIT Sloan Sports Analytics Conference, http://www. sloansportsconference. com.*

Madgwick, S. (2010). "An efficient orientation filter for inertial and inertial/-magnetic sensor arrays". In: *Report x-io and University of Bristol (UK)* 25, pp. 1–32.

Malone, J., R. Di Michele, R. Morgans, D. Burgess, J. Morton, and B. Drust (2015). "Seasonal training-load quantification in elite English premier league soccer players". In: *International journal of sports physiology and performance* 10.4, pp. 489–497.

Malone, J., R. Lovell, M. Varley, and A. Coutts (2017). "Unpacking the black box: applications and considerations for using GPS devices in sport". In: *International journal of sports physiology and performance* 12.Suppl 2, S2–18.

Malone, S., A. Owen, M. Newton, B. Mendes, L. Tiernan, B. Hughes, and K. Collins (2018). "Wellbeing perception and the impact on external training

output among elite soccer players". In: *Journal of science and medicine in sport* 21.1, pp. 29–34.

Marek, P., B. Šedivá, and T. Ťoupal (2014). "Modeling and prediction of ice hockey match results". In: *Journal of quantitative analysis in sports* 10.3, pp. 357–365.

McKinney, W. (2010). "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX, pp. 51–56.

McLaren, S., A. Smith, I. Spears, and M. Weston (2017). "A detailed quantification of differential ratings of perceived exertion during team-sport training". In: *Journal of science and medicine in sport* 20.3, pp. 290–295.

Meinshausen, N. and P. Bühlmann (2010). "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4, pp. 417–473.

Messier, S., C. Legault, C. Schoenlank, J. Newman, D. Martin, and P. Devita (2008). "Risk factors and mechanisms of knee injury in runners". In: *Medicine & Science in Sports & Exercise* 40.11, pp. 1873–1879.

Miller, A., L. Bornn, R. Adams, and K. Goldsberry (2014). "Factorized point process intensities: A spatial analysis of professional basketball". In: *International Conference on Machine Learning*, pp. 235–243.

Mitchell, E., A. Ahmadi, N. O'Connor, C. Richter, E. Farrell, J. Kavanagh, and K. Moran (2015). "Automatically detecting asymmetric running using time and frequency domain features". In: *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, pp. 1–6.

Mitchell, T. (1980). *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research ...

Moalla, W., M. Fessi, F. Farhat, S. Nouira, D. Wong, and G. Dupont (2016). "Relationship between daily training load and psychometric status of professional soccer players". In: *Research in Sports Medicine* 24.4, pp. 387–394.

Moe-Nilssen, R. and J. Helbostad (2004). "Estimation of gait cycle characteristics by trunk accelerometry". In: *Journal of Biomechanics* 37.1, pp. 121–126.

Morin, J., P. Samozino, and G. Millet (2011). "Changes in running kinematics, kinetics, and spring-mass behavior over a 24-h run." In: *Medicine and Science in Sports and Exercise* 43.5, pp. 829–36. ISSN: 1530-0315. URL: http://www.ncbi.nlm.nih.gov/pubmed/20962690.

Nassis, G., M. Hertzog, and J. Brito (2017). "Workload assessment in soccer: an open-minded, critical thinking approach is needed". In: *The Journal of Strength & Conditioning Research* 31.8, e77–e78.

Nédélec, M., A. McCall, C. Carling, F. Legall, S. Berthoin, and G. Dupont (2012). "Recovery in soccer". In: *Sports medicine* 42.12, pp. 997–1015.

Nielsen, R., M. Bertelsen, M. Møller, A. Hulme, J. Windt, E. Verhagen, M. Mansournia, M. Casals, and E. Parner (2018). *Training load and structure-specific load: applications for sport injury causality and data analyses.*

Op De Beéck, T., A. Hommersom, J. Van Haaren, M. van der Heijden, J. Davis, P. Lucas, L. Overbeek, and I. Nagtegaal (2015). "Mining hierarchical pathology data using inductive logic programming". In: *Conference on Artificial Intelligence in Medicine in Europe*. Springer, pp. 76–85.

Opta Sports (2018). http://www.optasports.com. Accessed: 2018-08-03.

Pasteur, R. and K. Cunningham-Rhoads (2014). "An expectation-based metric for NFL field goal kickers". In: *Journal of Quantitative Analysis in Sports* 10.1, pp. 49–66.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg (2011). "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12.Oct, pp. 2825–2830.

Polar (2019). https://www.polar.com/be-nl. Accessed: 2019-03-07.

Richman, J. and J. Moorman (2000). "Physiological time-series analysis using approximate entropy and sample entropy". In: *American Journal of Physiology-Heart and Circulatory Physiology* 278.6, pp. 2039–2049.

Rubin, E. (1958). "Questions and Answers: An Analysis of Baseball Scores by Innings". In: *The American Statistician* 12.2, pp. 21–22.

Saw, A., M. Kellmann, L. Main, and P. Gastin (2017). "Athlete self-report measures in research and practice: considerations for the discerning reader and fastidious practitioner". In: *International journal of sports physiology and performance* 12.Suppl 2, S2–127.

Saw, A., L. Main, and P. Gastin (2016). "Monitoring the athlete training response: subjective self-reported measures trump commonly used objective measures: a systematic review". In: *Br J Sports Med* 50.5, pp. 281–291.

Schulte, O., M. Khademi, S. Gholami, Z. Zhao, M. Javan, and P. Desaulniers (2017). "A Markov Game model for valuing actions, locations, and team performance in ice hockey". In: *Data Mining and Knowledge Discovery* 31.6, pp. 1735–1757.

Schütte, K., J. Aeles, T. Op De Beéck, B. van der Zwaard, R. Venter, and B. Vanwanseele (2016). "Surface effects on dynamic stability and loading during outdoor running using wireless trunk accelerometry". In: *Gait & posture* 48, pp. 220–225.

Schütte, K., E. Maas, V. Exadaktylos, D. Berckmans, R. Venter, and B. Vanwanseele (2015). "Wireless tri-axial trunk accelerometry detects deviations in dynamic center of mass motion due to running-induced fatigue". In: *PloS One* 10.10, e0141957.

Schütte, K., E. Maas, R. Venter, D. Berckmans, and B. Vanwanseele (2015). "Identifying treadmill running fatigue using trunk accelerometry based

measures". In: *International Society of Biomechanics Congress, Date: 2015/07/12-2015/07/16, Location: Glasgow, Scotland*.

Schütte, K., S. Seerden, R. Venter, and B. Vanwanseele (2016). "Fatigue-related asymmetry and instability during a 3200-m time-trial performance in healthy runners". In: *ISBS-Conference Proceedings Archive*. Vol. 34, pp. 933–936.

– (2018). "Influence of outdoor running fatigue and medial tibial stress syndrome on accelerometer-based loading and stability". In: *Gait & posture* 59, pp. 222–228.

Scikit-neuralnetwork (2019). https://scikit-neuralnetwork.readthedocs.io. Accessed: 2019-05-30.

SciSports (2018). http://www.scisports.com. Accessed: 2018-08-03.

Scott, B., R. Lockie, T. Knight, A. Clark, and X. Janse de Jonge (2013). "A comparison of methods to quantify the in-season training load of professional soccer players". In: *International Journal of Sports Physiology and Performance* 8.2, pp. 195–202.

Scott, M., T. Scott, and V. Kelly (2016). "The validity and reliability of global positioning systems in team sport: a brief review". In: *The Journal of Strength & Conditioning Research* 30.5, pp. 1470–1490.

Shimmer (2019). http://www.shimmersensing.com/. Accessed: 2019-03-07.

Smyth, B. and P. Cunningham (2018). "Marathon Race Planning: A Case-Based Reasoning Approach." In: *IJCAI*, pp. 5364–5368.

Soligard, T., M. Schwellnus, J. Alonso, R. Bahr, B. Clarsen, H. Dijkstra, T. Gabbett, M. Gleeson, M. Hägglund, and M. Hutchinson (2016). "How much is too much?(Part 1) International Olympic Committee consensus statement on load in sport and risk of injury". In: *Br J Sports Med* 50.17, pp. 1030–1041.

Soriano-Maldonado, A., L. Romero, P. Femia, C. Roero, J. Ruiz, and A. Gutierrez (2014). "A learning protocol improves the validity of the Borg 6–20 RPE scale during indoor cycling". In: *International Journal of Sports Medicine* 35.05, pp. 379–384.

Spanias, D. and W. Knottenbelt (2013). "Predicting the outcomes of tennis matches using a low-level point model". In: *IMA Journal of Management Mathematics* 24.3, pp. 311–320.

Spencer, M., S. Lawrence, C. Rechichi, D. Bishop, B. Dawson, and C. Goodman (2004). "Time–motion analysis of elite field hockey, with special reference to repeated-sprint activity". In: *Journal of sports sciences* 22.9, pp. 843–850.

STATS' SportVU (2018). http://www.stats.com/sportvu. Accessed: 2018-08-03.

STATSports (2018). https://statsports.com/. Accessed: 2018-08-03.

Stoudemire, N., L. Wideman, K. Pass, C. Mcginnes, G. Gaesser, and A. Weltman (1996). "The validity of regulating blood lactate concentration during running by ratings of perceived exertion". In: *Medicine and Science in Sports and Exercise* 28.4, pp. 490–495. ISSN: 0195-9131.

Taunton, J., M. Ryan, D. Clement, D. McKenzie, D. Lloyd-Smith, and B. Zumbo (2002). "A retrospective case-control analysis of 2002 running injuries". In: *British journal of sports medicine* 36.2, pp. 95–101.

Thorpe, R., A. Strudwick, M. Buchheit, G. Atkinson, B. Drust, and W. Gregson (2015). "Monitoring fatigue during the in-season competitive phase in elite soccer players". In: *International journal of sports physiology and performance* 10.8, pp. 958–964.

– (2017). "The influence of changes in acute training load on daily sensitivity of morning-measured fatigue variables in elite soccer players". In: *International journal of sports physiology and performance* 12.Suppl 2, S2–107.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Tochigi, Y., N. Segal, T. Vaseenon, and T. Brown (2012). "Entropy analysis of tri-axial leg acceleration signal waveforms for measurement of decrease of physiological variability in human gait". In: *Journal of Orthopaedic Research* 30.6, pp. 897–904.

Um, T., V. Babakeshizadeh, and D. Kulić (2017). "Exercise motion classification from large-scale wearable sensor data using convolutional neural networks". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 2385–2390.

Van Craenendonck, T., T. Op De Beéck, W. Meert, B. Vanwanseele, and J. Davis (2014). "Monitoring the Crus for Physical Therapy". In: *1st International Workshop on Machine Learning for Urban Sensor Data*, pp. 1–16.

Van Haaren, J. and J. Davis (2015). "Predicting the final league tables of domestic football leagues". In: *Proceedings of the 5th international conference on mathematics in sport*, pp. 202–207.

Van Haaren, J., V. Dzyuba, S. Hannosset, and J. Davis (2015). "Automatically Discovering Offensive Patterns in Soccer Match Data". In: *Advances in Intelligent Data Analysis XIV*, pp. 286–297.

Van Haaren, J., B. Horesh, J. Davis, and P. Fua (2016). "Analyzing volleyball match data from the 2014 World Championships using machine learning techniques". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 627–634.

Vanrenterghem, J., N. Nedergaard, M. Robinson, and B. Drust (2017). "Training load monitoring in team sports: a novel framework separating physiological and biomechanical load-adaptation pathways". In: *Sports Medicine* 47.11, pp. 2135–2142.

Varley, M., T. Gabbett, and R. Aughey (2014). "Activity profiles of professional soccer, rugby league and Australian football match play". In: *Journal of sports sciences* 32.20, pp. 1858–1866.

Varley, M., A. Jaspers, W. Helsen, and J. Malone (2017). "Methodological considerations when quantifying high-intensity efforts in team sport using global positioning system technology". In: *International journal of sports physiology and performance* 12.8, pp. 1059–1068.

Verhagen, E. and T. Gabbett (2019). *Load, capacity and health: critical pieces of the holistic performance puzzle*.

Vroonen, R., T. Decroos, J. Van Haaren, and J. Davis (2017). "Predicting the potential of professional soccer players". In: *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2017 workshop*.

Wei, X., P. Lucey, S. Morgan, P. Carr, M. Reid, and S. Sridharan (2015). "Predicting serves in tennis using style priors". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 2207–2215.

Weston, M., J. Siegler, A. Bahnert, J. McBrien, and R. Lovell (2015). "The application of differential ratings of perceived exertion to Australian Football League matches". In: *Journal of Science and Medicine in Sport* 18.6, pp. 704–708.

Wetherell, C. (1986). "The Log Percent (L%): An Absolute Measure of Relative Change". In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 19.1, pp. 25–26.

Whelan, D., M. O'Reilly, T. E. Ward, E. Delahunt, and B. Caulfield (2016). "Evaluating performance of the lunge exercise with multiple and individual inertial measurement units". In: *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and . . ., pp. 101–108.

Williams, T., G. Krahenbuhl, and D. Morgan (1991). "Mood state and running economy in moderately trained male runners". In: *Medicine & Science in Sports & Exercise* 23.6, pp. 727–31.

Willy, R. (2018). "Innovations and pitfalls in the use of wearable devices in the prevention and rehabilitation of running related injuries". In: *Physical Therapy in Sport* 29, pp. 26–33.

Windt, J., B. Zumbo, B. Sporer, K. MacDonald, and T. Gabbett (2017). *Why do workload spikes cause injuries, and which athletes are at higher risk? Mediators and moderators in workload–injury investigations*.

Zimmermann, A., S. Moorthy, and Z. Shi (2013). "Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned". In: *Proceedings of the 1st Workshop on Machine Learning and Data Mining for Sports Analytics*.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.

List of publications

Journal Papers

- Jaspers, A.[*], **Op De Beéck, T.**[*], Brink, M., Frencken, W., Staes, F., Davis, J.[+], Helsen, W.[+] (2018). Relationships between the External and Internal Training Load in Professional Soccer: What can We Learn from Machine Learning? International journal of sports physiology and performance, 13(5), 625-630.
- **Op De Beéck, T.**[*], Jaspers, A.[*], Brink, M., Frencken, W., Staes, F., Davis, J.[+], Helsen, W.[+] (2019). Predicting Future Perceived Wellness in Professional Soccer: The Role of Preceding Load and Wellness. International journal of sports physiology and performance, 1-25.
- De Brabandere, A., **Op De Beéck, T.**, Schütte, K., Meert, W., Vanwanseele, B., Davis, J. (2018). Data Fusion of Body-worn Accelerometers and Heart Rate to Predict VO2max during Submaximal Running. PloS one, 13(6).
- Schütte, K. H., Aeles, J., **Op De Beéck, T.**, van der Zwaard, B., Venter, R., Vanwanseele, B. (2016). Surface Effects on Dynamic Stability and Loading During Outdoor Running using Wireless Trunk Accelerometry. Gait & posture, 48, 220-225.

([*]) denotes shared first authorship
([+]) denotes shared last authorship

6.8 Conference Papers

- **Op De Beéck, T.**, Meert, W., Schütte, K., Vanwanseele, B., Davis, J. (2018). Fatigue Prediction in Outdoor Runners Via Machine Learning and Sensor Fusion. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 606-615). ACM.
- **Op De Beéck, T.**, Hommersom, A., Van Haaren, J., van der Heijden, M., Davis, J., Lucas, P., Nagtegaal, I. (2015). Mining Hierarchical Pathology Data using Inductive Logic Programming. In Conference on Artificial Intelligence in Medicine in Europe (pp. 76-85). Springer, Cham.
- Decroos, T., Schütte, K., **Op De Beéck, T.** , Vanwanseele, B., Davis, J. (2018). AMIE: Automatic Monitoring of Indoor Exercises. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 424-439). Springer, Cham.

6.9 Workshop Papers

- Van Craenendonck, T., **Op De Beéck, T.**, Meert, W., Vanwanseele, B., Davis, J. (2014). Monitoring the Crus for Physical Therapy. In 1st International Workshop on Machine Learning for Urban Sensor Data (pp. 1-16).

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
DTAI
Celestijnenlaan 200A box 2402
B-3001 Leuven
first.last@cs.kuleuven.be
http://www.dtai.cs.kuleuven.be